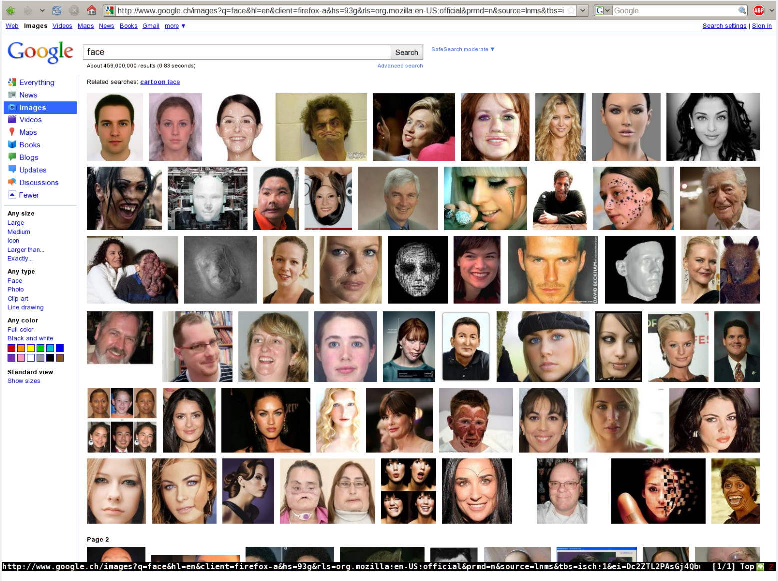
Pattern Recognition: Probability Theory

Sandro Schönborn

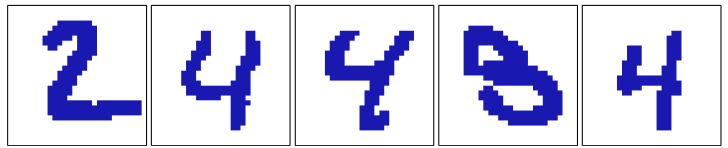
Department of Mathematics and Computer Science
University of Basel

Autumn Semester 2012

Variability

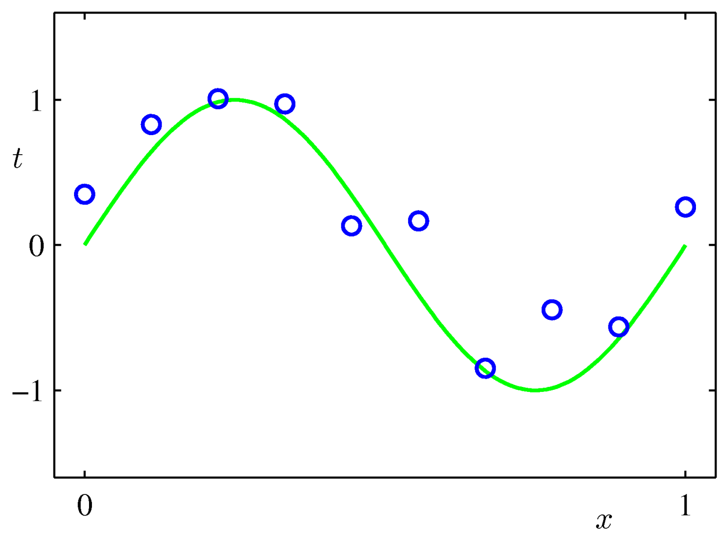


Variability



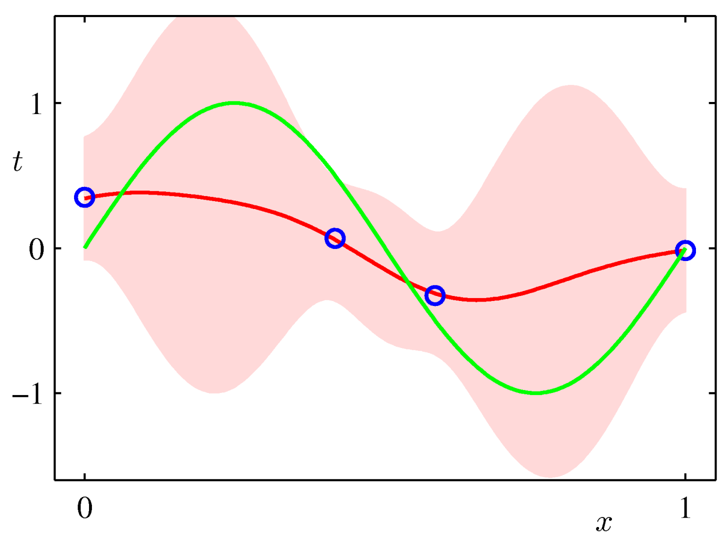
Bishop 2009

Noise



Bishop 2009

Uncertainty



Bishop 2009

Motivation

Why do we need probability theory??

Probability and Statistics

To model

- Variability of pattern itself
- Variability of measurement / context (noise)
- Uncertainty in our models and methods

$\Rightarrow$ A short repetition of probability theory

- First Part: Dry theory $\rightarrow$ quick reference for you
- Second Part: Multivariate Gaussian serving as example

Discrete Random Variables

Random Variable X with possible Realisations $x \in \{1, 2, 3, \ldots\}$:

Cummulative Distribution Function (cdf)

$$P[X < x] = F(x)$$

Probability Mass Function

$$P[X = x] = P_x$$

Normalisation and Positivity

$$\sum_x P_x = 1 \qquad P_x \geq 0$$

Discrete Random Variables — Examples

Binomial – A coin flip

$$x \in \{0, 1\}$$

$$P_0 = P[X = 0] = p, \qquad P_1 = P[X = 1] = q$$

$$p \in [0, 1], \quad q = 1 - p$$

Poisson – Rare events

$$x \in \{0, 1, 2, \ldots\}$$

$$P_x = P[X = x] = \frac{\lambda^x e^{-\lambda}}{x!}$$

$\lambda > 0$: Rate of events occuring per interval

Continouos Random Variables

Random Variable X with possible Realisations $x \in \mathbb{R}$:

Cummulative Distribution function (cdf)

$$F(x): \qquad P[X < x] = F(x)$$

Probability Density Function (pdf)

$$p(x): \qquad P[x < X < x + \mathrm{d}x] = p(x)\,\mathrm{d}x \qquad = \mathrm{d}F(x)$$

Normalisation and Positivity

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \qquad p(x) \geq 0$$

Continuous Random Variables — Examples

Gaussian

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad x \in \mathbb{R}$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean μ, Variance σ^2 ▸ Examples

Gamma Distribution

$$X \sim \Gamma(k, \theta), \quad x \in [0, \infty)$$

$$p(x) = x^{k-1} \frac{e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k}$$

Shape $k > 0$, Scale $\theta > 0$

Mean

- The mean is a measure for *central tendency*

Expected Value, Mean, Expectation

$$E[X] = \sum_x x P_x \qquad E[X] = \int x p(x)\,\mathrm{d}x$$

Linearity

$$E[aX + bY] = a\ E[X] + b\ E[Y]$$

a, b Real *constants*,
X, Y Random variables (same space)

Variance

- The variance is a measure for *spread*

Variance / Standard Deviation

$$V[X] = E[(X - E[X])^2]$$
$$\mathrm{sd}[X] = \sigma_X = \sqrt{V[X]}$$

Hint: $V[X] = E[X^2] - E[X]^2$

Properties

$$V[aX + bY] = a^2 V[X] + b^2 V[Y] + 2ab\ \mathrm{Cov}(X, Y)$$

Mean and Variance — Examples

Binomial

$$E[X] = q$$

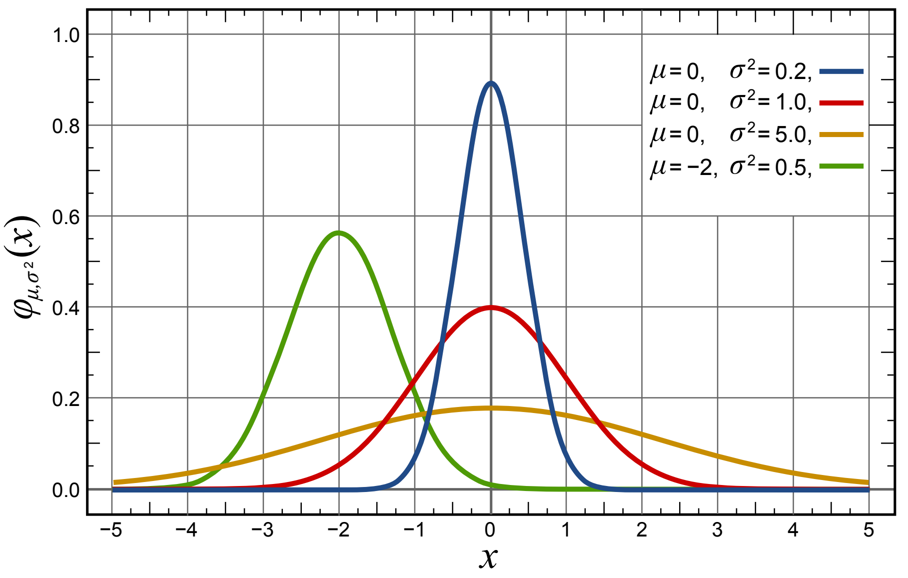
$$V[X] = q(1-q) = p(1-p)$$

Gaussian

$$E[X] = \mu$$

$$V[X] = \sigma^2$$

Example: Gaussian



Example

More than one Random Variable, e.g.

Length L and Weight W of a fish

$$\vec{X} = [L, W]^{\mathsf{T}}$$

Joint Probability

$$P[X = x \ \wedge \ Y = y] = P_{xy}$$

$$P[x < X < x + \mathrm{d}x \ \wedge \ y < Y < y + \mathrm{d}y] = p(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Marginals and Conditionals

Marginalisation

$$P[X = x] = \sum_y P[X = x, Y = y]$$

$$p(x) = \int p(x, y) \, \mathrm{d}y$$

Conditional Probability

$$P[X = x \mid Y = y] = \frac{P[X = x, Y = y]}{P[Y = y]} \qquad P[Y = y] > 0$$

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

Bayes' Rules

Use the factorization for the joint probability density / distribution:

$$p(x, y) = p(x \mid y)\; p(y)$$

$$p(x, y) = p(y \mid x)\; p(x)$$

Bayes' Rule

$$P_{x|y} = \frac{P_{y|x} P_x}{P_y}$$

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

- *Bayesian talk*: "Prior adapted to data leads to posterior"

Covariance and Independence

Covariance

$$\mathrm{Cov}(X,Y) = E[(X - E[X])\ (Y - E[Y])]$$
$$\mathbf{\Sigma}(\mathbf{X}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^{\mathsf{T}}]$$

Independence

$$p(x,y) = p(x)p(y) \iff X \text{ and } Y \text{ are independent}$$

Covariance $\neq$ Independence

$$X \text{ and } Y \text{ are independent, } X \perp Y \implies \mathrm{Cov}(X,Y) = 0$$

Multivariate Gaussian Distribution

- This distribution occurs very frequently
  - Central Limit Theorem
  - Maximum Entropy Principle
  - Ease of use
- Simple enough to demonstrate these concepts

Multivariate Gaussian Distribution

$$p\left(\vec{x}\right) = \frac{1}{\sqrt{(2\pi)^d \det\left(\mathbf{\Sigma}\right)}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^{\mathsf{T}} \mathbf{\Sigma}^{-1} \left(\vec{x}-\vec{\mu}\right)\right)$$

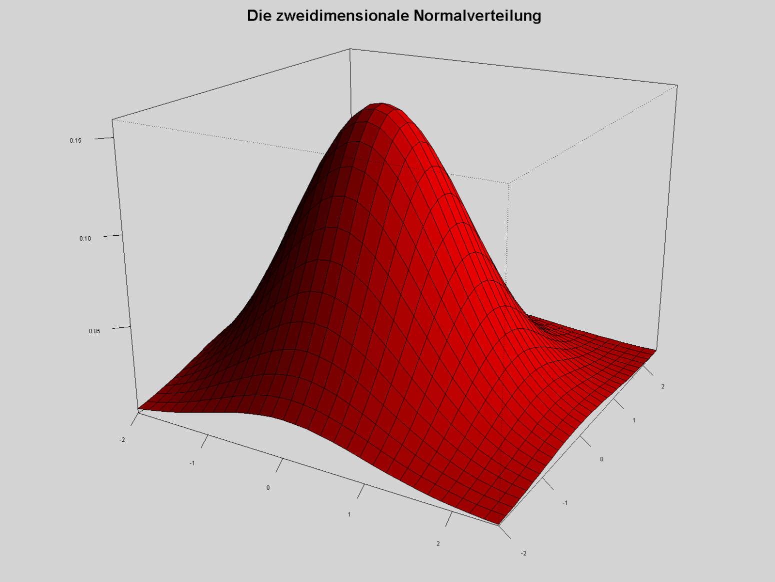
$\vec{\mu}$ Mean

$\mathbf{\Sigma}$ Covariance Matrix ($d \times d$, positive definite, symmetric)

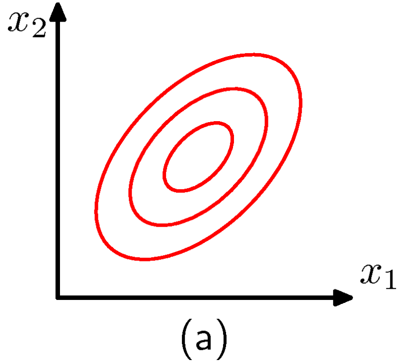
d Number of dimensions

$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$$

2D Gaussian — Surface Plot



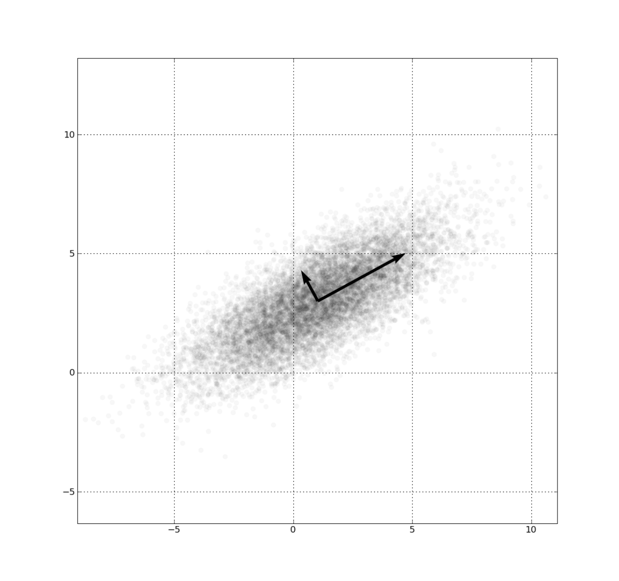
2D Gaussian — Contour Plot



- Points on a contour have equal probability density - *equidensity* lines
- Contours are ellipsoids

Figure: Bishop 2009

2D Gaussian — Samples / Scatter



Equidensity lines are Ellipsoids

- The ellipsoids are determined by the quadratic form

$$(\vec{x} - \vec{\mu})^{\mathsf{T}} \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu})$$

- $\mathbf{\Sigma}$ is positive definite and symmetric $\Rightarrow$ Ellipsoid
- Center at $\vec{\mu}$
- Eigenvectors and eigenvalues of Σ

$$\mathbf{\Sigma}\vec{e}_i = \lambda_i \vec{e}_i$$

- Direction of semi-axes is determined by eigenvectors $\vec{e}_i$
- λ_i measures the variance along the corresponding eigendirection $\vec{e}_i$

Moments

Mean

$$E[\vec{X}] = \vec{\mu} \qquad E[X_i] = \mu_i$$

Covariance

$$V[\vec{X}] = \mathbf{\Sigma} \qquad \mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$$

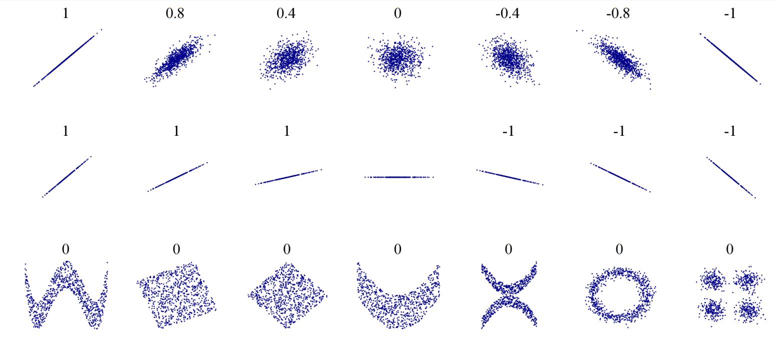
Correlation

$$\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i \sigma_j} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, \qquad \sigma_i = \sqrt{\Sigma_{ii}}$$

Correlation and Covariance

- Correlation measures strength of linear relations between variables
- It does not measure independence
- It does not tell you anything about causal relations
- Correlation is normalized and dimensionless

Example



Marginals

- Marginal: *Randverteilung*
- Removing unknown variables — *"projection"*
- $p(x) = \int p(x,y)dy$

Marginal of a Gaussian

$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$$

$$\vec{X} = \begin{bmatrix} \vec{X}_a \\ \vec{X}_b \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{bmatrix}$$

$$p(\vec{x}_a) = \mathcal{N}(\vec{x}_a \mid \vec{\mu}_a, \mathbf{\Sigma}_{aa})$$

Conditionals

- Conditional: *Bedingte Verteilung*
- Fixing a variable to a certain value — *"slices"*
- $p(x \mid y) = \dfrac{p(x, y)}{p(y)}$

Conditional of a Gaussian

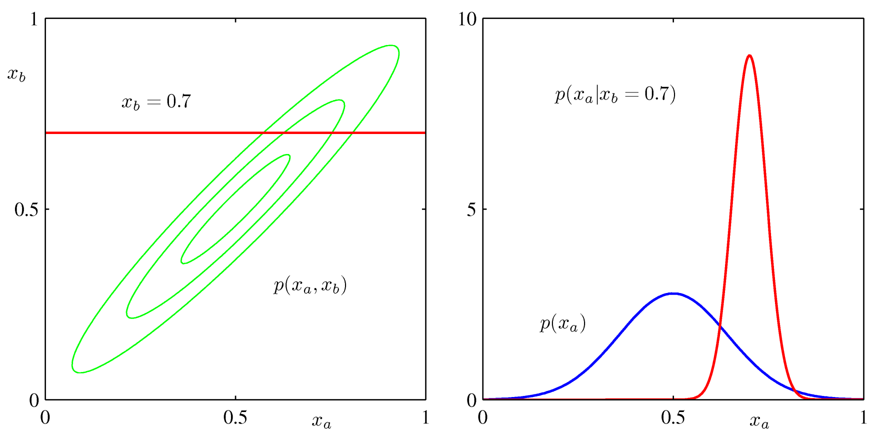
$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$$

$$\vec{X} = \begin{bmatrix} \vec{X}_a \\ \vec{X}_b \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{bmatrix}$$

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

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

Marginal and Conditional of a Gaussian



Bishop 2009

Affine Transformations

- Gaussians are stable under affine transforms
- Affine transformation: $\vec{Y} = \mathbf{A}\vec{X} + \vec{b} \quad (\mathbf{A}$ and $\vec{b}$ are constant)

Affine Transform

$$\begin{aligned} \vec{X} &\sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma}) && \vec{X} \in \mathbb{R}^d \\ \vec{Y} &= \mathbf{A}\vec{X} + \vec{b} && \vec{Y} \in \mathbb{R}^n,\ \mathbf{A} \in \mathbb{R}^{n \times d},\ \vec{b} \in \mathbb{R}^n \end{aligned}$$

$$\vec{Y} \sim \mathcal{N}(\vec{y} \mid \vec{\mu}_Y, \Sigma_Y)$$

$$\begin{aligned} \vec{\mu}_Y &= \mathbf{A}\vec{\mu} + \vec{b} \\ \mathbf{\Sigma}_Y &= \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathsf{T}} \end{aligned}$$

Standard Normal

Univariate Standard Normal

$$X \;\sim \mathcal{N}(0, 1)$$

$$\mu = 0 \qquad \sigma = 1$$

Multivariate Standard Normal

$$\vec{X} \;\sim \mathcal{N}(0, \mathbf{I}_d)$$

$$\vec{\mu} = 0 \qquad \sigma = \mathbf{I}$$

Standardizing

- Transform a normal distributed variable X into a standard normal Z:
- Also called *whitening* or *Z transform / score*

Univariate

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \rightarrow \quad Z = \frac{X - \mu}{\sigma} \quad \rightarrow \quad Z \sim \mathcal{N}(0, 1)$$

Multivariate

$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma}) \quad \rightarrow \quad \vec{Z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\vec{X} - \vec{\mu}) \quad \rightarrow \quad \vec{Z} \sim \mathcal{N}(0, \mathbf{I})$$

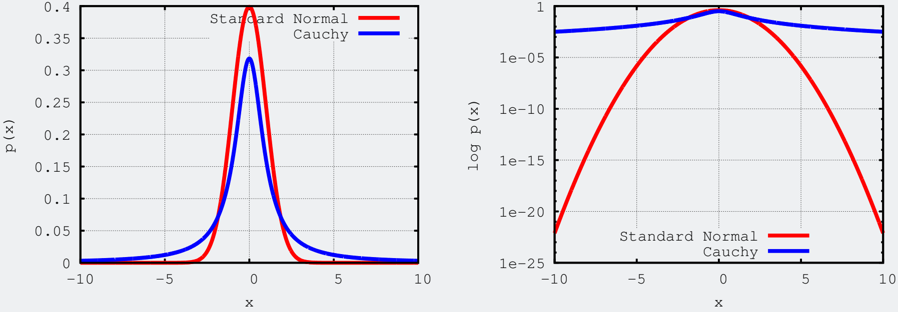
use $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}^2\mathbf{U}^\mathsf{T} \Rightarrow \mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{U}\mathbf{D}$

When to Stop using Gaussians

Gaussians are very handy and can be used in a lot of situations, but be careful if one of the these points applies to your problem:

- Gaussians do not have heavy tails
  - In many real world (empirical) distributions extreme events occur far more often than a Gaussian would allow
- Gaussians have only a single mode
  - Can use a mixture of Gaussians here (see lecture)
- The central limit theorem is only valid for sums of independent random variables
  - For products use a log-normal distribution
  - The variables need to have finite mean and variance
- If you only know the mean and you know nothing about the variance
  - Use an exponential distribution in this case (maximum entropy)

Heavy Tails



Standard Normal
Cauchy
p(x)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−10
−5
0
5
10
x
log p(x)
1
1e−05
1e−10
1e−15
1e−20
1e−25
Standard Normal
Cauchy
−10
−5
0
5
10
x