T-61.6020 Special Course in Computer and Information Science II
Machine Learning: Basic Principles

Probability Distributions

Mikko Korpela
Jan 22, 2007

Outline

- An overview of Chapter 2 of the book [1]
  - Binary variables
  - Multinomial variables
  - The Gaussian distribution
  - The exponential family
  - Nonparametric methods

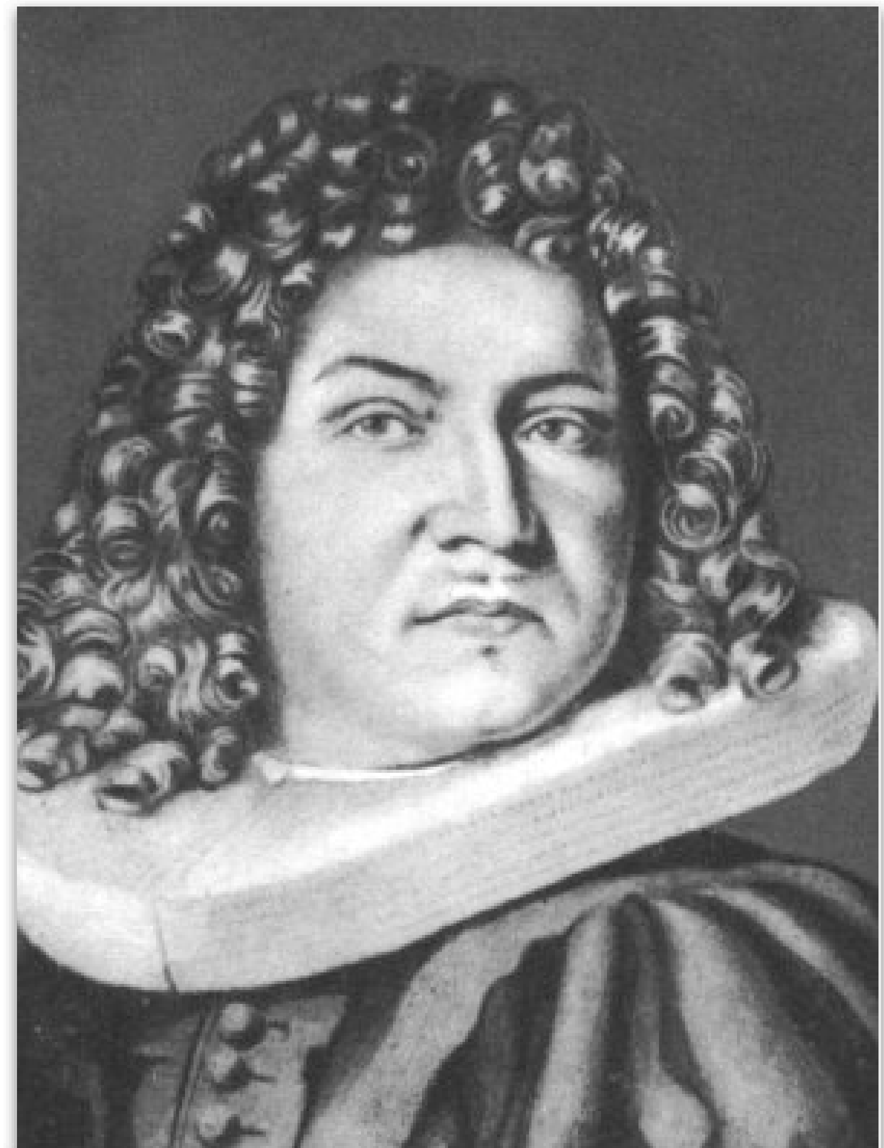
[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 0387310738.

Introduction

- Probability distributions and their properties
- (Probability distributions are) “of great interest in their own right”
- Also “building blocks for more complex models” (later in the book)
- Basic (ill-posed) problem: *density estimation* of a random variable given observations
- Parametric and nonparametric methods

Binary Variables

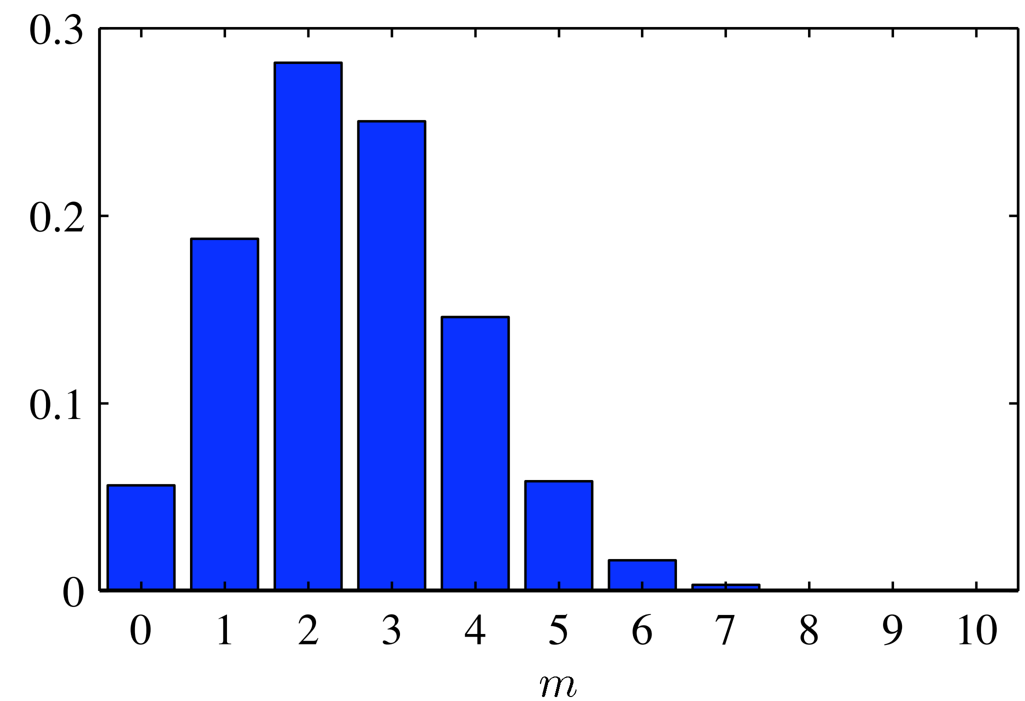
- First, consider a single binary random variable $x \in \{0, 1\}$
- Probability distribution $\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$
- $\mathbb{E}[x] = \mu, \ \mathrm{var}[x] = \mu(1-\mu)$
- Likelihood function for data set $\mathcal{D} = \{x_1, \dots, x_N\}$ is $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n}$
- Maximum likelihood estimator $\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$ or $\mu_{ML} = \frac{m}{N}$ where m is the number of observations $x = 1$



Jacob Bernoulli
1654–1705

Binary Variables
The binomial distribution

- The distribution of the number of ones m in N trials is
  $\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$
- Mean and variance are given by
  $\mathbb{E}[x] = N\mu,\ \mathrm{var}[x] = N\mu(1-\mu)$
  (independence of repeated trials ⇒ means and variances add up)



Binomial distribution

$N = 10$ and $\mu = 0.25$

Binary Variables

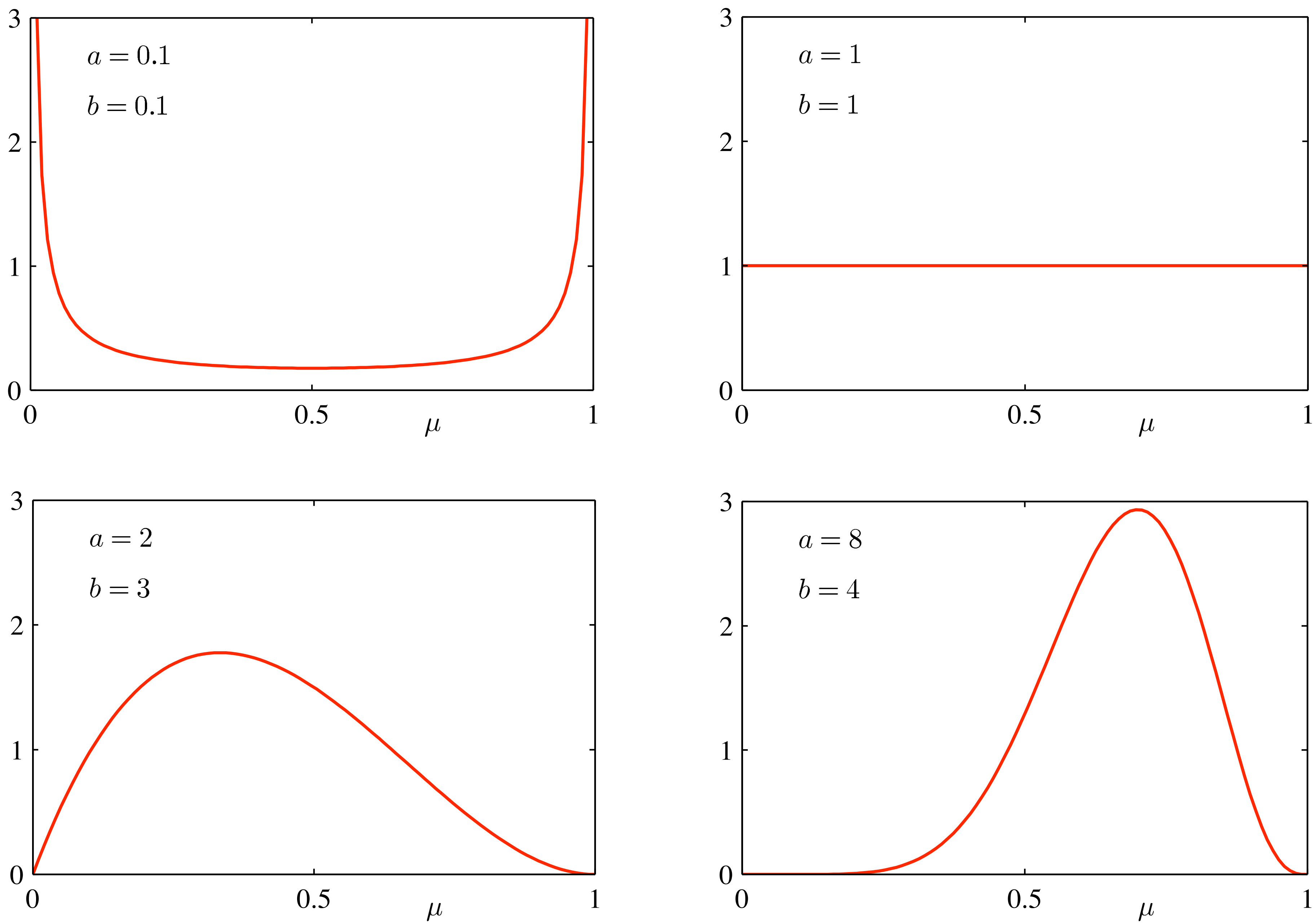
Overfitting and a proposed fix

- Maximum likelihood estimation can results in overfitting
  - Example: Flipping a coin 3 times and observing 3 heads ⇒ $\mu_{ML} = 1$
- Overfitting can be fixed with Bayesian treatment
  - Prior distribution $p(\mu)$ needed

Binary Variables
The beta distribution (1/4)

- The beta distribution
  $$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$
  is a *conjugate prior* for the binomial distr.
  - Conjugacy means that the posterior has *the same functional form* as the prior
- Posterior (where $l = N - m$)
  $$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$$
- Interpretation: a and b in the prior are effective number of observations $x = 1$ and $x = 0$, respectively

Binary Variables
The beta distribution (2/4)



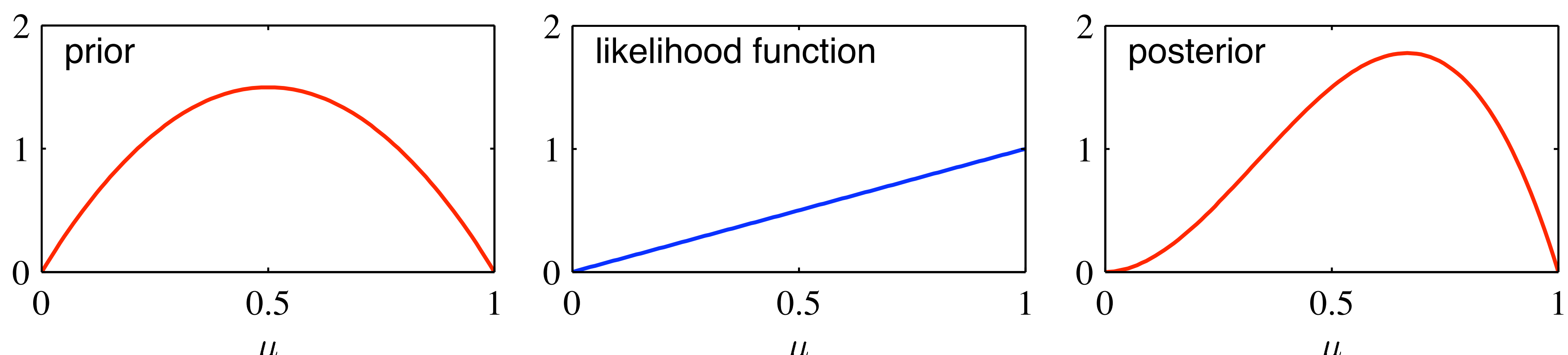
Plots of the beta distribution

Binary Variables
The beta distribution (3/4)

- Predictions, given the prior and observations, can be made with
  $p(x = 1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$
- Result agrees with ML in the limit of infinitely large number of observations
  - General property of Bayesian learning

Binary Variables
The beta distribution (4/4)

- In the Bayesian setting, a *sequential* approach is possible
  - Observations are taken in one at a time or in small batches
  - Old posterior becomes new prior



One step of sequential Bayesian inference. Prior is beta with a=2, b=2. Likelihood corresponds to an observation x=1.

Multinomial Variables

- Consider a discrete variable that can take one of K possible values
- Convenient representation with a vector where one element equals 1, others 0, e.g. $\mathbf{x} = (0,0,1,0,0,0)^T$
- Prob. distrib. $p(\mathbf{x}|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$ with $\sum_k \mu_k = 1$
- Generalization of the Bernoulli distribution
- Likelihood $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k}$ where m_k is the number of observations belonging to category k

Multinomial Variables (2)

- Maximum likelihood estimators $\mu_k^{ML} = \frac{m_k}{N}$
- Distribution of different categories in N observations, the *multinomial distribution*:
  $\mathrm{Mult}(m_1, m_2, \ldots, m_K | \mu, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$
  where $\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$
- Constraint $\sum_{k=1}^{K} m_k = N$

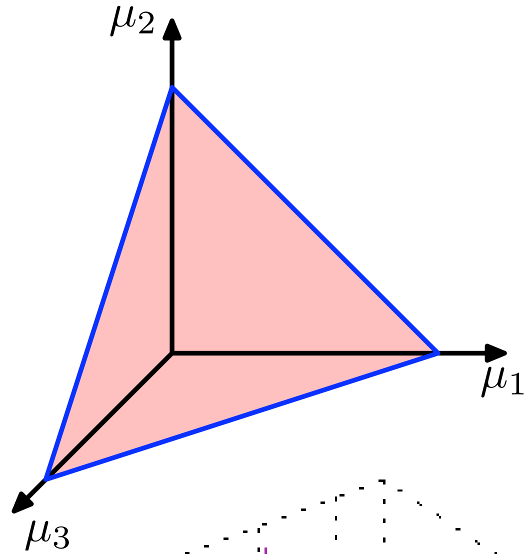
Multinomial Variables
The Dirichlet distribution (1)

- Family of conjugate prior distributions for the parameters $\{\mu_k\}$
- $\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_{})\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$
  where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
- From the posterior (omitted), we see that the α_k are the effective number of observations in each category

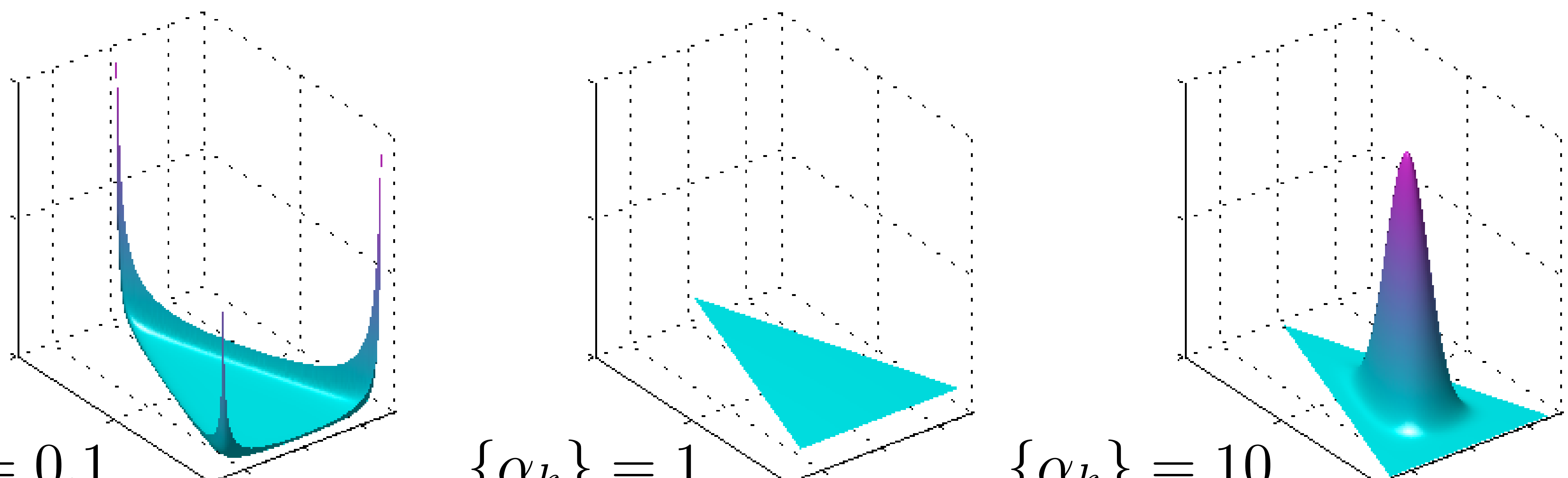
Johann Peter Gustav Lejeune Dirichlet 1805–1859

Multinomial Variables

The Dirichlet distribution (2)



The domain of the Dirichlet distribution, K=3, is the red plane



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

Plots of the Dirichlet distribution (K=3).
The horizontal axes are coordinates in the red plane.

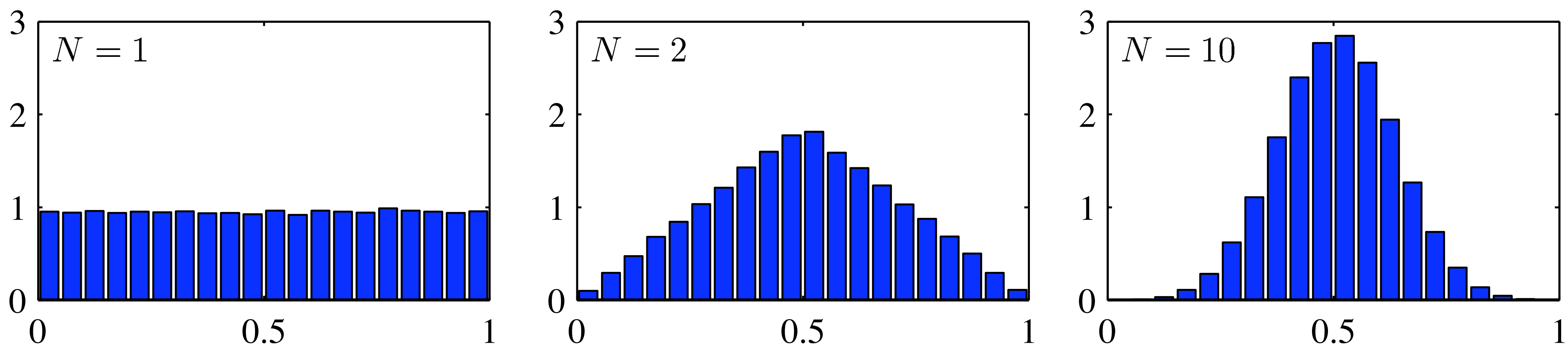
The Gaussian Distribution

Carl Friedrich Gauss
1777–1855

- Also known as the normal distribution
- Can be motivated from a variety of perspectives
  - Maximizes entropy
  - Sum of multiple random variables approaches Gaussian
- $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$

The Gaussian Distribution (2)

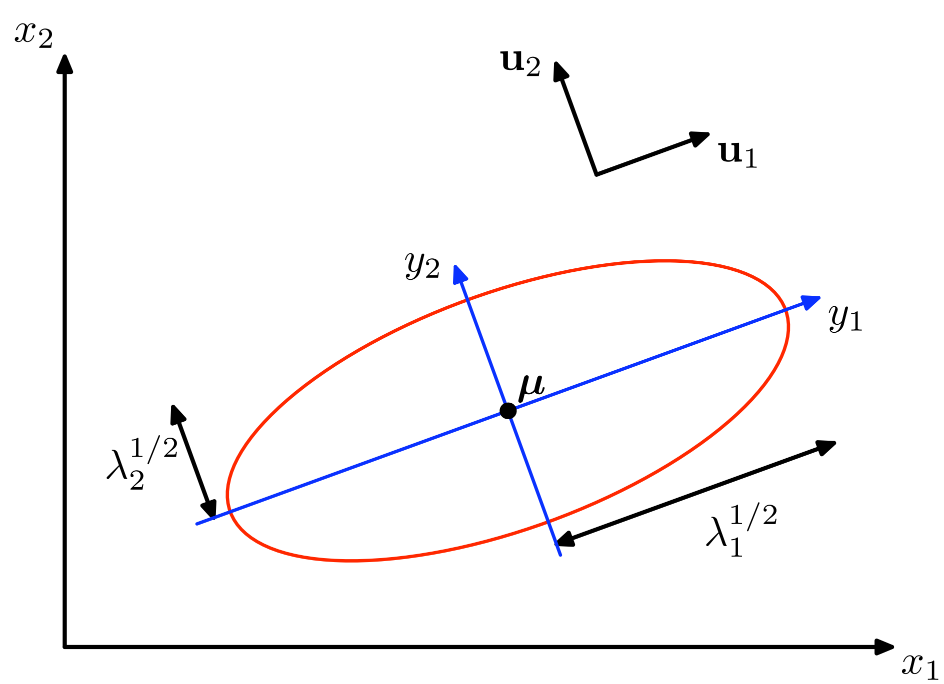
- The *central limit theorem* contains the result about the sum of random variables approaching Gaussian
- The rate of convergence depends on the distributions of the variables



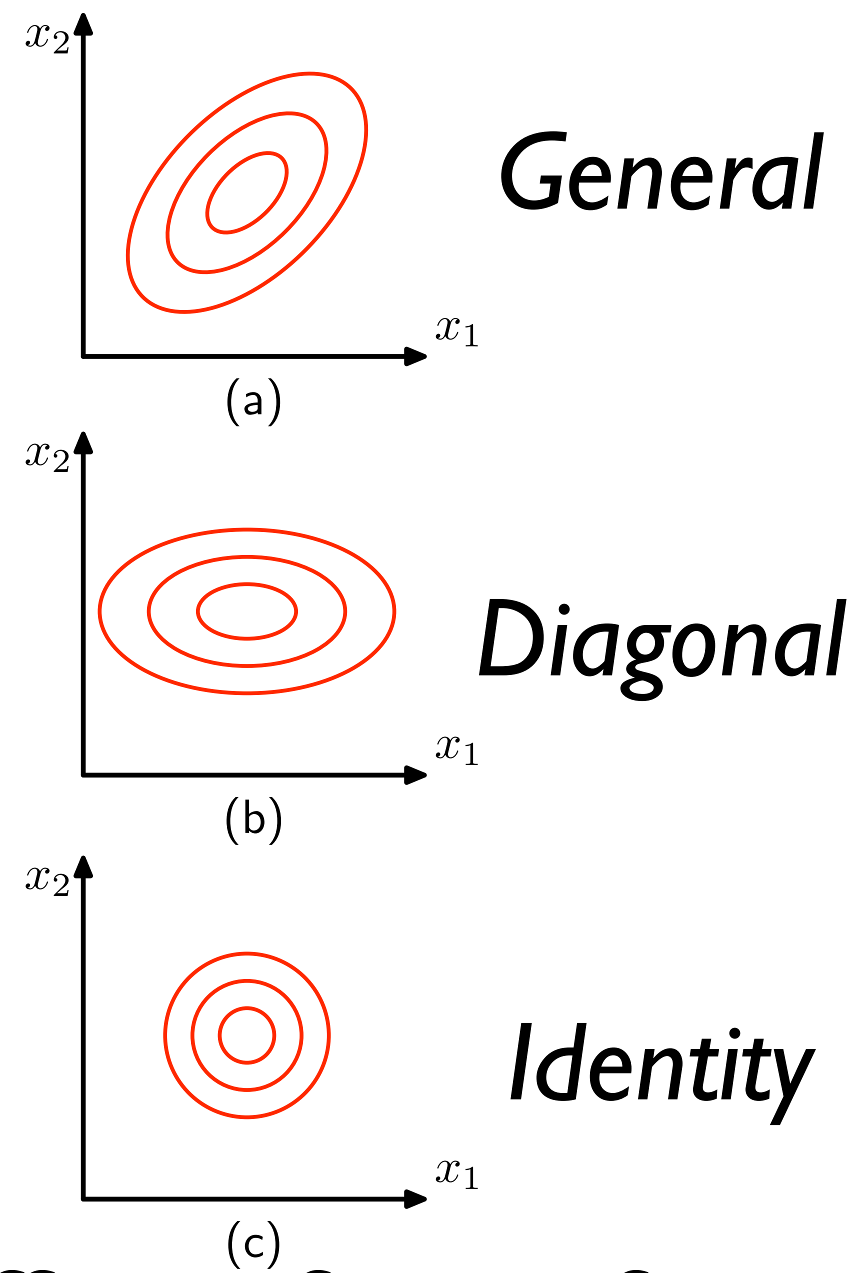
Histogram of the mean of N uniform random variables

The Gaussian Distribution (3)

- The Gaussian density is constant on elliptical surfaces



The major axes of the ellipse are given by the eigenvectors of the covariance matrix



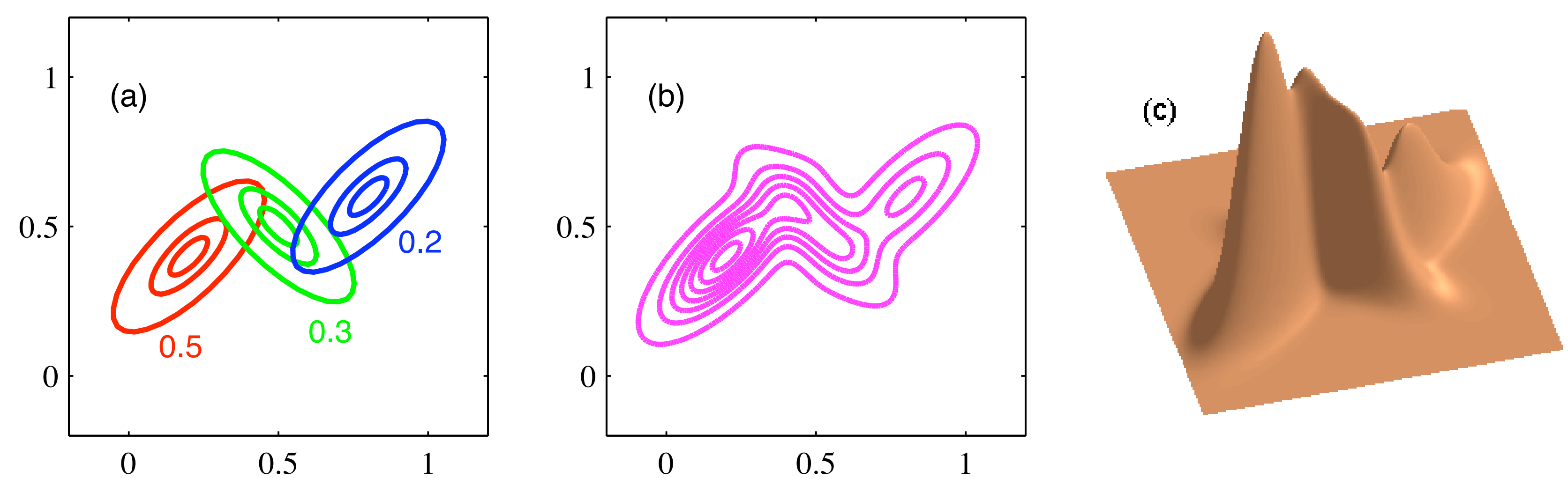
Different forms of covariance matrix

The Gaussian Distribution (4)
Some properties

- Consider two distinct sets of variables, $\mathbf{x}_a$ and $\mathbf{x}_b$, with $p(\mathbf{x}_a, \mathbf{x}_b)$ jointly Gaussian
- Conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian
- Marginal distribution $p(\mathbf{x}_a)$ is Gaussian
- Conjugate priors:
  - Gaussian for the mean
  - *Gamma* for the variance
  - Product of Gaussian and gamma, if both are estimated

The Gaussian Distribution (4)
Mixtures of Gaussians

- Multimodal data can be handled with a mixture of multiple Gaussian distributions
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$
- $0 \leq \pi_k \leq 1$ are the mixing coefficients
- $\sum_{k=1}^{K} \pi_k = 1$



A mixture of Gaussians

The Gaussian Distribution (5)
Themes not covered here...

- Handling of periodic variables
  - The *von Mises* distribution
- *Student's t-distribution* as the posterior distribution of a Gaussian variable, when the precision of the Gaussian has a Gamma prior
- See the book for more information

The Exponential Family

- A broad class of probability distributions
  - Gaussian, Bernoulli, multinomial...
  - Probability is $p(\mathbf{x}|\eta) = h(\mathbf{x})g(\eta)\exp\{\eta^T \mathbf{u}(\mathbf{x})\}$ where η are *natural parameters* and $g(\eta)$ is a normalization constant
- There are general results for the family
  - Sufficient statistics for ML estimators
  - Existence and form of conjugate priors
  - etc.

Noninformative priors

- In Bayesian inference, prior information about the problem is used
- Sometimes, there is little information
- Then, a *noninformative prior* may be used
  - Designed to have as little influence on the posterior as possible
- Example: Gaussian prior with $\sigma_0^2 \to \infty$ for estimating the mean μ of a Gaussian
- *Improper* priors can be used, if posterior is proper

Nonparametric Methods

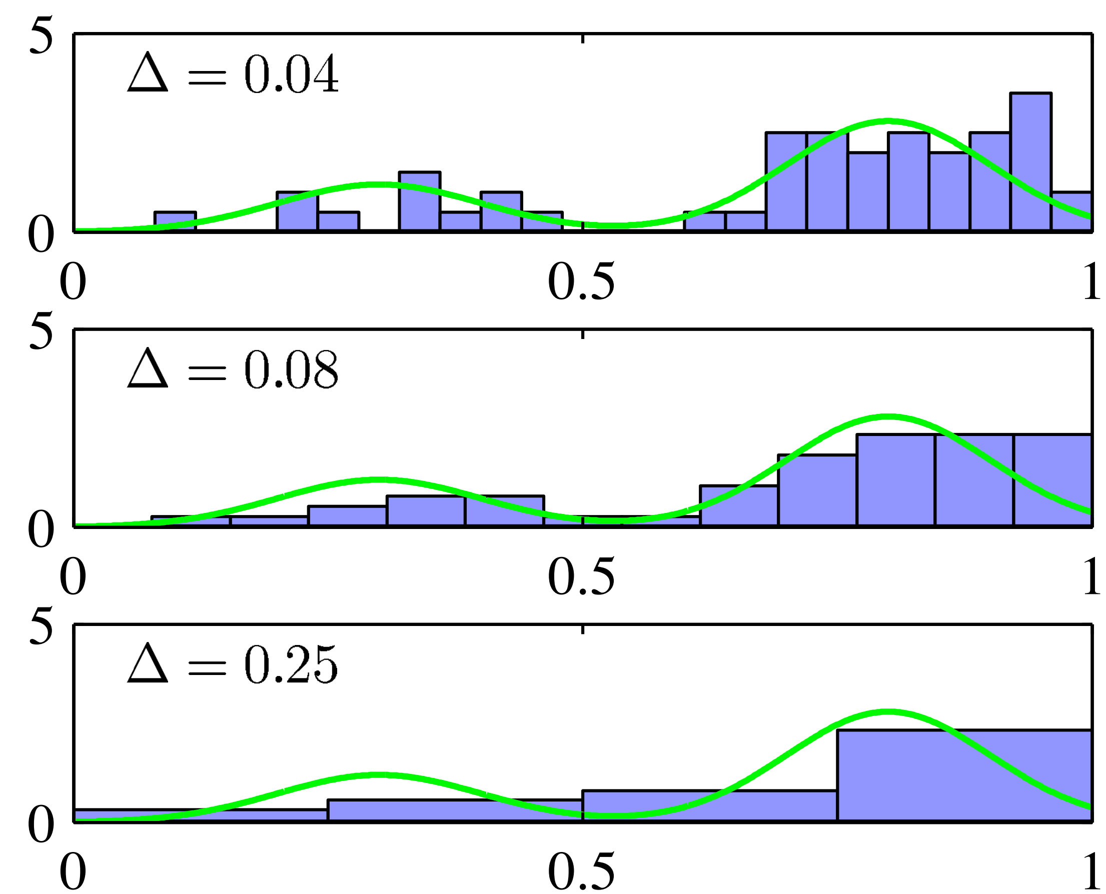
- In the *parametric* methods seen so far, a bad choice of the model may result in poor predictive performance
    - Example: trying to model multimodal data with a Gaussian distribution
- *Nonparametric* methods make fewer assumptions

Nonparametric Methods

Histograms for density estimation

1. Partition the input space into bins of width Δ_i (often $\Delta_i = \Delta$)
2. Count observations in each bin n_i
3. Count probabilities $p_i = \frac{n_i}{N\Delta_i}$

- Example for one dimension
- Can also be used for quick visualization in two dimensions
- Unsuitable for most applications
  - Discontinuities at bin edges
  - Poor scaling with increasing dimensionality

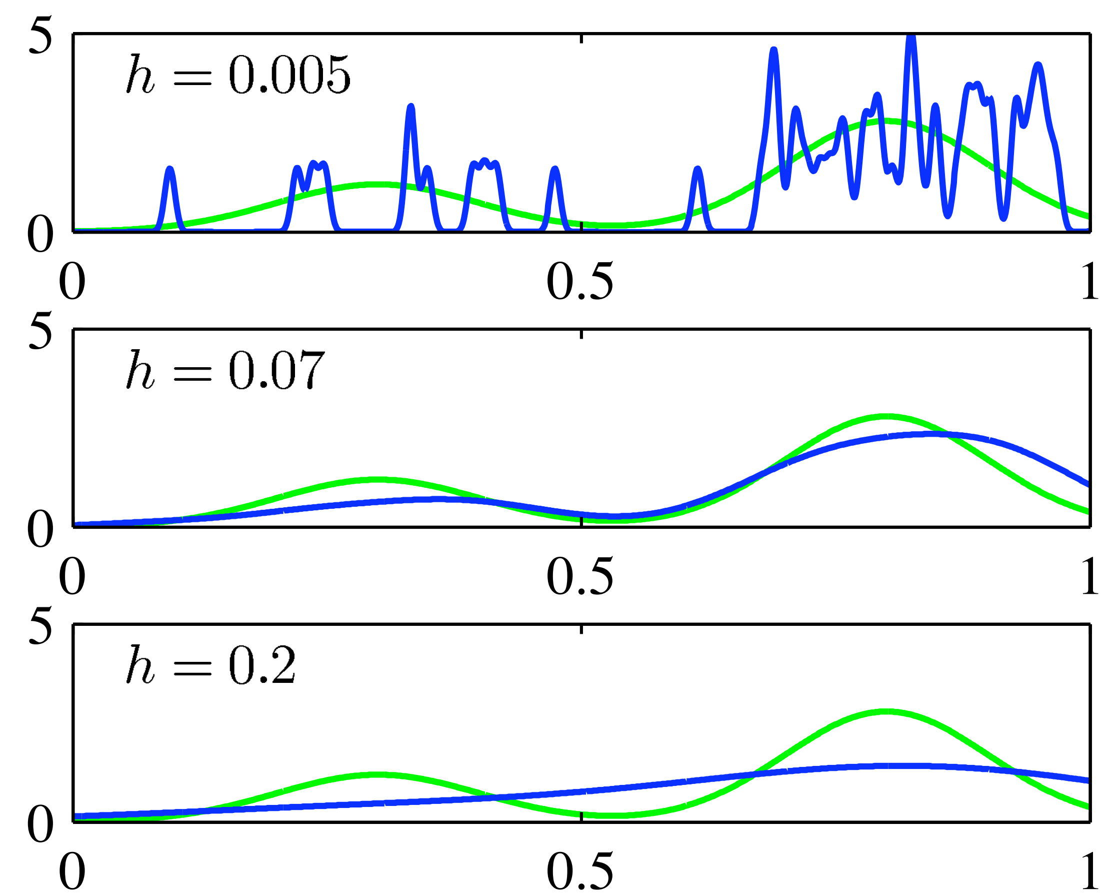


Histograms with different bin widths. Underlying distribution drawn in green.

Nonparametric Methods

Kernel density estimators

- Given N observations, place a probability mass $\frac{1}{N}$ centered on each observation
- Different kinds of kernels can be used as the probability mass
  - Constant in a hypercube
  - (Symmetric) Gaussian
- The smoothness of the model can be adjusted
  - Size of the hypercube
  - Variance of the Gaussian



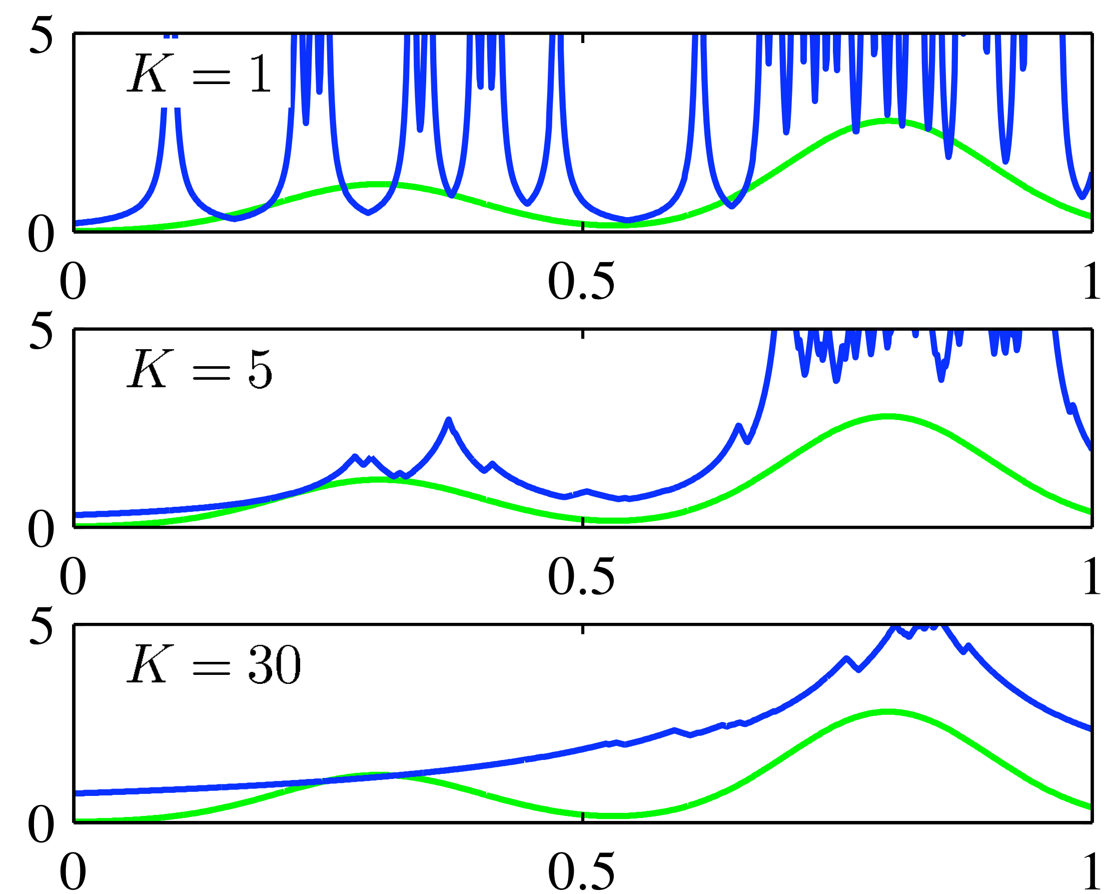
Gaussian kernel density estimation. Underlying distribution drawn in green.

Nonparametric Methods

Nearest-neighbour density estimation

1. Choose a point $\mathbf{x}$ where to estimate the density
2. Grow a sphere centered at $\mathbf{x}$ until it contains K points
3. Density estimate is $p(\mathbf{x}) = \frac{K}{NV}$, V is the volume of the sphere

- Not a true density model (intregral over all space diverges)

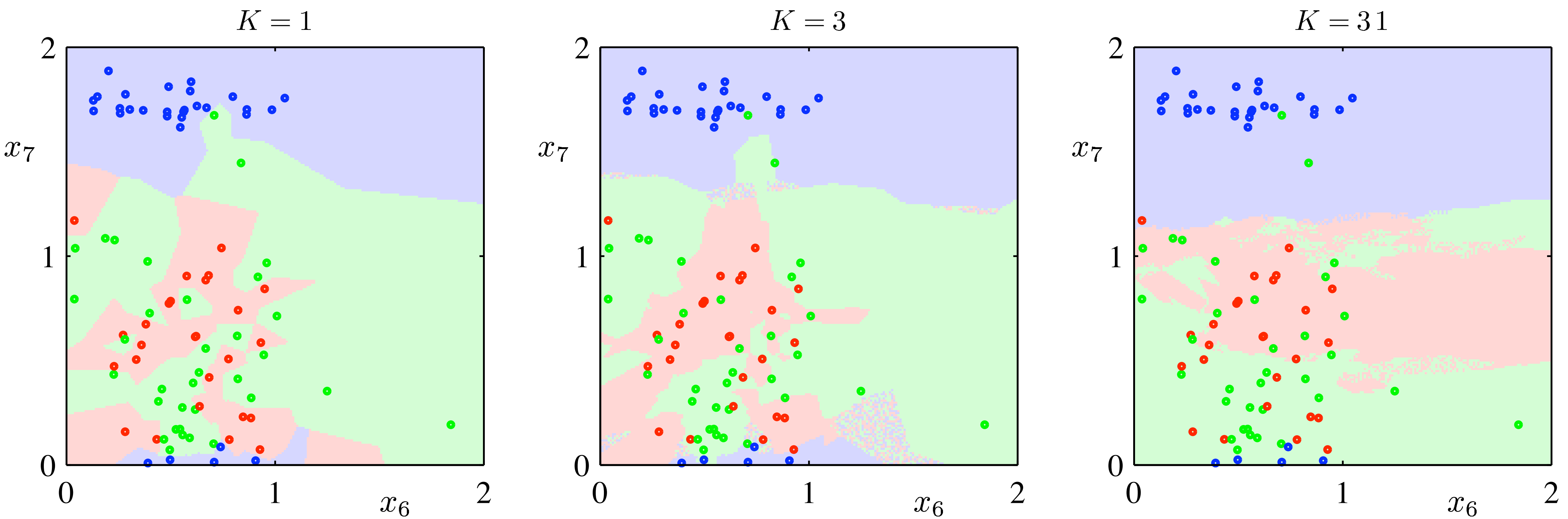


K-nearest-neighbour density estimation. Underlying distribution drawn in green.

Nonparametric Methods

Nearest-neighbour classification

1. Choose a point $\mathbf{x}$ to classify
2. Grow a sphere centered at $\mathbf{x}$ until it contains K points
3. Classify $\mathbf{x}$ according to the majority class in the K points



K-nearest-neighbour classification

Nonparametric Methods
There is a problem

- In both the K-nearest-neighbours method and the kernel density estimator, <u>all training data</u> needs to be stored
  - Heavy computational requirements with a large data set
- Compromise between accuracy and efficiency: tree-based search structures

Summary

- Various probability distributions
- Especially the Gaussian distribution has great practical significance
- Parametric and nonparametric methods
  - Both have advantages and disadvantages
- This chapter of the book is basic knowledge, required later in the book and "in real life" (if your life happens to be research...)