Slides modified from:

PATTERN RECOGNITION
AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

and:

Computer vision: models, learning and inference.

Bayesian probabilities

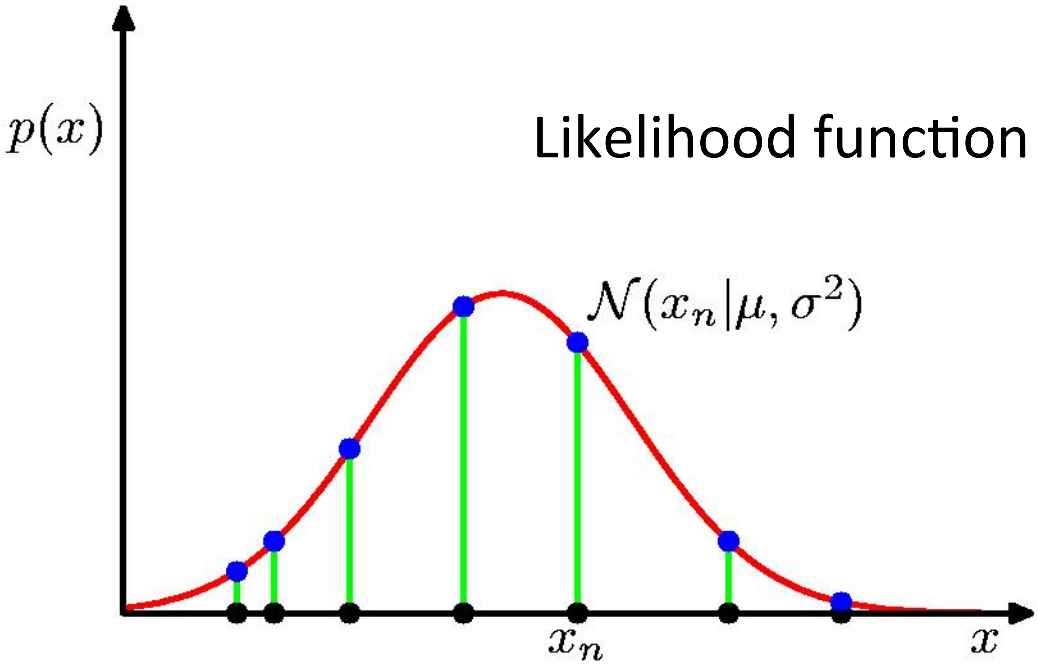
$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

posterior $\propto$ likelihood x prior

Likelihood (function): viewed as function of parameters μ

Expresses how probable the observed data set is for different settings of the parameter μ

Gaussian Parameter Estimation



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

Likelihood for the Gaussian

Assume σ 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).

Maximum (Log) Likelihood

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

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

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

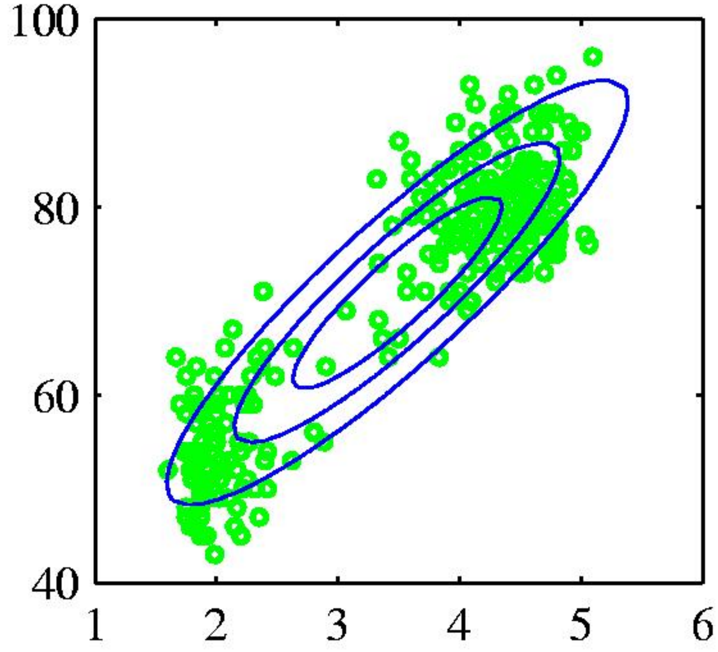
$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

Completing the square over μ, we see that

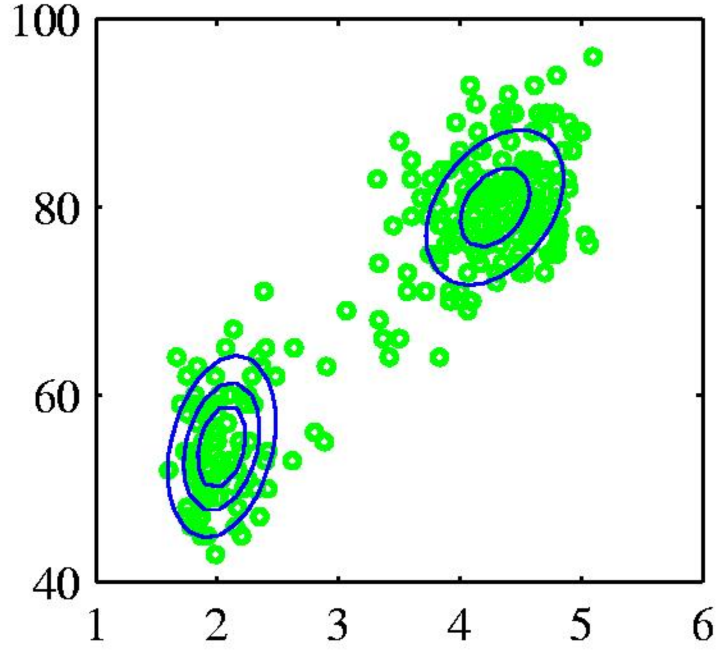
$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

Mixtures of Gaussians (1)

Old Faithful data set



Single Gaussian



Mixture of two Gaussians

Mixtures of Gaussians (2)

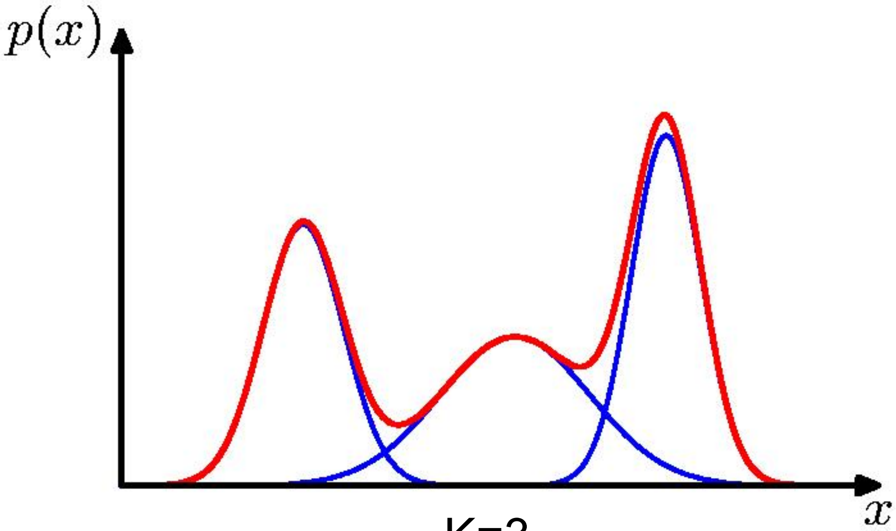
Combine simple models into a complex model:

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

Component

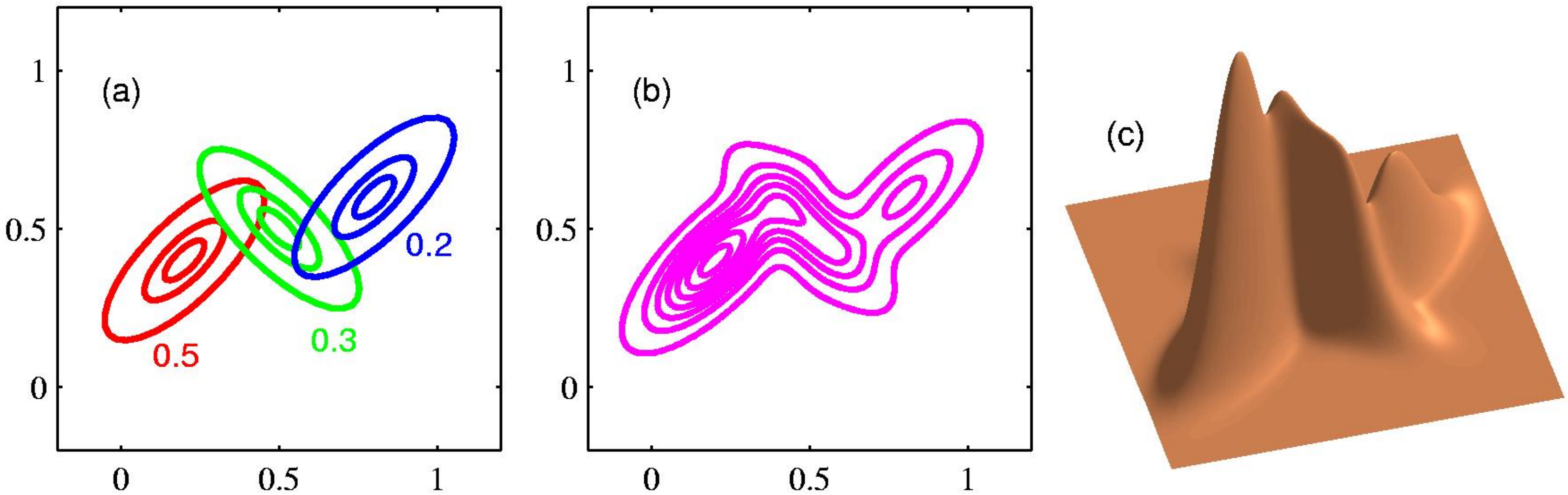
Mixing coefficient

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



K=3

Mixtures of Gaussians (3)



Mixtures of Gaussians (4)

Determining parameters **μ**, **Σ**, and **π** using maximum log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}}$$

Log of a sum; no closed form maximum.

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

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.

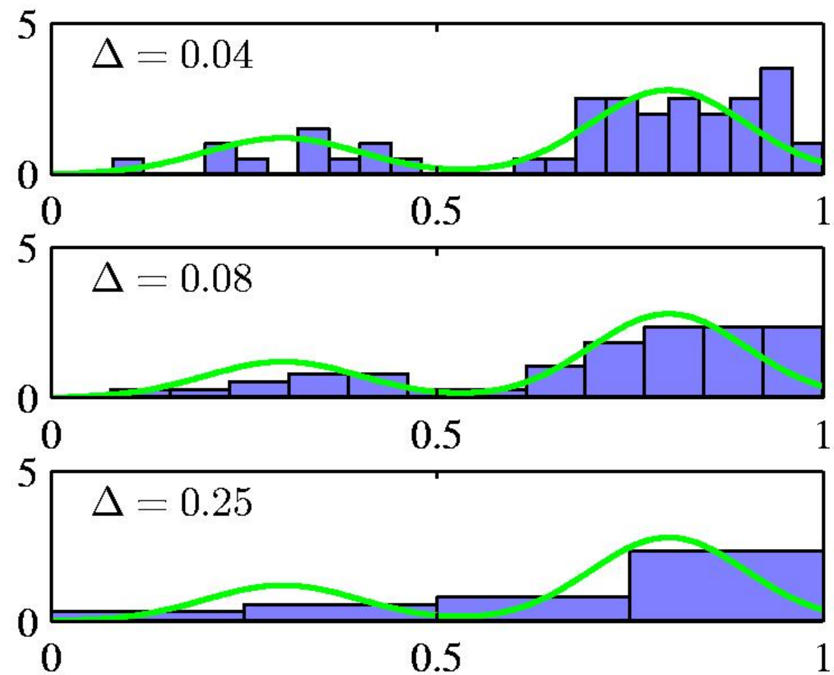
Nonparametric Methods (2)

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

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

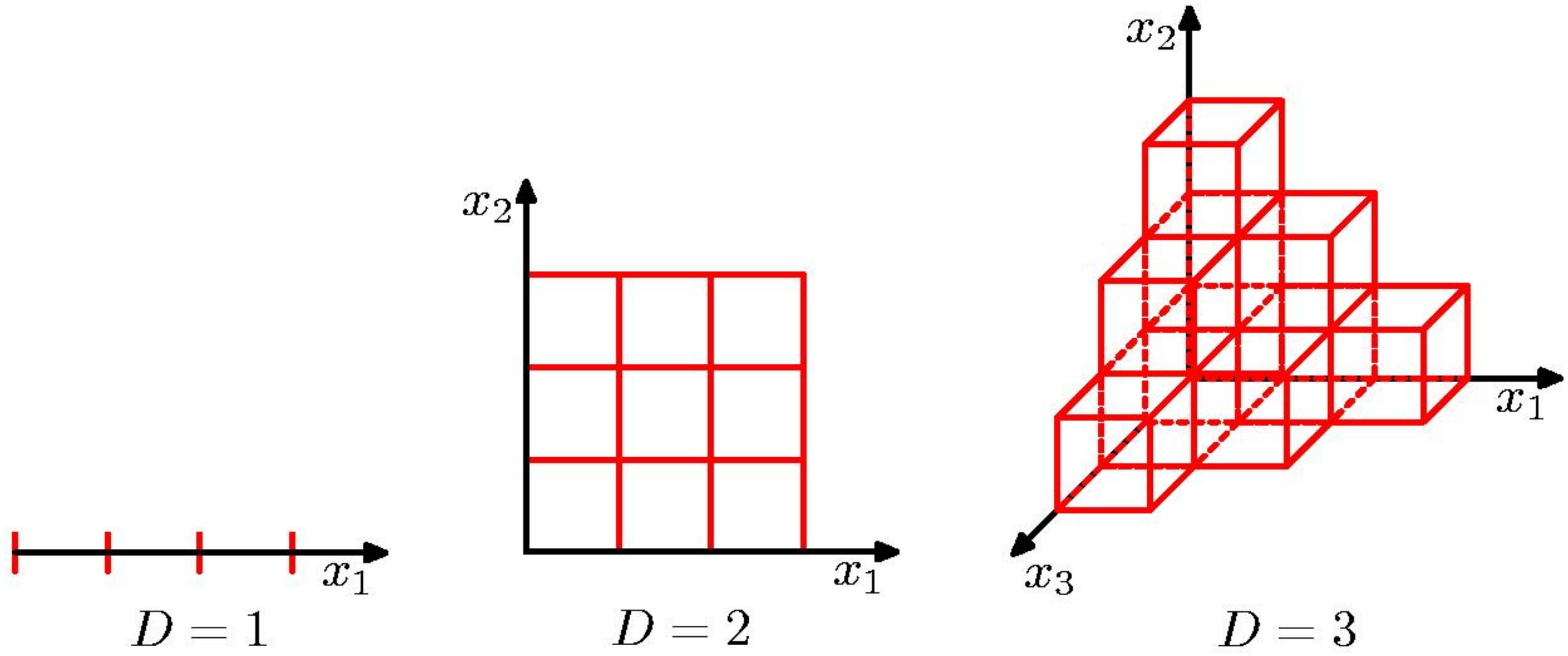
•Often, the same width is used for all bins, $\Delta_i = \Delta$.

•Δ acts as a smoothing parameter.



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

Curse of Dimensionality (1)

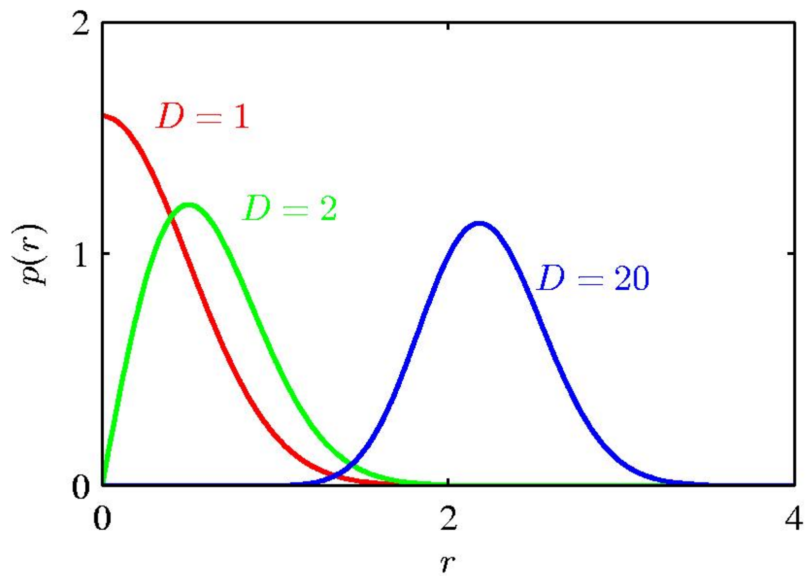


Curse of Dimensionality (2)

Polynomial curve fitting, M = 3

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Gaussian Densities in higher dimensions



Nonparametric Methods (3)

Assume observations drawn from a density p(x) and consider a small region $\mathcal{R}$ 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 $\mathcal{R}$ is Bin(K|N,P) and if N is large

$$K \simeq NP.$$

If the volume of $\mathcal{R}$, V, is sufficiently small, p(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 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)

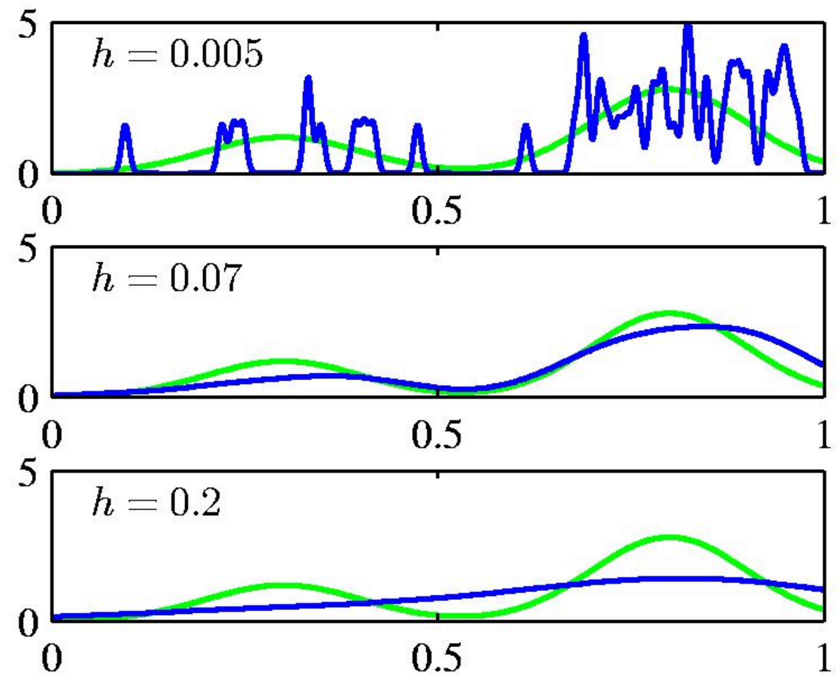
To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

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

Any kernel such that

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

will work.

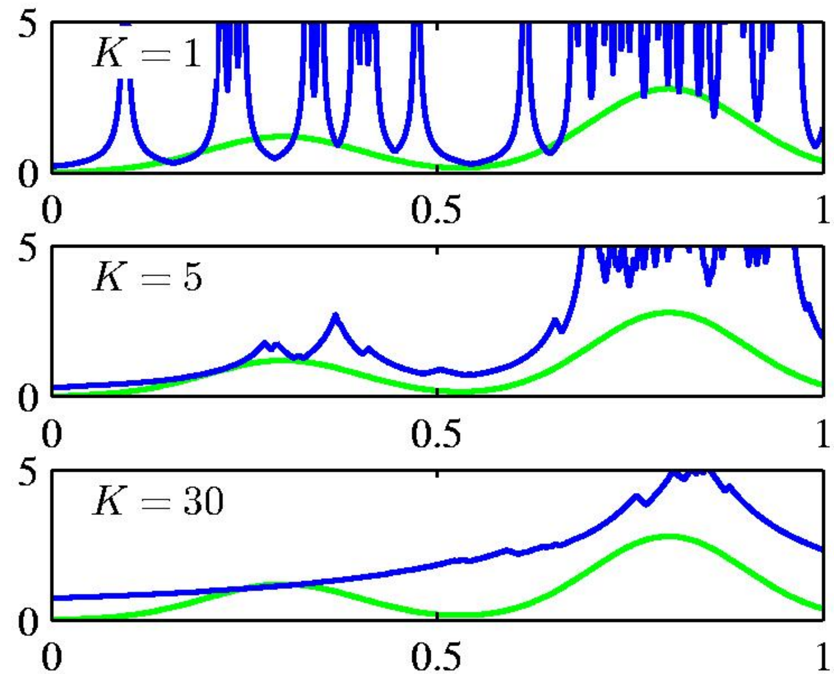


h acts as a smoother.

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.

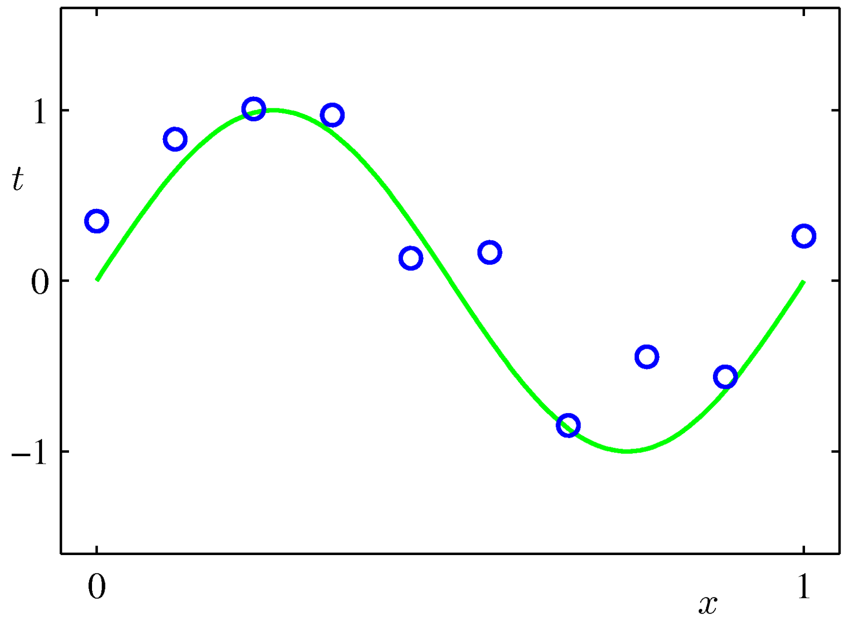
Nonparametric Methods (7)

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.

Linear Basis Function Models (1)

Example: Polynomial Curve Fitting



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Linear Basis Function Models (2)

Generally

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Where $\phi_j(x)$ are known as *basis functions*.

Typically, $\phi_0(x) = 1$, so that w_0 acts as a bias.

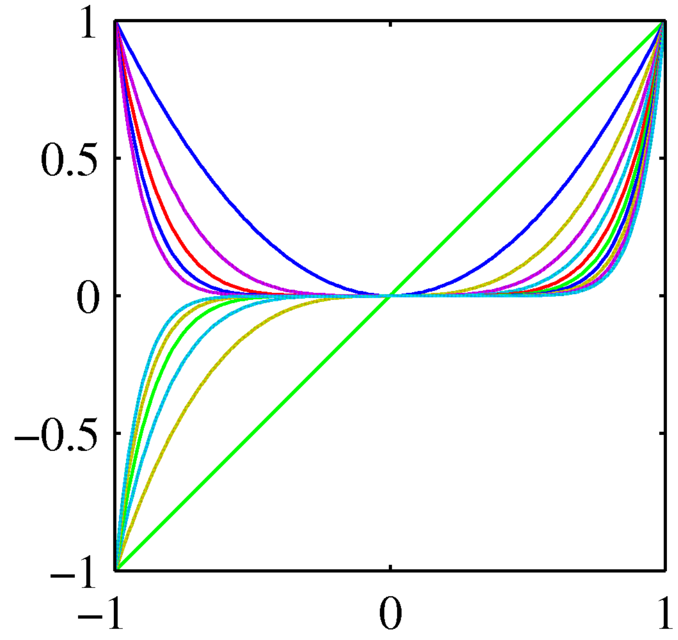
In the simplest case, we use linear basis functions : $\phi_d(x) = x_d$.

Linear Basis Function Models (3)

Polynomial basis functions:

$$\phi_j(x) = x^j.$$

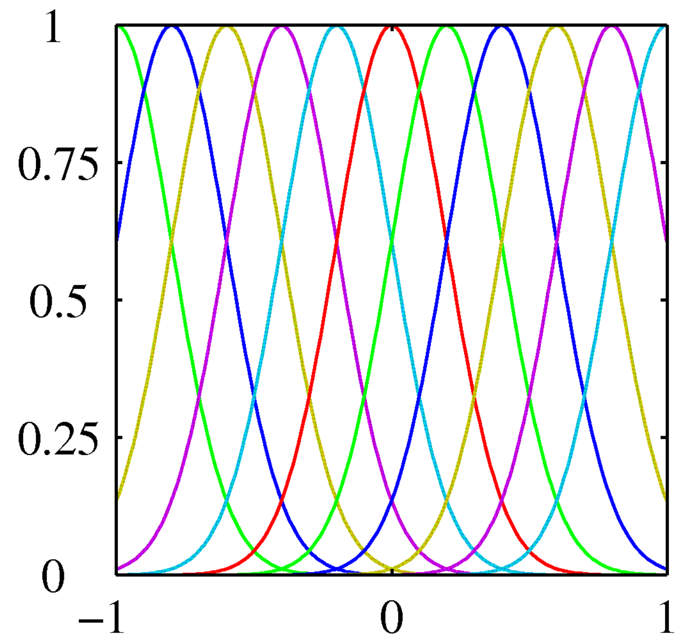
These are global; a small change in *x* affect all basis functions.



Linear Basis Function Models (4)

Gaussian basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



These are local; a small change in *x* only affect nearby basis functions. μ_j and *s* control location and scale (width).

Linear Basis Function Models (5)

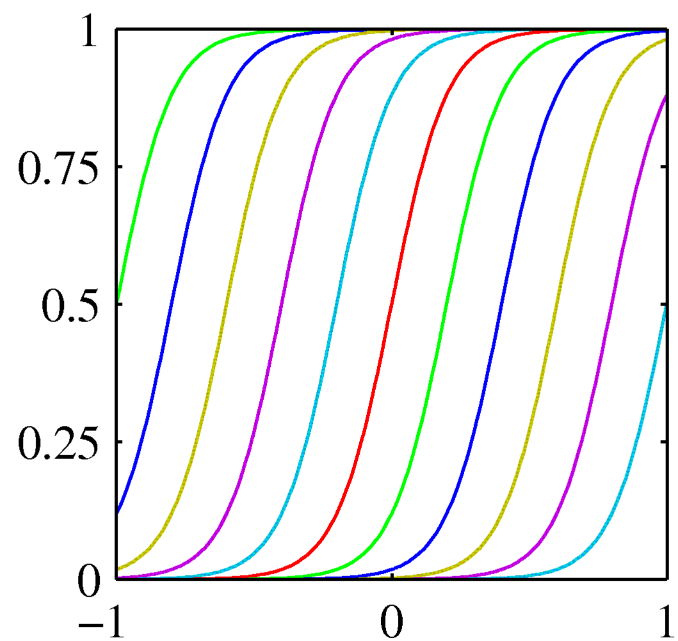
Sigmoidal basis functions:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

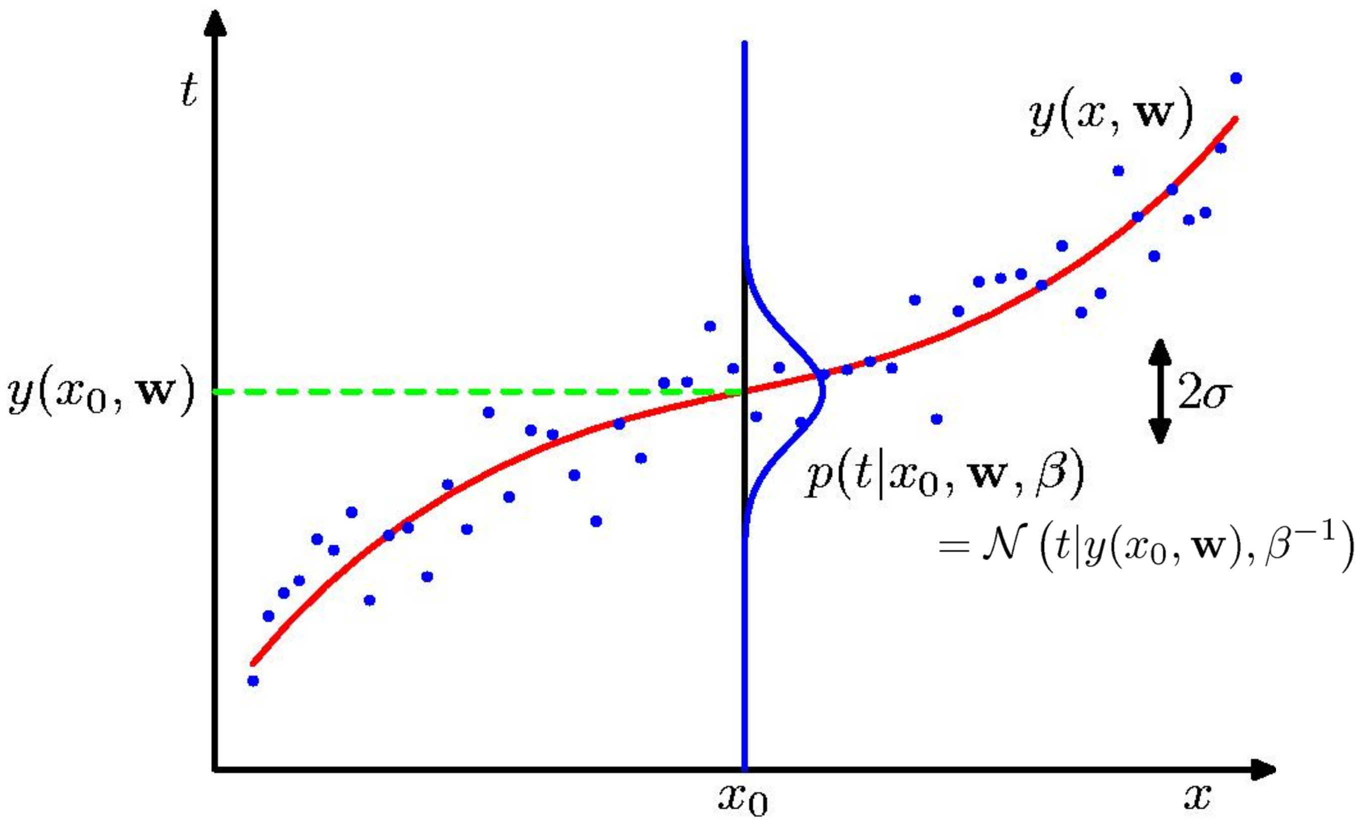
where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

Also these are local; a small change in *x* only affect nearby basis functions. μ_j and *s* control location and scale (slope).



Curve Fitting Re-visited



Maximum Likelihood and Least Squares (1)

Assume observations from a deterministic function with added Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad \text{where} \qquad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

which is the same as saying,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$, we obtain the likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

Maximum Likelihood and Least Squares (2)

Taking the logarithm, we get

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned}$$

where

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

is the sum-of-squares error.

Sum-of-Squares Error Function