DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Probability Theory and Estimation

Kim Steenstrup Pedersen

Plan for lectures on probability and estimation (this and the next lecture)

- ~~Why *Statistical* Machine learning?~~
- ~~Probability theory 101 (crash course or reminder)~~
- The Gaussian / Normal distribution
- Bayesian probabilities
- Parametric and non-parametric estimation of probability distributions
  - Maximum likelihood and maximum a posteriori estimation
  - Examples of non-parametric methods (more to come later in the course)
- The curse of dimensionality

Parametric estimation of probability distributions

The Gaussian (a.k.a. Normal) Distribution

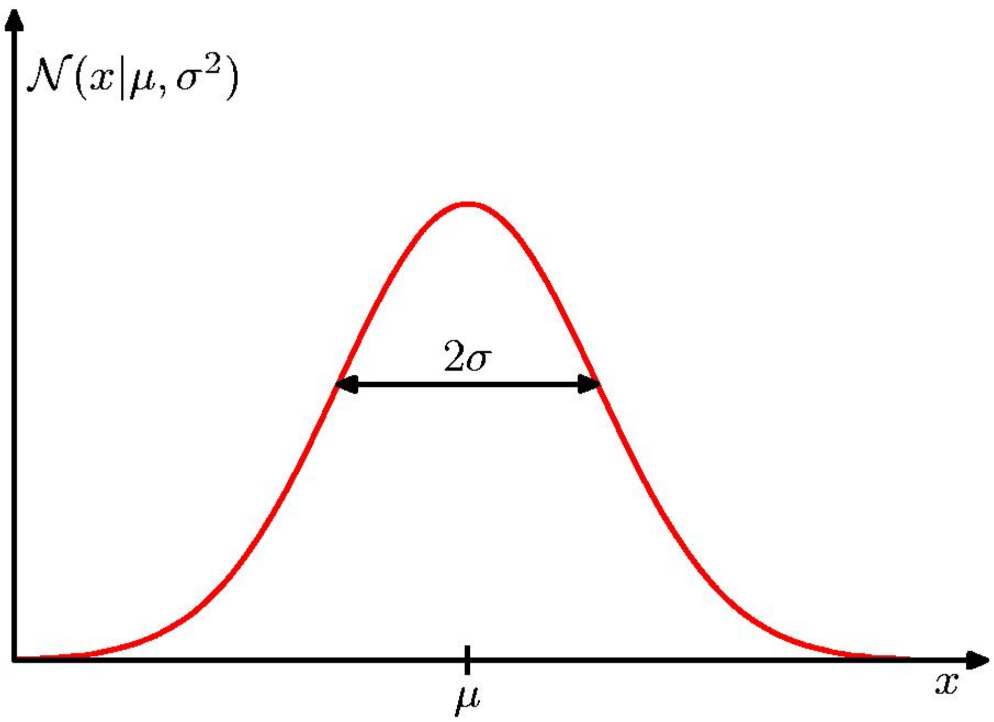
The 1-dimensional Gaussian probability density:

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

Fulfills the density conditions:

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

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

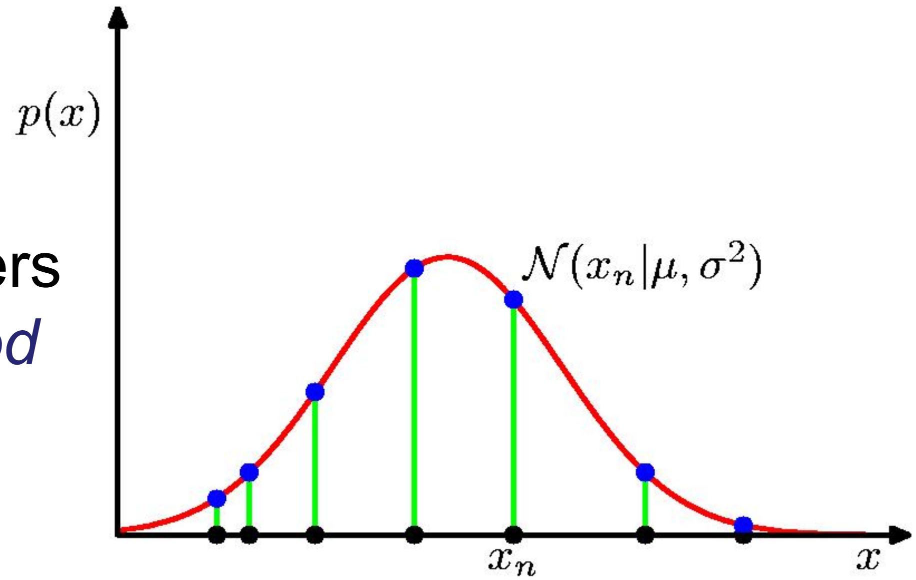


Maximum likelihood estimation for Gaussian

- Assume we have *N* observations: $\mathbf{X} = (x_1, \ldots, x_N)$
- Assume data is drawn independently from the same Gaussian distribution with μ and σ, $\mathcal{N}(x_n \mid \mu, \sigma^2)$.
  i.i.d. = independent and identically distributed
- Lets find that model which maximizes the joint probability density of the observations:

$$p(\mathbf{X} \mid \mu, \sigma^2) = p(x_1, \ldots, x_N \mid \mu, \sigma^2)$$

This function of the parameters μ and σ is called the *likelihood function*. **X** is fixed!



Maximum likelihood estimation for Gaussian

- Due to independence of the observations the *likelihood function* can be written as

$$p(\mathbf{X} \mid \mu, \sigma^2) = p(x_1, \ldots, x_N \mid \mu, \sigma^2) = \prod_{n=1}^{N} p(x_n \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$$

- The maximum likelihood estimates is obtained by

$$(\mu_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}) = \arg\max_{\mu,\sigma^2} p(\mathbf{X} \mid \mu, \sigma^2) \qquad \text{or equivalently}$$

$$(\mu_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}) = \arg\max_{\mu,\sigma^2} \log p(\mathbf{X} \mid \mu, \sigma^2)$$

Maximum likelihood estimates for Gaussian

$$p(\mathbf{X} \mid \mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu,\sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

$$\log p(\mathbf{X} \mid \mu,\sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi)$$

- Maximize with respect to μ

$$\frac{d}{d\mu} \log p(\mathbf{X} \mid \mu,\sigma^2) = +\frac{2}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu) = \frac{1}{\sigma^2} \left(\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} \mu \right) = 0$$

$$\Rightarrow \sum_{n=1}^{N} x_n - N\mu = 0$$

$$\Rightarrow \quad \mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{(Sample mean)}$$

Maximum likelihood estimates for Gaussian

$$p(\mathbf{X} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

$$\log p(\mathbf{X} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi)$$

- Maximize with respect to σ^2

$$\frac{d}{d\sigma^2} \log p(\mathbf{X} \mid \mu_{\mathrm{ML}}, \sigma^2) = +\frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2} \frac{1}{\sigma^2} = 0$$

$$\Rightarrow \frac{N}{2} \frac{1}{\sigma^2} = \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2$$

$$\Rightarrow \quad \sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 \quad \text{(Sample variance)}$$

Maximum likelihood estimates for Gaussian

- Maximize with respect to μ

$$\frac{d}{d\mu} \log p(\mathbf{X} \mid \mu, \sigma^2) = 0 \quad \Rightarrow$$

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{(Sample mean)}$$

- Maximize with respect to σ^2

$$\frac{d}{d\sigma^2} \log p(\mathbf{X} \mid \mu_{\mathrm{ML}}, \sigma^2) = 0 \quad \Rightarrow$$

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

Back to Probability Theory 101

Multivariate distributions

- *D*-dimensional random vector: $\mathbf{x} = (x_1, \ldots, x_D)^T$
- Joint probability density: $p(\mathbf{x}) = p(x_1, \ldots, x_D)$
- Probability of falling in a subset of the domain:

$$p(\mathbf{x} \in \Omega) = \int_{\Omega} p(\mathbf{x}) d\mathbf{x} = \iiint_{\Omega} p(x_1, \ldots, x_D) dx_1 \cdots dx_D$$

- Again these conditions must be fulfilled:

$$p(\mathbf{x}) \geq 0$$

$$\int_{\mathcal{D}} p(\mathbf{x}) d\mathbf{x} = 1$$

Definition of covariance and mean

- Expectation: $E[f(\mathbf{x})] \equiv \iiint f(\mathbf{x}) p(\mathbf{x}) \, dx_1 \cdots dx_D$

- Mean: $E[\mathbf{x}] \in \mathbb{R}^D$

- Multivariate covariance: $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = E_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y}^T - E[\mathbf{y}^T])] = E_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T]$$

$$\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}, \mathbf{x}]$$

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] \in \mathbb{R}^D$$

Multivariate Gaussian distribution

- Multivariate Gaussian density in *D*-dimensions

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

$$\mathbf{x}, \mu \in \mathbb{R}^D \quad , \quad \Sigma \in \mathbb{R}^{D \times D}$$

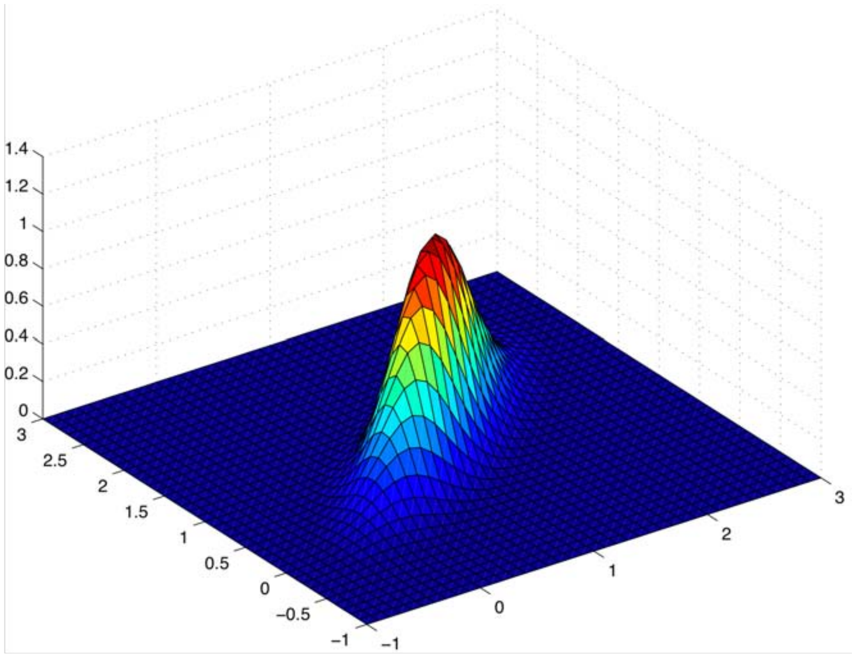
$E[\mathbf{x}] = \mu$: Mean vector

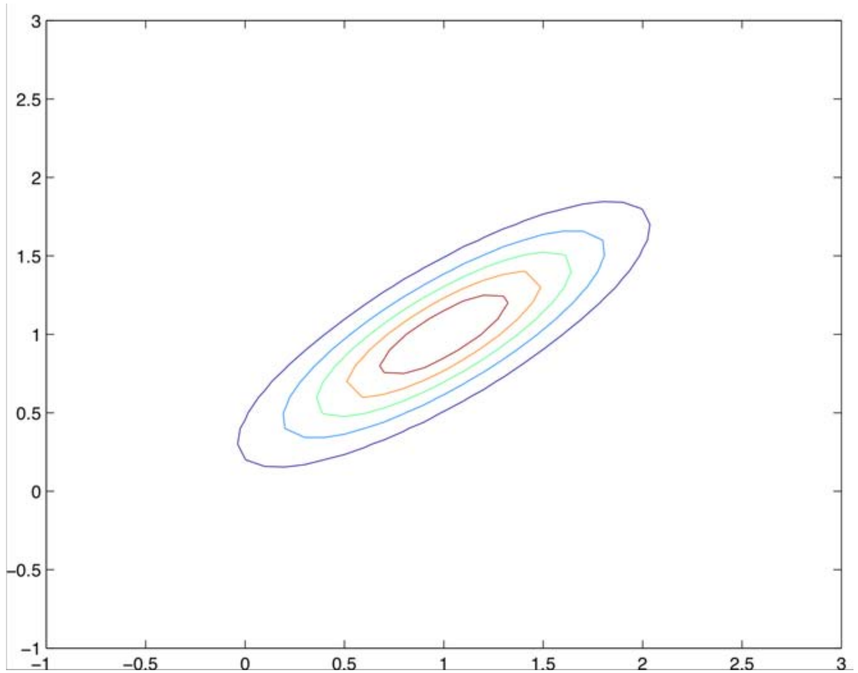
$\mathrm{cov}[\mathbf{x}] = \Sigma$: Covariance matrix

$\Lambda = \Sigma^{-1}$: Precision matrix

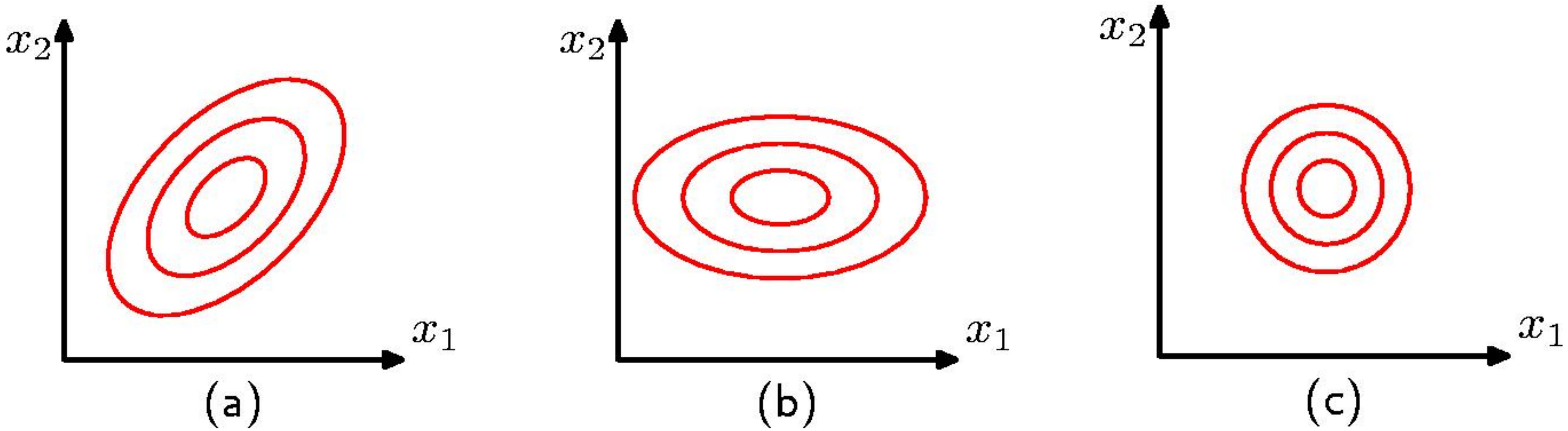
$|\Sigma| = \det(\Sigma)$: Determinant of the covariance matrix

Example: 2D Gaussian distribution





Effect of different covariance on the Gaussian



$$\Sigma \qquad\qquad \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2) \qquad\qquad \Sigma = \sigma^2 \mathbf{I}_D$$

The covariance matrix of the Gaussian

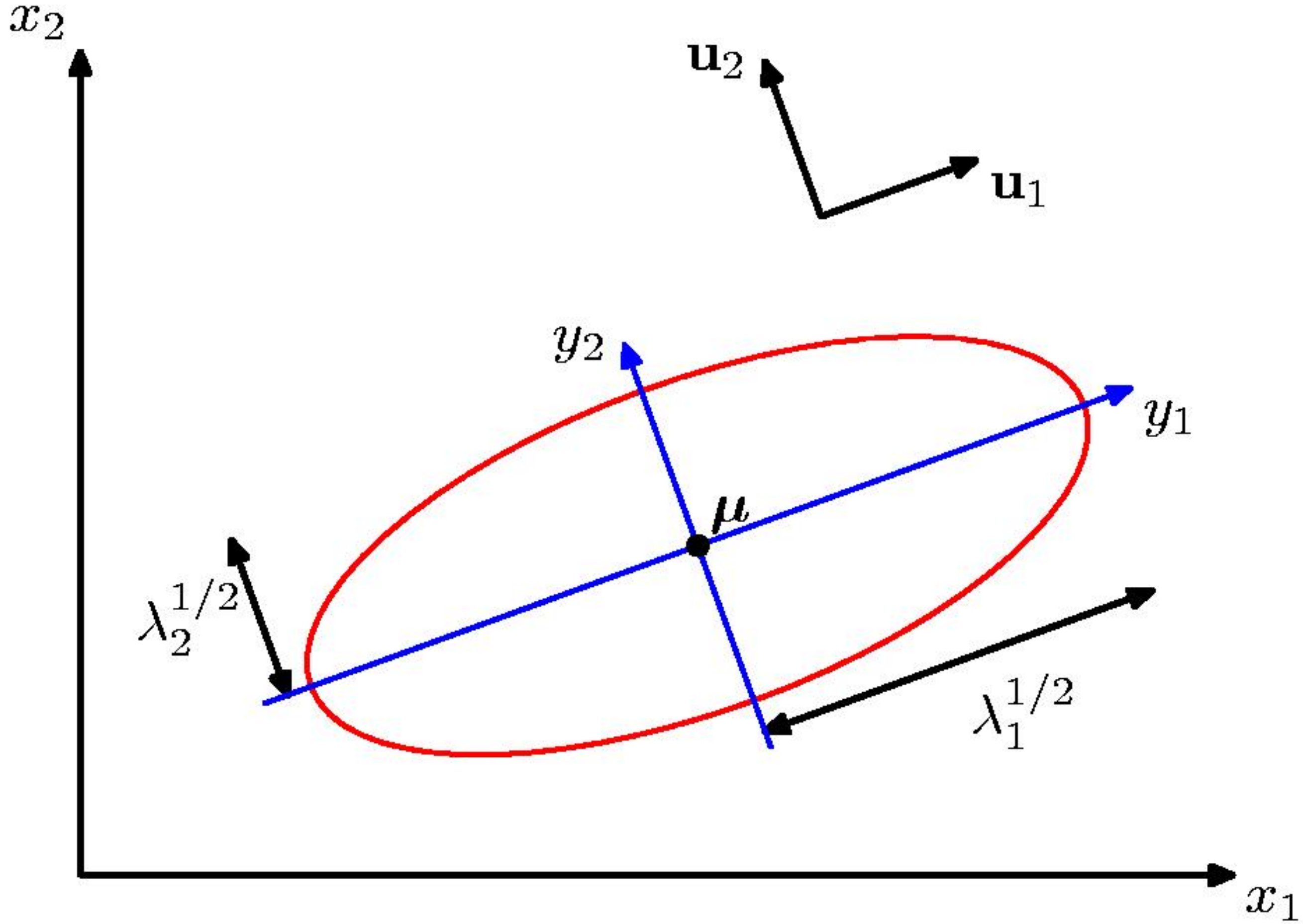
- Mahalanobis distance (the content of the exponential function in the Gaussian)

  $$\Delta^2_{\text{Mah}} = (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)$$

  Distance between **x** and μ weighted by the covariance Σ.
- The covariance matrix must be symmetric and positive definite (all eigenvalues are strictly positive).
- Consider eigenvectors and eigenvalues:

  $$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad \text{(The eigenvalue problem)}$$

Eigenvectors and Eigenvalues



The eigenvalue λ_i is the variance in the direction of the eigenvector $\mathbf{u}_i$

$\sqrt{\lambda_i}$ is the units or scale in the eigenvector coordinate system

Change of variables (whitening)

- Transform into the eigenvector coordinate system:

$$\mathbf{y} = \mathbf{D}^{-1}\mathbf{U}(\mathbf{x} - \mu)$$

where

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_D^T \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_D} \end{pmatrix}$$

- Notice that $\mathrm{cov}[\mathbf{y}] = \mathbf{I}$ and $E[\mathbf{y}] = 0$.
- Aside: This transformation applied to data is called whitening of the data (preprocessing).
- Since **U** is orthogonal we have the inverse transform

$$\mathbf{x} = \mathbf{U}^T\mathbf{D}\mathbf{y} + \mu$$

Maximum likelihood estimation for multivariate Gaussian

$$p(\mathbf{X} \mid \mu, \Sigma) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mid \mu, \Sigma) \quad \text{(Likelihood function)}$$

- Maximizing with respect to μ gives

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \quad \text{(Sample mean vector)}$$

- Maximizing with respect to Σ gives

$$\Sigma_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^T \quad \text{(Sample covariance matrix)}$$

Bayesian probabilities

Bayes' theorem and Bayesian probabilities

- Bayes' theorem

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$

- Bayesian probabilities:
  - Probabilities do not have to be based on the frequencies of outcome of experiments (contrary to frequentists view-point)
  - We may assign probabilities to e.g. parameters and models
  - Probabilities express degree of uncertainty in our knowledge of a variable

Interpretation of Bayes' theorem

Assume we want to learn parameters *w* from a data set *D*.

- Bayes' theorem allows us to update our belief of uncertainty in the model parameters given observations.

$$p(w \mid D) = \frac{p(D \mid w)\, p(w)}{p(D)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

  – Our knowledge prior to the experiment is coded in the prior.
  – After the experiment our uncertainty about w has been updated and is given by the posterior distribution.

Approaches to parameter estimation

- Maximum likelihood (ML) estimation
  Choose *w* that maximizes
  $$p(D \mid w) \qquad \text{(likelihood function)}$$

- Maximum a posteriori (MAP) estimation
  Choose *w* that maximizes
  $$p(w \mid D) \qquad \text{(posterior probability)}$$

- Can you see the difference?

Maximum a posteriori estimation for Gaussian

- Consider the posterior: $p(\mu,\sigma^2 \mid \mathbf{X}) \propto p(\mathbf{X} \mid \mu,\sigma^2)\, p(\mu,\sigma^2)$
  where
  $$p(\mathbf{X} \mid \mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu,\sigma^2) \quad \text{(Likelihood function)}$$
  $$\mathbf{X} = (x_1,\ldots,x_N) \qquad \text{(i.i.d. Observations)}$$

- How to choose the prior $p(\mu,\sigma^2)$?
- A choice: *Conjugated prior*
  - Choose prior with same functional form as the likelihood functions dependence on the parameters.
  - Posterior then has the same functional form as the prior.
  - Can make it possible to find an analytical solution to the estimation problem.

Maximum a posteriori estimation for Gaussian
But only for the mean

- Consider 1-dim. Gaussian with known variance σ²

$$p(\mathbf{X} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right]$$

- The Likelihood function is exponential quadratic in μ, hence the conjugated prior is a Gaussian distribution in μ:

$$p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2) \qquad \text{(Prior)}$$

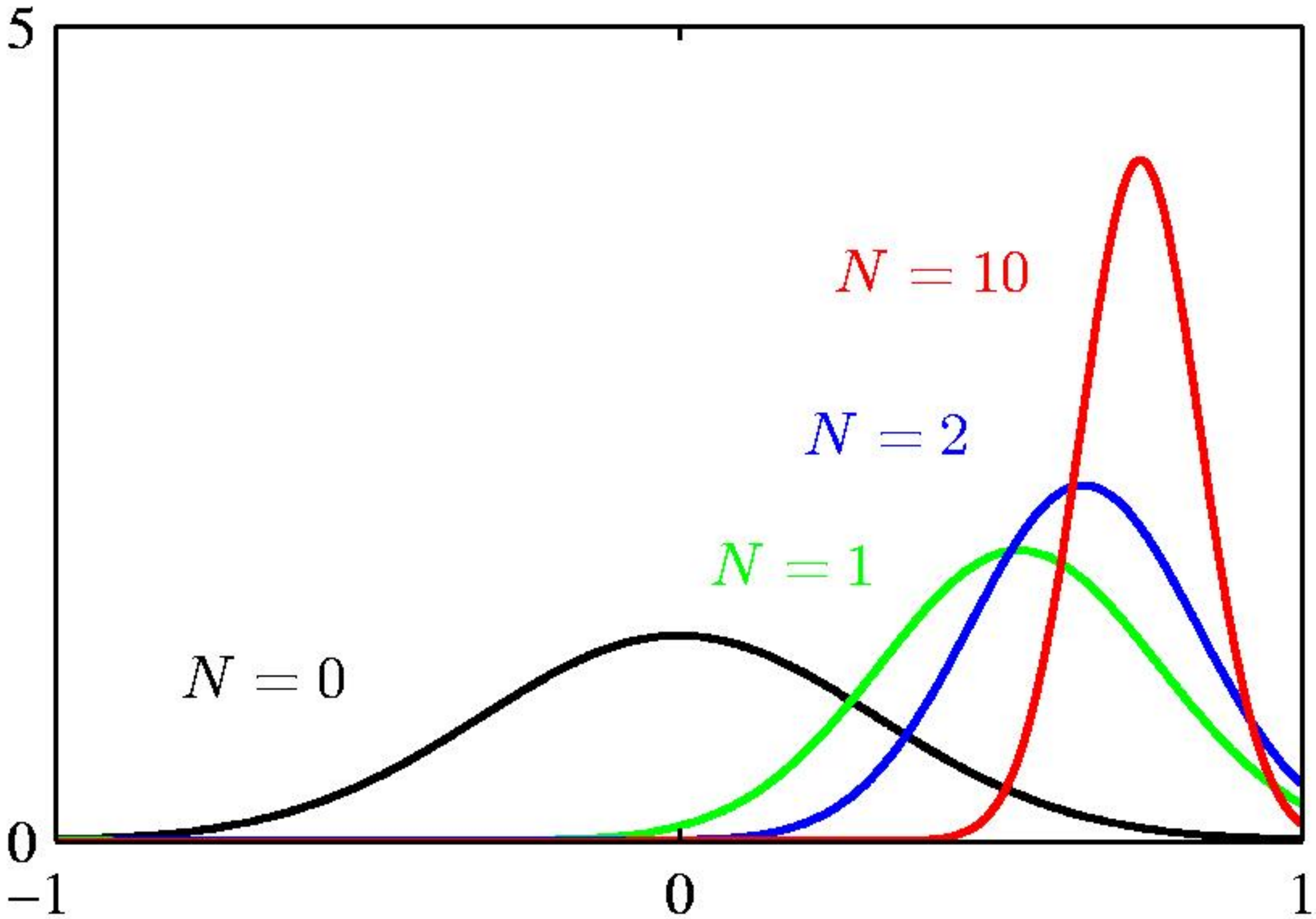
$$p(\mu \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \mu)\, p(\mu)}{p(\mathbf{X})} = \mathcal{N}(\mu \mid \mu_N, \sigma_N^2) \qquad \text{(Posterior)}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} \qquad \text{(Posterior mean)}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \qquad \text{(Posterior precision)}$$

$$\mu_{\mathrm{MAP}} = \arg\max_{\mu}\; p(\mu \mid \mathbf{X}) = \mu_N \qquad \text{(MAP estimate)}$$

Bayesian (Sequential) Inference

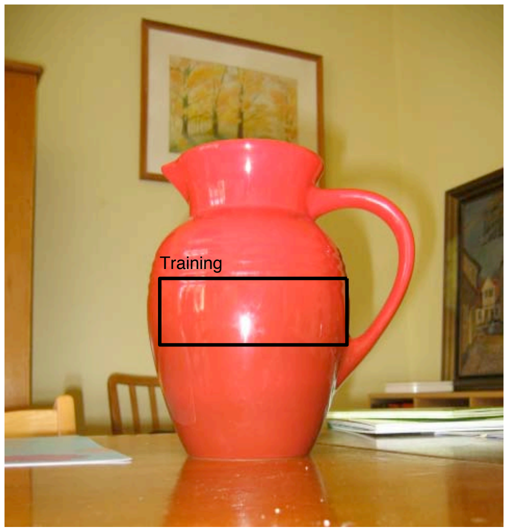


$$p(\mu \mid \mathbf{X}) \propto \left[p(\mu) \prod_{n=1}^{N-1} p(x_n \mid \mu) \right] p(x_N \mid \mu)$$

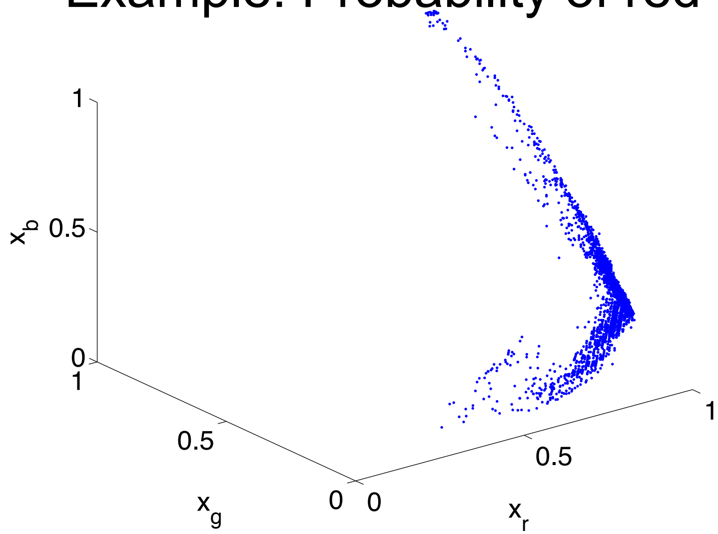
Non-Parametric Density Estimation

Non-Parametric Probability Density Estimation

- If it is not possible or necessary to model a probability distribution parametrically (with a function), we may use non-parametric density estimation methods such as:
  - Histogram estimator
  - Kernel density estimator
  - K Nearest Neighbor (KNN) density estimator



Example: Probability of red



Non-Parametric Density Estimation: Histogram

- A histogram *H(X)* of the random variable *X* is a table of frequency counts of *N* experiments (or data points):
  1. Subdivide the domain of *X*, e.g. the set of real numbers, into *M* bins of width Δ (bin volume in D-dim.).
  2. For the i'th bin, let H_i be the frequency count of how many times *X* falls into the bin.

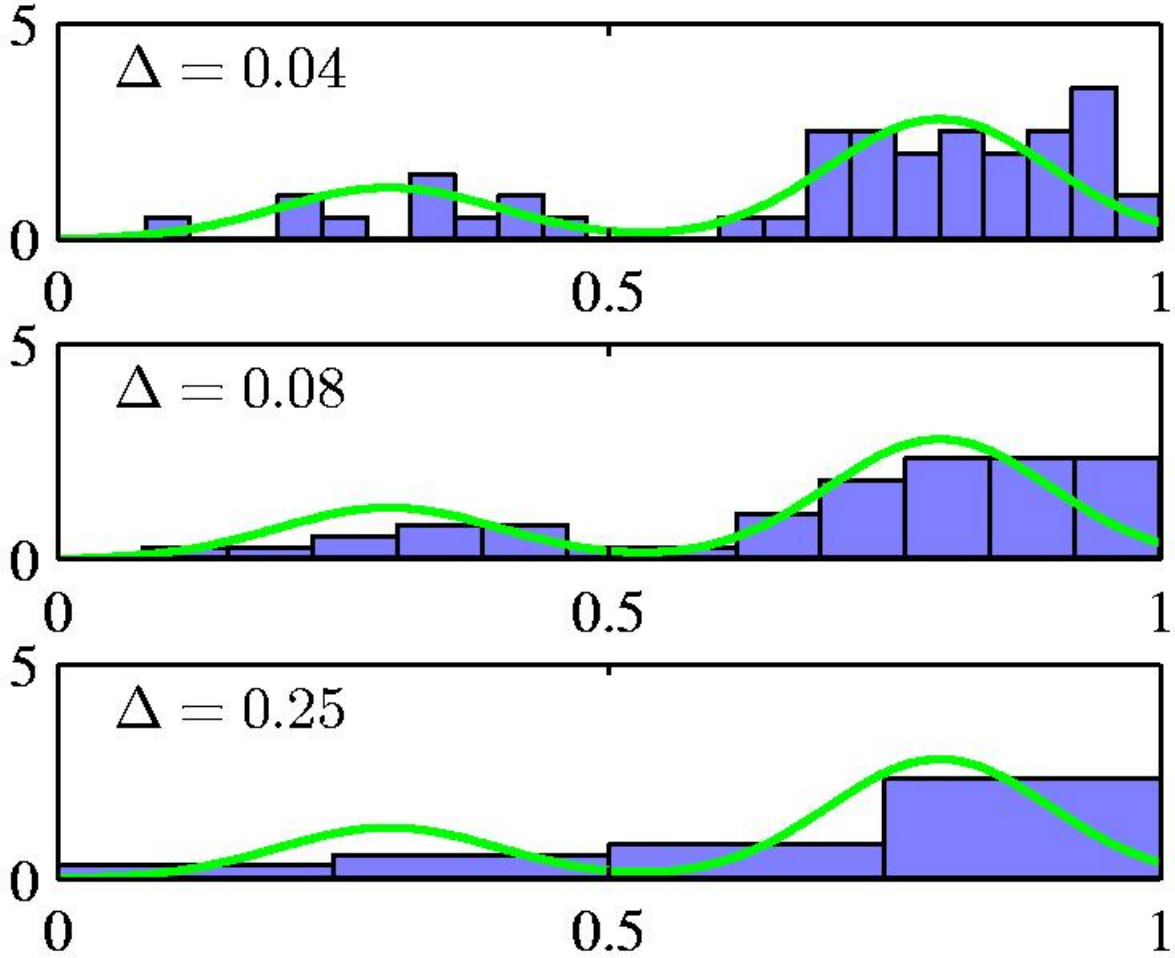
- Probability estimate: Probability of falling in the i'th bin

$$p(X \in \Delta_i) = \frac{H_i}{N} \qquad \text{(Probability estimator)}$$

- Probability density estimate:

$$p(x) = \frac{H_i}{N\Delta} \qquad \text{(Probability density estimator)}$$

Non-Parametric Density Estimation: Histogram



The bin width Δ controls the quality of the estimate.

Non-Parametric Density Estimation: Kernels

- Extending the idea of histograms to estimates around arbitrary points **x**.
  1. Count the number of points around **x** using a kernel function centered on **x** (kernel = bin) $K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$

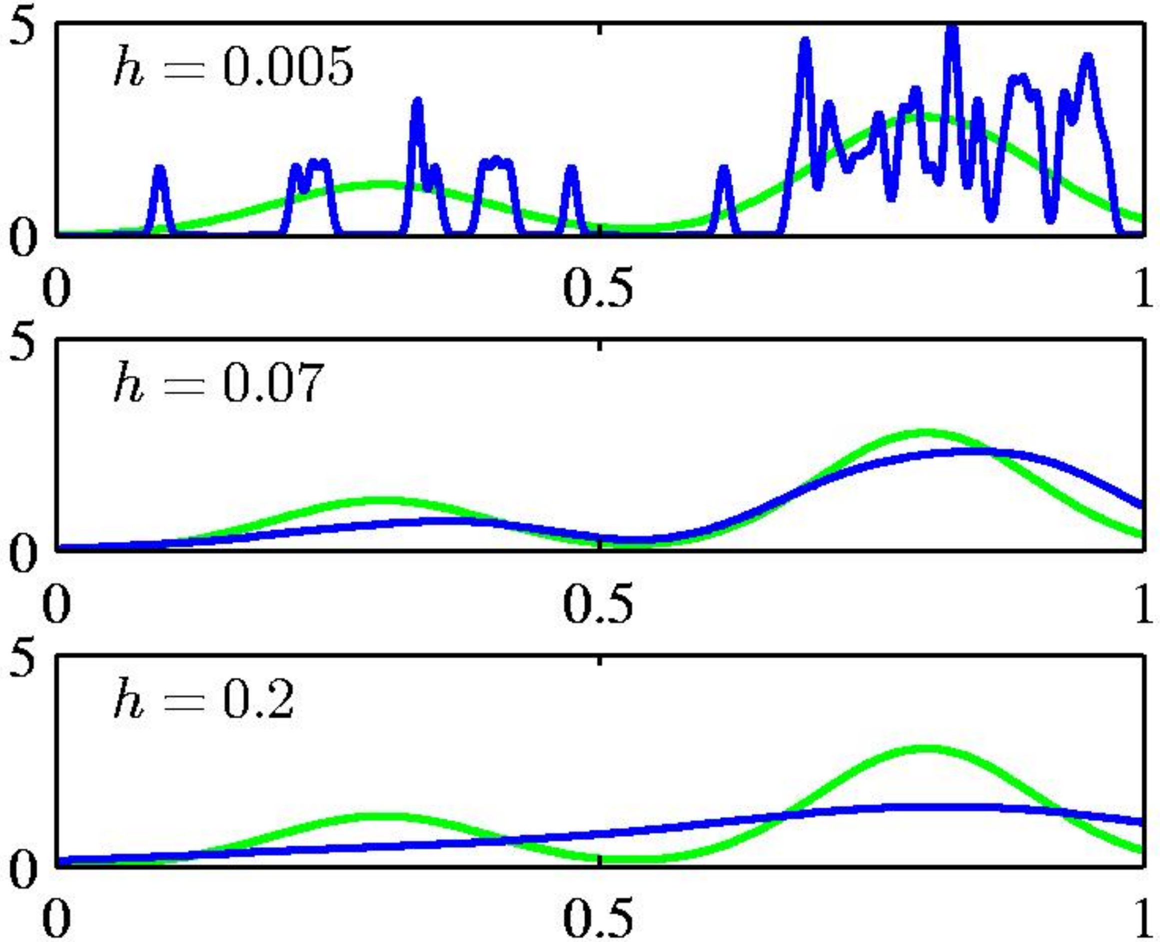
     Equivalently, put a kernel centered on each data point and sum the values of the kernel functions at **x**.
  1. The volume of the bin defined by the kernel is $V = h^D$
  2. Probability density kernel estimate

$$p(\mathbf{x}) = \frac{K}{NV} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right),\ k(\mathbf{u}) = \begin{cases}1\ , & |u_i| \le 1/2\ ,\ i = 1,\ldots,D \\ 0\end{cases} \qquad \text{(Parzen window)}$$

- Alternative kernel function: Gaussian kernel

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

Non-Parametric Density Estimation: Kernels

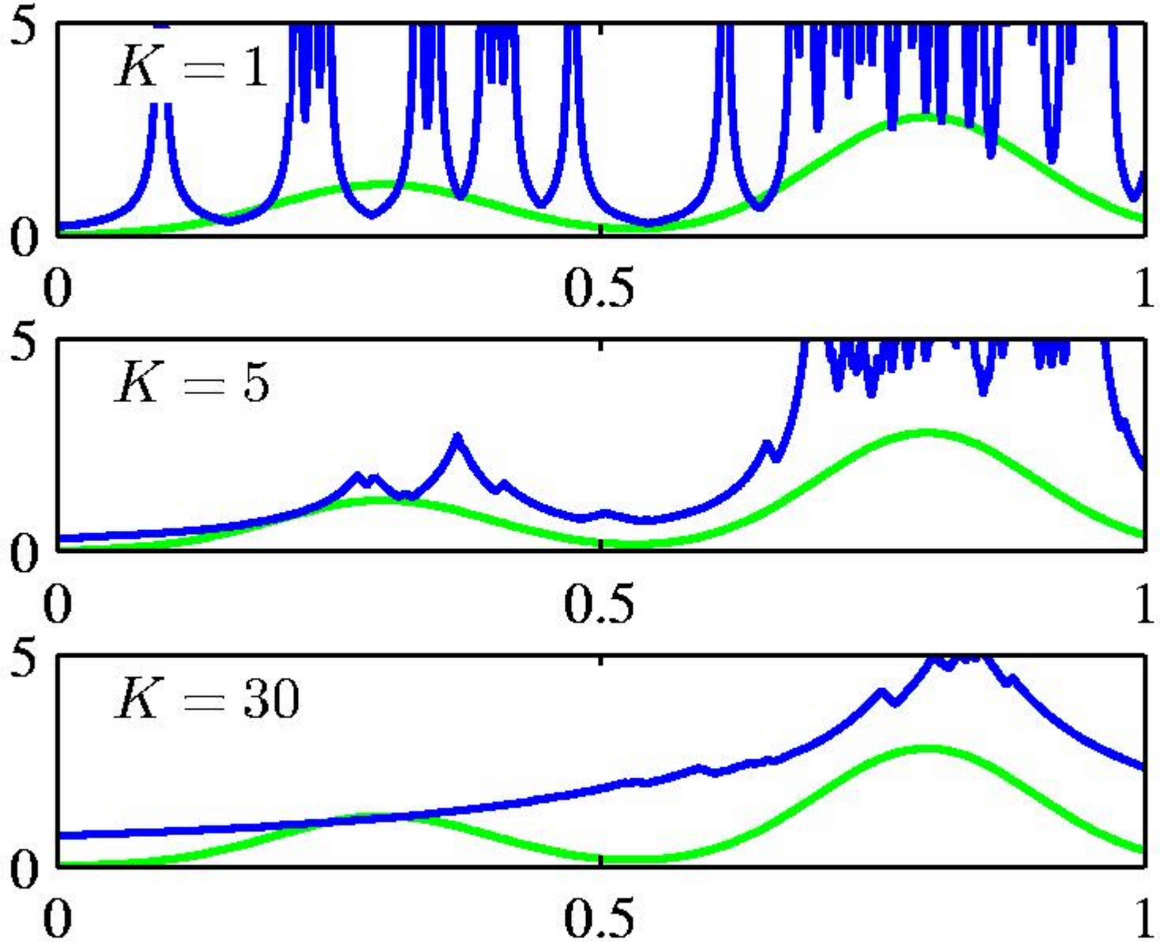


Non-Parametric Density Estimation: K-NN

- In the K nearest neighbor (KNN) estimator, the bin frequency is fixed to K and the bin volume varies.
    1. Expand a sphere centered at **x** until it encompass K data points (neighbors of **x**).
    2. Compute the sphere volume $V(\mathbf{x})$.
    3. Probability density estimate:

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

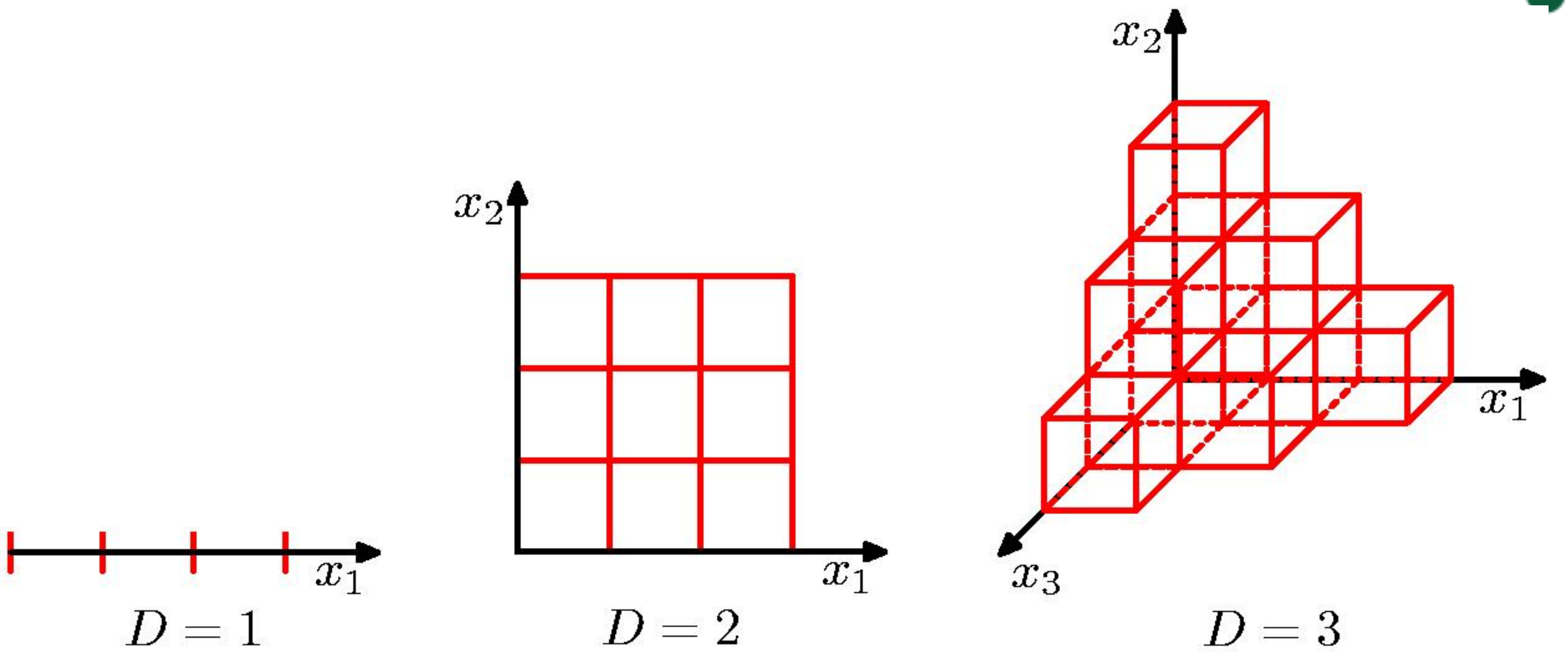
Non-Parametric Density Estimation: K-NN



Non-Parametric Density Estimation: A comparison

- Histograms:
  - Easy to implement, running time is proportional to number of data points N and can be used for both off-line and online processing.
  - When dimensionality grows $H(x_1, \ldots, x_D)$ the number of bins grows as M^D, and amount of memory needed does the same.
  - Only tractable for small dimensionality D.
  - How to choose number of bins M (model order) and bin width Δ?

Curse of Dimensionality



Histograms: As dimensionality D of data space grows, the amount of bins grows exponentially M^D (fixed size bins). Hence amount of data points N has to grow exponentially to keep the same estimation error.

The curse of dimensionality haunts all machine learning methods!

Non-Parametric Density Estimation: A comparison

- Histograms:
  - Easy to implement, running time is proportional to number of data points *N* and can be used for both off-line and online processing.
  - When dimensionality grows $H(x_1,\ldots, x_D)$ the number of bins grows as M^D, and amount of memory needed does the same.
  - Only tractable for small dimensionality D.
  - How to choose number of bins *M* (model order) and bin width Δ?
- Kernels:
  - Allows estimates at arbitrary **x** and (smooth) bins defined by kernel function.
  - Provides poor estimates in low density areas (too few samples to get good estimates).
  - Requires access to all data (to find points under the kernel), hence only allows batch / off-line processing.

Non-Parametric Density Estimation: A comparison

- KNN:
  - Allows estimates at arbitrary **x**.
  - Adapts the volume to improve estimates in low density areas at the cost of smoothing.
  - Requires access to all data, hence only allows batch / off-line processing.
  - Worse case: Requires a search in N data points for the K nearest neighbors of **x**.
- Histograms and kernel methods estimate the frequency of the bin / under the kernel:

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

- KNN estimates the volume of the bin with fixed frequency: $p(\mathbf{x}) = \dfrac{K}{NV(\mathbf{x})}$

Summary

- Multivariate Gaussian distribution
- Maximum likelihood maximum a posteriori parameter estimation
- Non-parametric probability density estimations

Literature

- Probability theory: Sec. 1.2, 1.4
- Gaussian distribution and ML, MAP, and non-parametric estimation: Sec. 2.3, 2.5