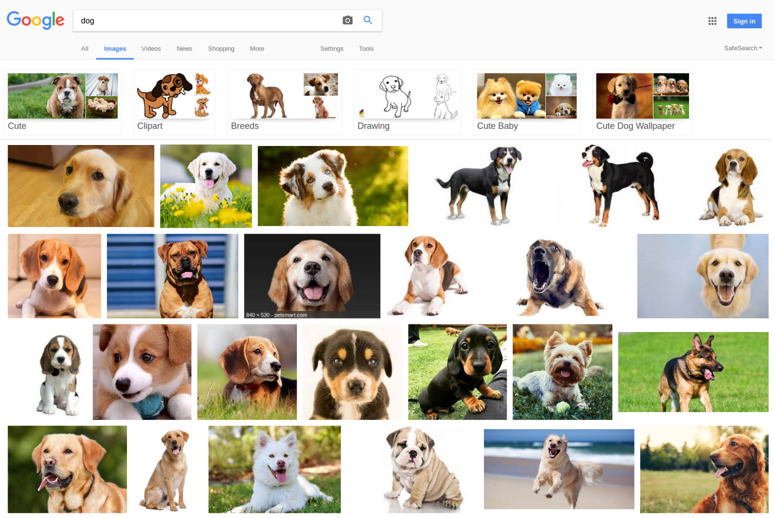
Pattern Recognition: Probability Theory

Dennis Madsen

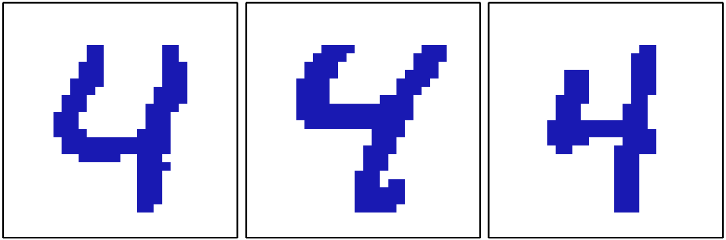
Department of Mathematics and Computer Science
University of Basel

Fall Semester 2018

Variability of a pattern - Dog

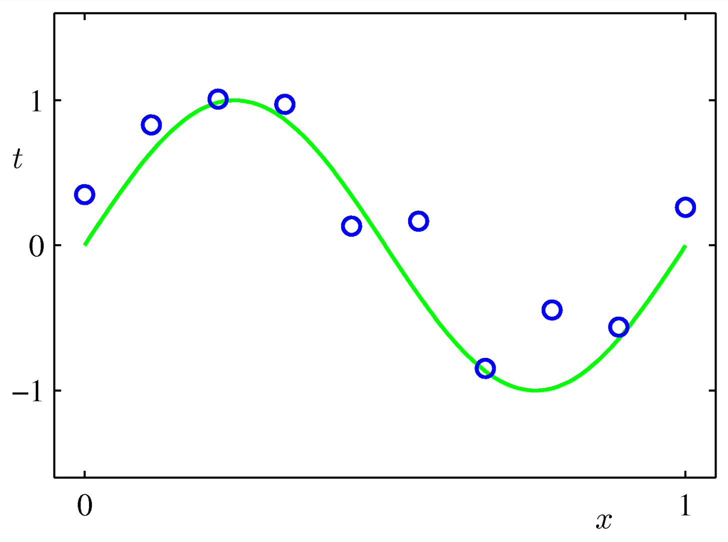


Variability of a pattern - Digit 4



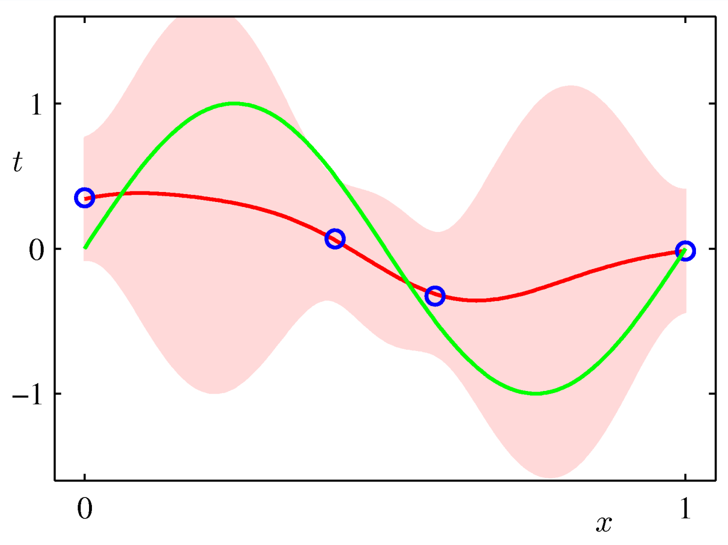
Bishop 2009

Variability of measurement (noise)



Bishop 2009

Uncertainty in the model



Bishop 2009

Motivation

Why do we need probability theory??

Probability and Statistics

To model

- Variability of pattern itself
- Variability of measurement (noise)
- Uncertainty in our model

Motivation

Why do we need probability theory??

Probability and Statistics

To model

- Variability of pattern itself
- Variability of measurement (noise)
- Uncertainty in our model

$\Rightarrow$ A short repetition of probability theory in the context of pattern recognition

- First Part: Theory $\rightarrow$ quick reference for you
- Second Part: Multivariate Gaussian as an example

Discrete Random Variables

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

Discrete Random Variables

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

Cumulative Distribution Function (cdf)

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

Discrete Random Variables

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

Cumulative Distribution Function (cdf)

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

Probability Mass Function

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

Discrete Random Variables

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

Cumulative Distribution Function (cdf)

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

Probability Mass Function

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

Normalization and Positivity

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

Discrete Random Variables — Examples

Binomial – A coin flip

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

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

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

Continuous Random Variables

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

Continuous Random Variables

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

Cumulative Distribution function (cdf)

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

Continuous Random Variables

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

Cumulative Distribution function (cdf)

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

Probability Density Function (pdf)

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

Continuous Random Variables

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

Cumulative 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

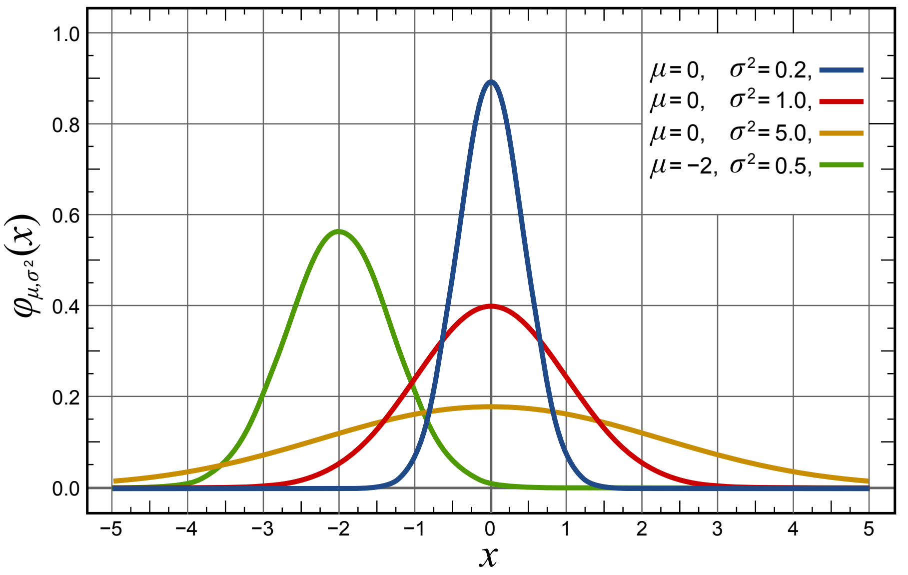
Gaussian

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

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

Mean μ, Variance σ^2

Example: Gaussian



Mean

- The mean is a measure for *central tendency*

Expected Value, Mean, Expectation

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

Variance

- The variance is a measure for *spread*

Variance / Standard Deviation

$$V[X] = E[(X - E[X])^2]$$

$$\mathrm{sd}[X] = \sigma_X = \sqrt{V[X]}$$

Variance

- The variance is a measure for *spread*

Variance / Standard Deviation

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

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

Multivariate Case

Multiple Random Variables

Example

More than one Random Variable, e.g.

Length L and Weight W of an object

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

Multivariate Case

Multiple Random Variables

Example

More than one Random Variable, e.g.

Length L and Weight W of an object

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

Joint Probability

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

$$p(x, y)$$

Marginals and Conditionals

Marginalisation

$$P[X = x] = \sum_{y} P[X = x, Y = y]$$

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

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' Rules

Use the factorization for the joint probability density / distribution:

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

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

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

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

Bayes' Rules

Use the factorization for the joint probability density / distribution:

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

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

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

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

$$\Rightarrow \quad P(\omega_i \mid \underline{x}) = \frac{p(\underline{x} \mid \omega_i) P(\omega_i)}{p(\underline{x})}$$

Bayes' Rules

Use the factorization for the joint probability density / distribution:

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

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

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

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

$$\Rightarrow \quad P(\omega_i \mid \underline{x}) = \frac{p(\underline{x} \mid \omega_i) P(\omega_i)}{p(\underline{x})}$$

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

Covariance and Independence

Covariance

$$\mathrm{Cov}(X, Y) = E[(X - E[X])\ (Y - E[Y])]$$

$$\boldsymbol{\Sigma}(\mathbf{X}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^{\mathsf{T}}]$$

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 and Independence

Covariance

$$\mathrm{Cov}(X, Y) = E[(X - E[X])\,(Y - E[Y])]$$

$$\boldsymbol{\Sigma}(\mathbf{X}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^{\mathsf{T}}]$$

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
- Simple enough to demonstrate these concepts

Multivariate Gaussian Distribution

- This distribution occurs very frequently
- Simple enough to demonstrate these concepts

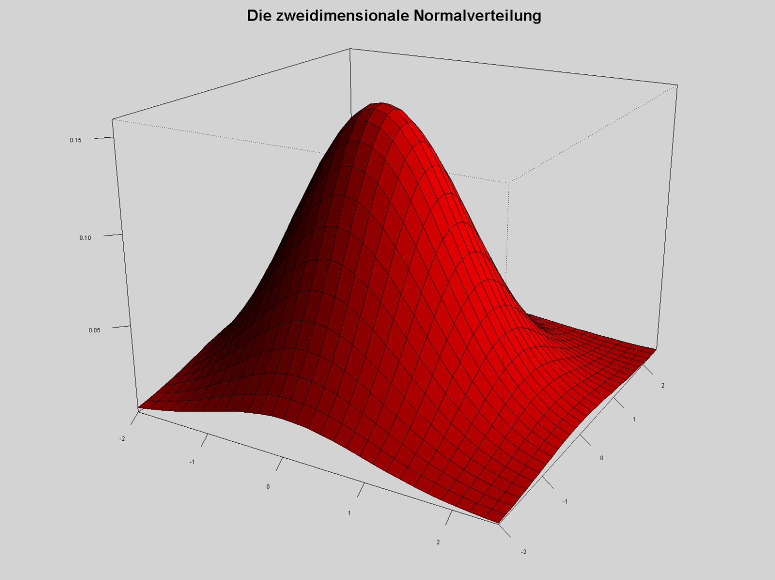
Multivariate Gaussian Distribution

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

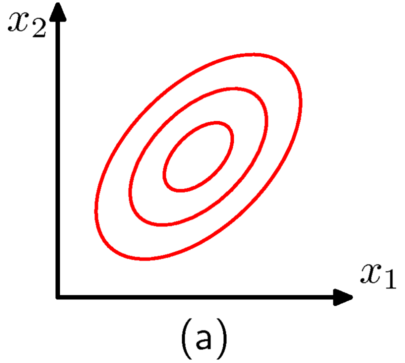
$\vec{\mu}$	Mean
$\mathbf{\Sigma}$	Covariance Matrix ($d \times d$, positive definite, symmetric)
$\lvert\mathbf{\Sigma}\rvert$	Determinant of Σ
d	Number of dimensions

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

2D Gaussian — Surface Plot



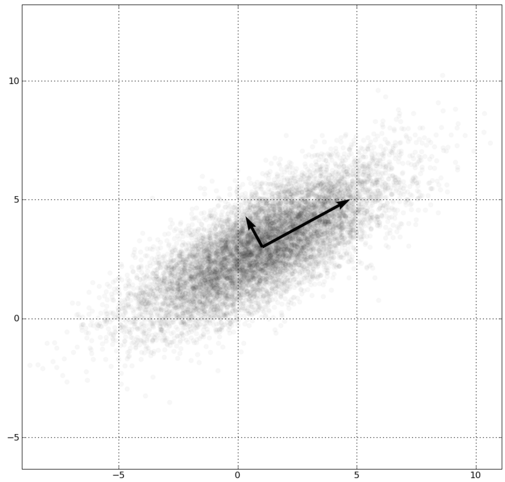
2D Gaussian — Contour Plot



- 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

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

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 $\vec{\mu}$

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 of a Multivariate Gaussian Distribution

Mean

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

Moments of a Multivariate Gaussian Distribution

Mean

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

Covariance

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

Moments of a Multivariate Gaussian Distribution

Mean

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

Covariance

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

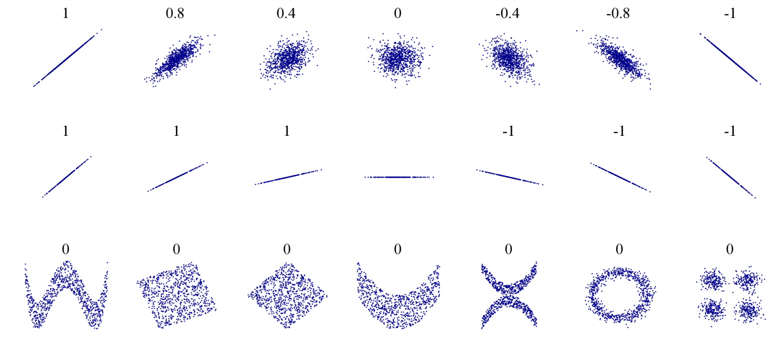
Correlation

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

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"*
- $p(x) = \int p(x, y) dy$

Marginals

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

Marginal of a Gaussian

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

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

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"*
- $p(x \mid y) = \dfrac{p(x, y)}{p(y)}$

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

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

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

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

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

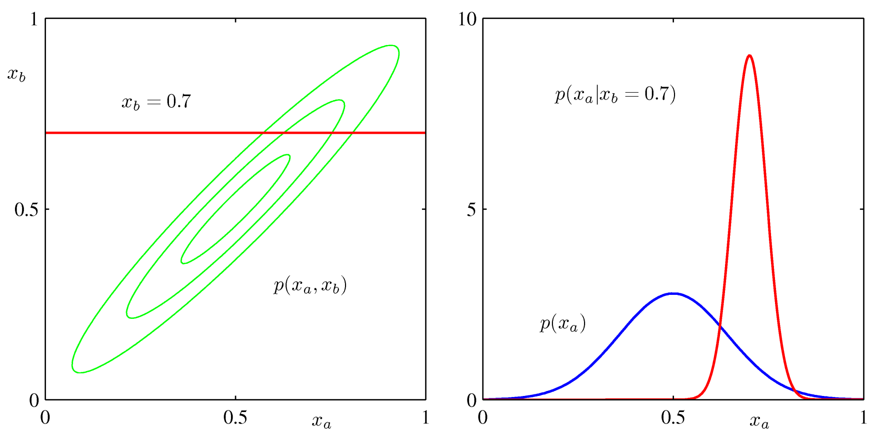
$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$$

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

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

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

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 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

$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma}) \qquad \vec{X} \in \mathbb{R}^d$$

$$\vec{Y} = \mathbf{A}\vec{X} + \vec{b} \qquad \vec{Y} \in \mathbb{R}^n,\ \mathbf{A} \in \mathbb{R}^{n\times d},\ \vec{b} \in \mathbb{R}^n$$

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

Affine Transformations

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

Affine Transform

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

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

$$\vec{\mu}_Y = \mathbf{A}\vec{\mu} + \vec{b}$$
$$\boldsymbol{\Sigma}_Y = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\mathsf{T}$$

Standard Normal

Univariate Standard Normal

$$X \sim \mathcal{N}(0, 1)$$
$$\mu = 0 \qquad \sigma = 1$$

Multivariate Standard Normal

$$\vec{X} \sim \mathcal{N}(0, \mathbf{I}_d)$$
$$\vec{\mu} = 0 \qquad \mathbf{\Sigma} = \mathbf{I}$$

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:

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:

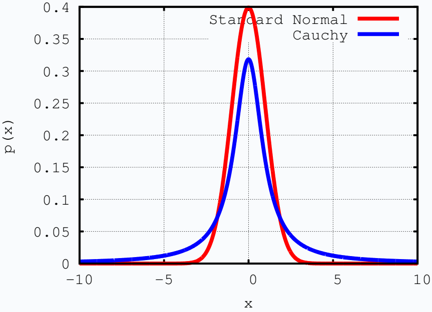
- Gaussians do not have heavy tails
  - In many real world (empirical) distributions extreme events occur far more often than a Gaussian would allow

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)

Heavy Tails



0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
p(x)
−10
−5
0
5
10
x
Standard Normal
Cauchy