UNIKASSEL VERSITÄT

Chair for Communication Technology

Prof. Dr.-Ing. Klaus David

ComTec

Communication Technologies 2 (CT2)

Machine Learning:

Applications and Algorithms

Gaussian Mixture Models

M. Sc. Judith Heinisch

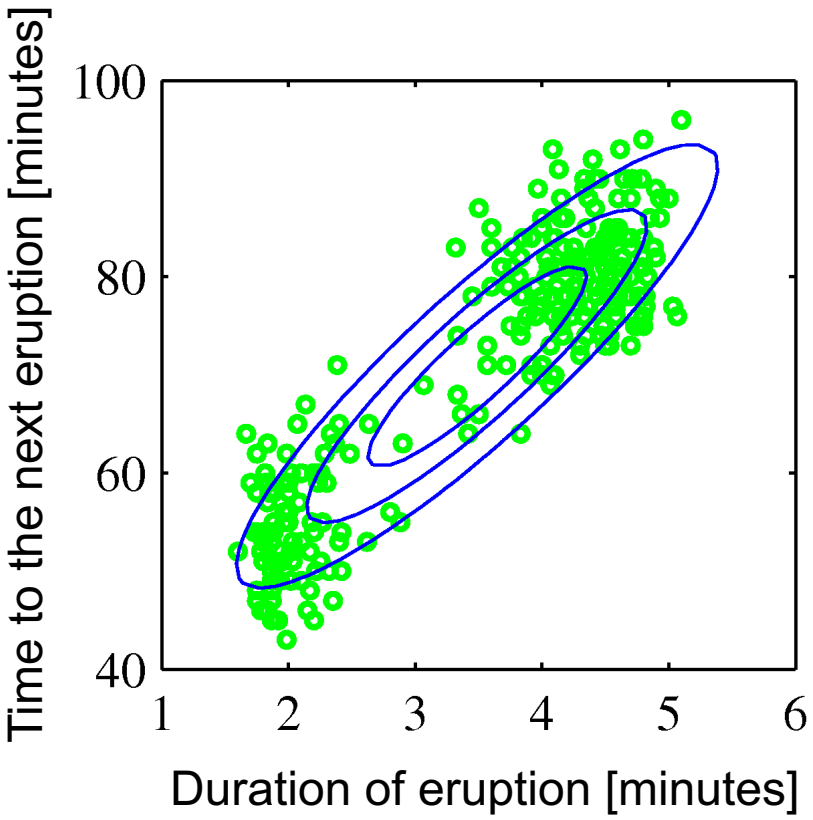
Lecture in WS 2018 / 2019
[13.12.2018]

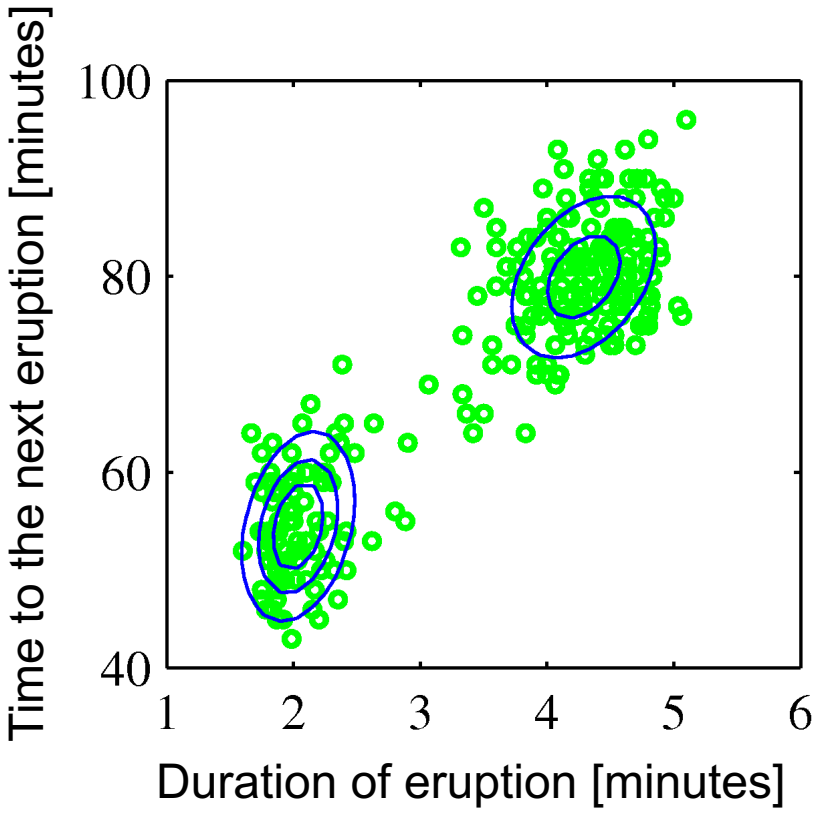
1. Motivation

ComTec

Damon Holley / pexels.com

1. Motivation





Plot of the „Old Faithful“ eruption data (Hydrothermal geyser in Yellowstone National Park). The horizontal axis shows the duration of the eruption in minutes and the vertical axis the time to the next eruption in minutes. [1]

2. Gaussian probability density function

Gaussian probability density function (normal distribution) of a continuous variable x for the class c:

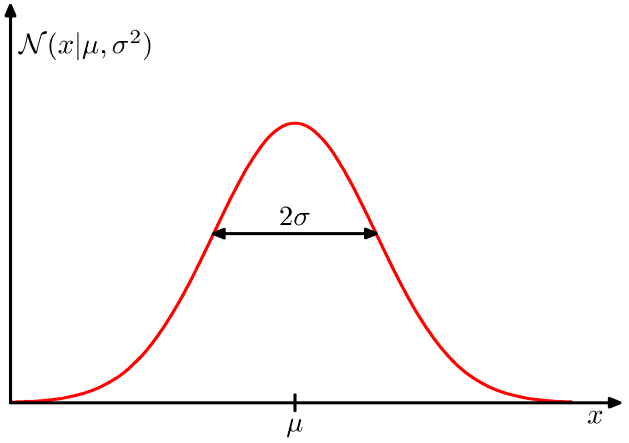
$$\mathcal{N}(x|\mu_c, \sigma_c^2) = \frac{1}{\sqrt{(2\pi\sigma_c^2)}} e^{-\frac{1}{2\sigma_c^2}(x-\mu_c)^2}$$

Mean (expectation value)

$$\mu_c = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu_c, \sigma_c^2) x \mathrm{d}x = \mathbb{E}[x_c]$$

Variance $\sigma_c^2 = \mathbb{E}[x_c^2] - \mathbb{E}[x_c]^2$

Standard deviation $\sigma_c = \sqrt{\sigma_c^2}$



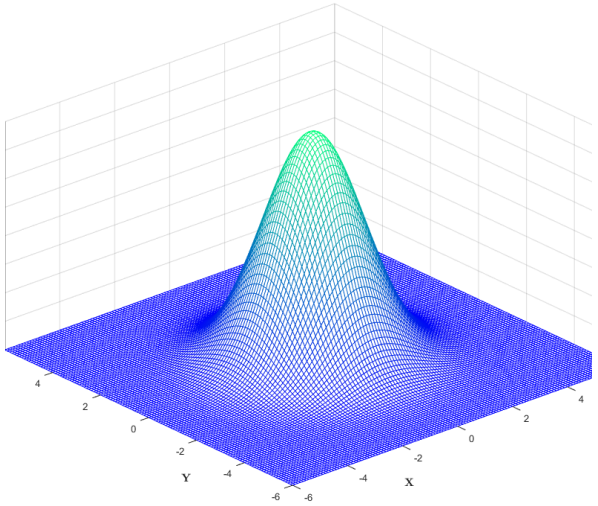
Plot of a Gaussian density [1]

2. Multivariate Gaussian probability density function

ComTec

Multivariate Gaussian probability density function of a D-dimensional vector $\vec{x}$ with continues variables for the class c:

$$\mathcal{N}(\vec{x}|\overrightarrow{\mu_c}, \Sigma_c) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_c|}} e^{\left(-\frac{1}{2}(\vec{x}-\overrightarrow{\mu_c})^T {\Sigma_c}^{-1} (\vec{x}-\overrightarrow{\mu_c})\right)}$$



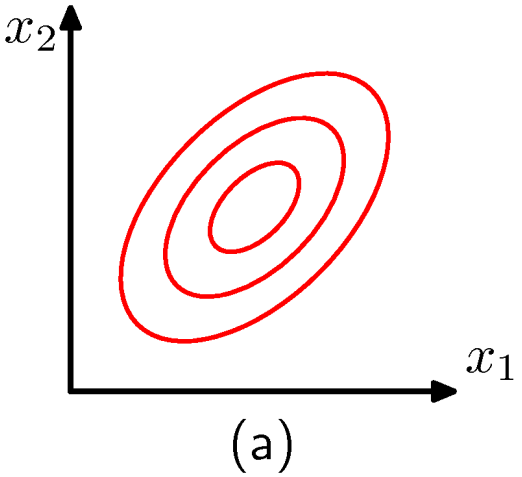
$\overrightarrow{\mu_c}$ is a D-dimensional mean vector

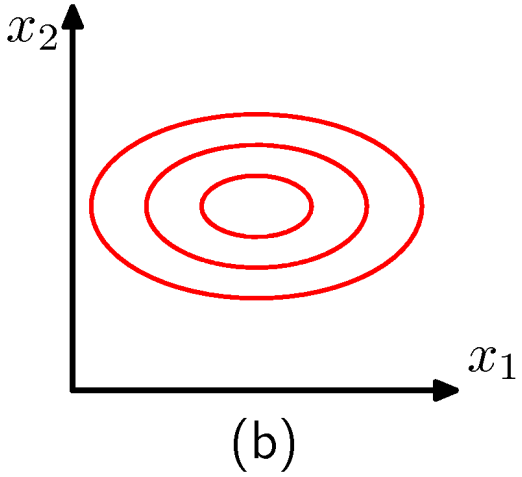
Σ_c is a DxD covariance matrix

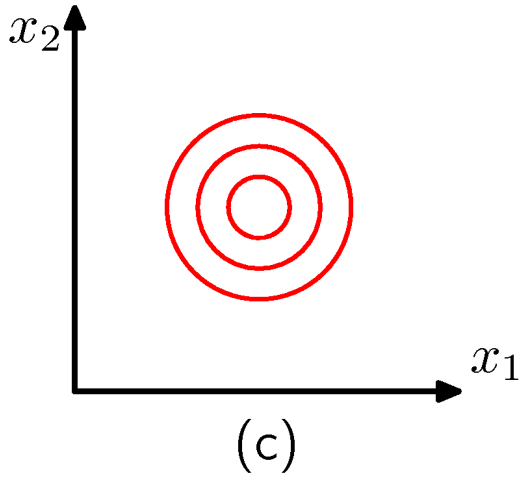
$|\Sigma_c|$ is the determinant of Σ_c

2. Covariance matrix - examples

Contours of a constant Gaussian probability density function in a two dimensional space.







a) General form of a covariance matrix or rather all fields of the covariance matrix are not zero.

$$\Sigma_c = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The values of the covariance matrix except the diagonal values are equal 0.

In this example:

$$\Sigma_c = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Covariance matrix is proportional to the identity matrix.

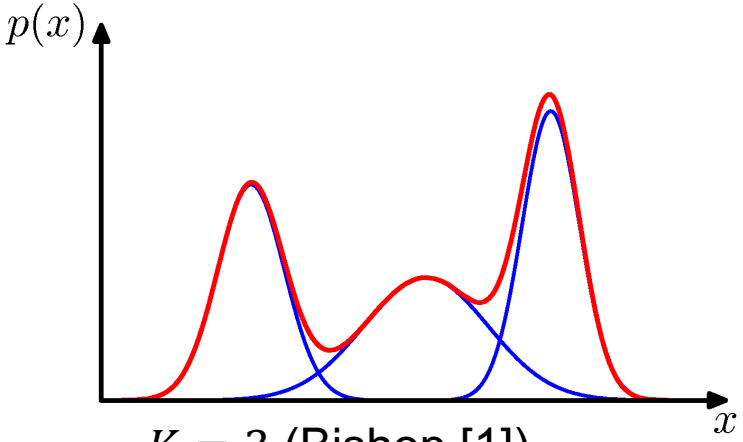
$$\Sigma_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

[1]

3. Gaussian Mixture Model (GMM)

For a D-dimensional vector of variables $\vec{x}$ the **GMM** is defined as a weighted sum of K Gaussian density functions:

$$p(\vec{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\vec{x}|\overrightarrow{\mu_k}, \Sigma_k)$$

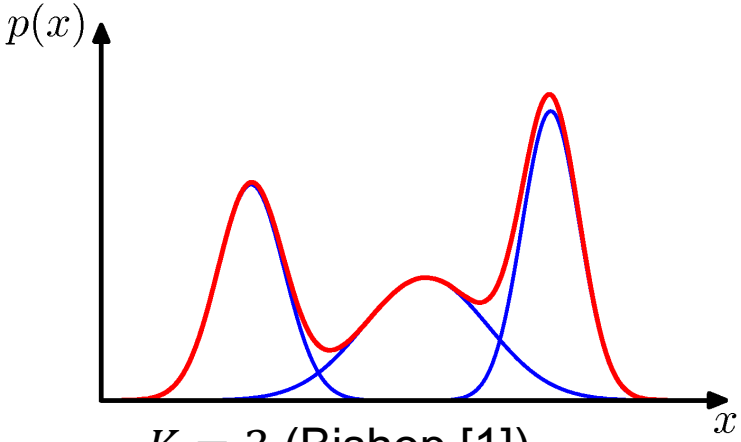


$K = 3$ (Bishop [1])
with one dimension

3. Gaussian Mixture Model (GMM)

For a D-dimensional vector of variables $\vec{x}$ the **GMM** is defined as a weighted sum of K Gaussian density functions:

$$p(\vec{x}) = \sum_{k=1}^{K} \omega_k \underbrace{\mathcal{N}(\vec{x}|\overrightarrow{\mu_k}, \Sigma_k)}_{\text{Component (Gaussian distribution)}}$$



$K = 3$ (Bishop [1])
with one dimension

3. Gaussian Mixture Model (GMM)

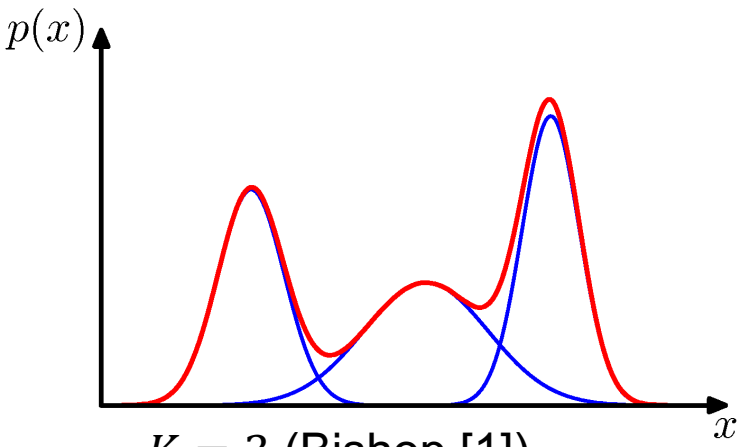
For a D-dimensional vector of variables $\vec{x}$ the **GMM** is defined as a weighted sum of K Gaussian density functions:

$$p(\vec{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\vec{x}|\overrightarrow{\mu_k}, \Sigma_k)$$

Component (Gaussian distribution)

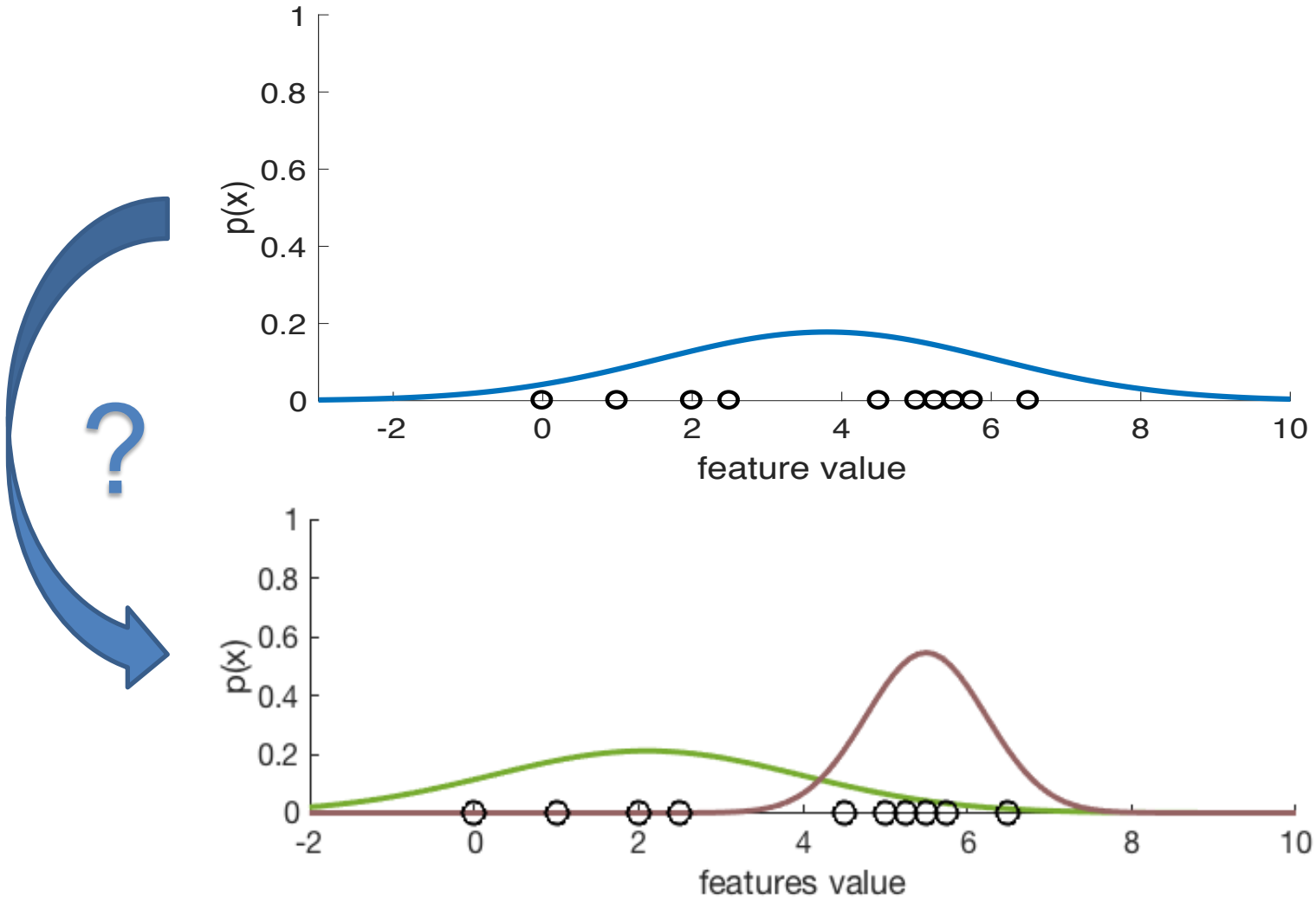
Mixing coefficients
Given for all ω_k:

$1 \geq \omega_k \geq 0$
and
$\sum_{k=1}^{K} \omega_k = 1$



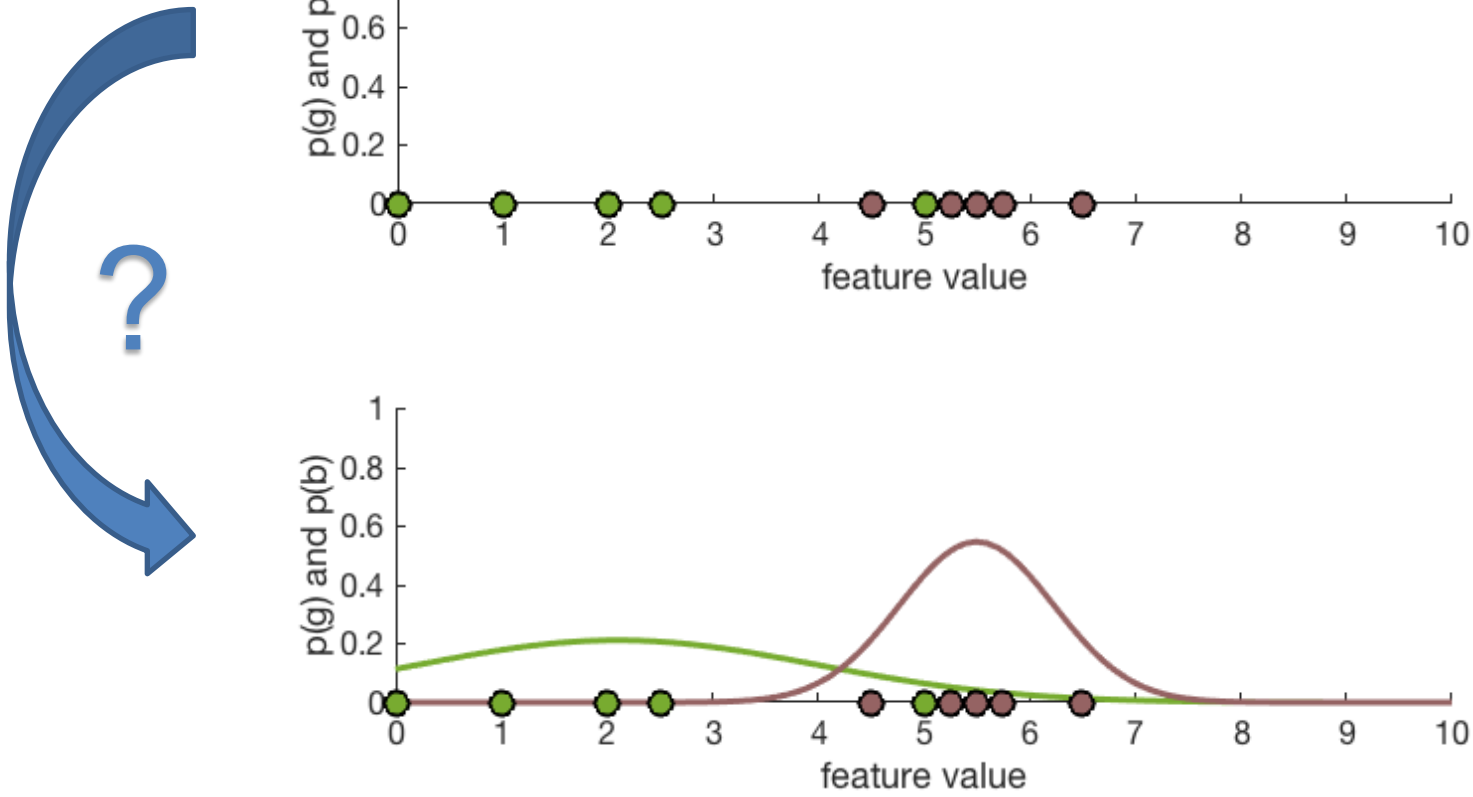
$K = 3$ (Bishop [1])
with one dimension

3. Example - GMM



(Victor Lavrenko [2])

3. Example 1 - GMM



(Victor Lavrenko [2])

3. Example 1 - GMM in 1-D

ComTec

- Labelled data set $x_1, x_2, \ldots x_n$
- Two classes $c_1 = green = g$ and $c_2 = brown = b$
- We don't know μ_g, σ_g^2 and μ_b, σ_b^2 to get the Gaussian bell curves

x	0	1	2	2.5	4.5	5	5.25	5.5	5.75	6.5
c	g	g	g	g	b	g	b	b	b	b

(Victor Lavrenko [2])

3. Example 1 - GMM in 1-D

ComTec

- Labelled data set $x_1, x_2, \dots x_n$
- Two classes $c_1 = green = g$ and $c_2 = brown = b$
- We don't know μ_g, σ_g^2 and μ_b, σ_b^2 to get the Gaussian bell curves
- The sum of data points containing to one component / class k is N

Mean:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

x	0	1	2	2.5	4.5	5	5.25	5.5	5.75	6.5
c	g	g	g	g	b	g	b	b	b	b

(Victor Lavrenko [2])

3. Example 1 – GMM in 1-D

ComTec

$$\mu_g = \frac{0 + 1 + 2 + 2.5 + 5}{5} = 2.1$$

$$\sigma_g^2 = \frac{(0 - 2.1)^2 + (1 - 2.1)^2 + (2 - 2.1)^2 + (2.5 - 2.1)^2 + (5 - 2.1)^2}{5} = 2.84$$

$$\sigma_g = \sqrt{\sigma_g^2} = \sqrt{2.84} = 1.685$$

$$\mu_b = 5.5 \qquad \sigma_b^2 = 0.425 \qquad \sigma_b = 0.652$$

(Victor Lavrenko [2])

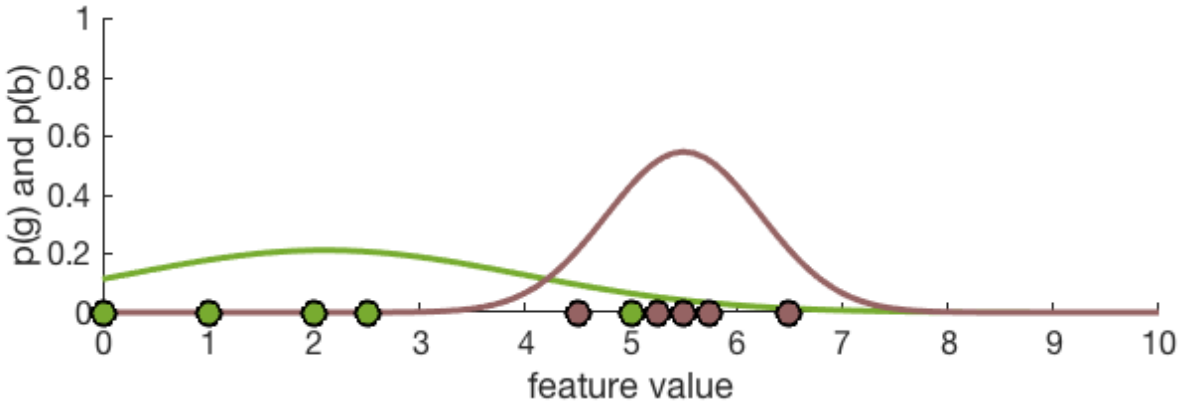
3. Example 1 – GMM in 1-D

$$\mu_g = \frac{0 + 1 + 2 + 2.5 + 5}{5} = 2.1$$

$$\sigma_g^2 = \frac{(0 - 2.1)^2 + (1 - 2.1)^2 + (2 - 2.1)^2 + (2.5 - 2.1)^2 + (5 - 2.1)^2}{5} = 2.84$$

$$\sigma_g = \sqrt{\sigma_g^2} = \sqrt{2.84} = 1.685$$

$$\mu_b = 5.5 \qquad \sigma_b^2 = 0.425 \qquad \sigma_b = 0.652$$



(Victor Lavrenko [2])

3. Example 1 – GMM in 1-D

ComTec

To get the GMM

$$p(x) = \sum_{k=1}^{K} \omega_k \mathcal{N}(x|\mu_k, \sigma^2)$$

we still need the mixing coefficients ω_g and ω_b.

$$\omega_g = \frac{1}{N} \sum_{n=1}^{N} p(g|x)$$

3. Example 1 – GMM in 1-D

ComTec

We still need the mixing coefficients ω_g and ω_b.

$$\omega_g = \frac{1}{N}\sum_{n=1}^{N} p(g|x)$$

Assume $p(g) = p(b) = 0.5$

$$p(g|x) = \frac{p(x|g)p(g)}{p(x|g)p(g) + p(x|b)p(b)}$$

$$p(x|g) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)}$$

$\mu_b = 5.5$	$\sigma_b^2 = 0.425$	$\sigma_b = 0.652$
$\mu_g = 2.1$	$\sigma_g^2 = 2.84$	$\sigma_g = 1.685$

3. Example 1 – GMM in 1-D

ComTec

Assume $p(g) = p(b) = 0.5$

$$\omega_g = \frac{1}{N}\sum_{n=1}^{N} p(g|x) = \frac{1}{10}\sum_{n=1}^{10} p(g|x) = 0.4578$$

$$\omega_b = \frac{1}{10}\sum_{n=1}^{10} p(b|x) = 0.5422$$

The final result we get is:

3. Example 1 – GMM in 1-D

ComTec

Assume $p(g) = p(b) = 0.5$

$$\omega_g = \frac{1}{N}\sum_{n=1}^{N} p(g|x) = \frac{1}{10}\sum_{n=1}^{10} p(g|x) = 0.4578$$

$$\omega_b = \frac{1}{10}\sum_{n=1}^{10} p(b|x) = 0.5422$$

The final result we get is:

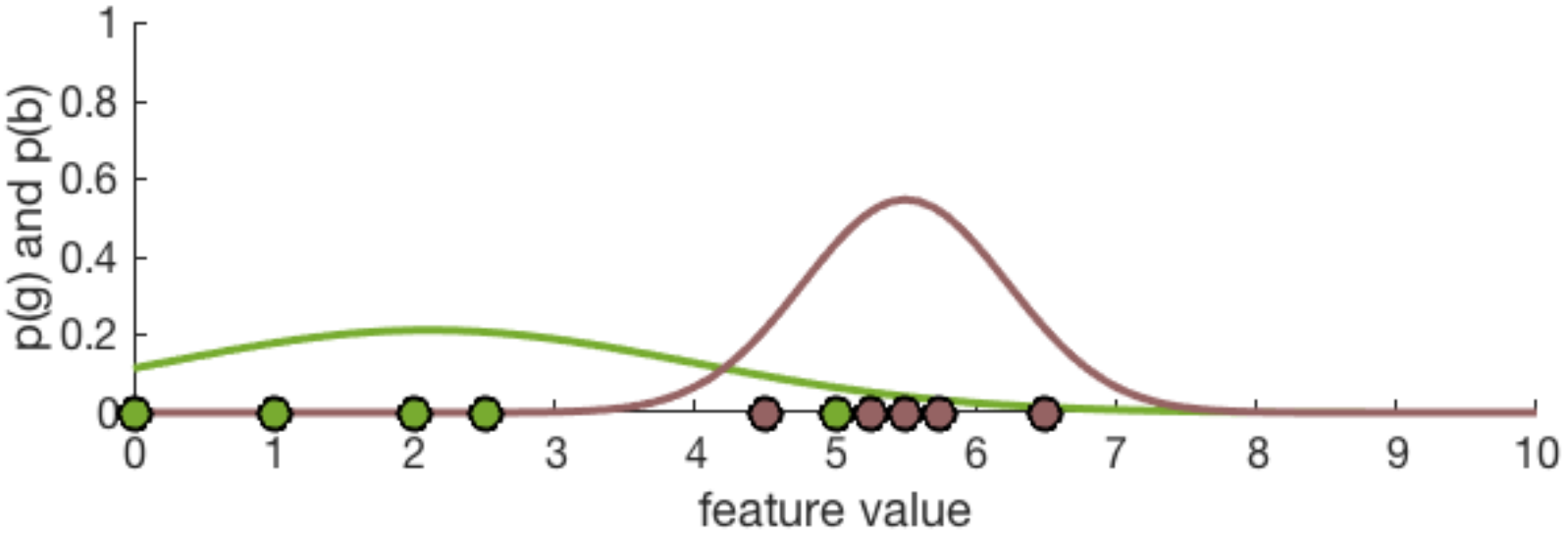
$$p(x) = 0.4578 * \mathcal{N}(x|2.1,2.84) + 0.5422 * \mathcal{N}(x|5.5,0.425)$$

3. Example 1 – GMM in 1-D

ComTec

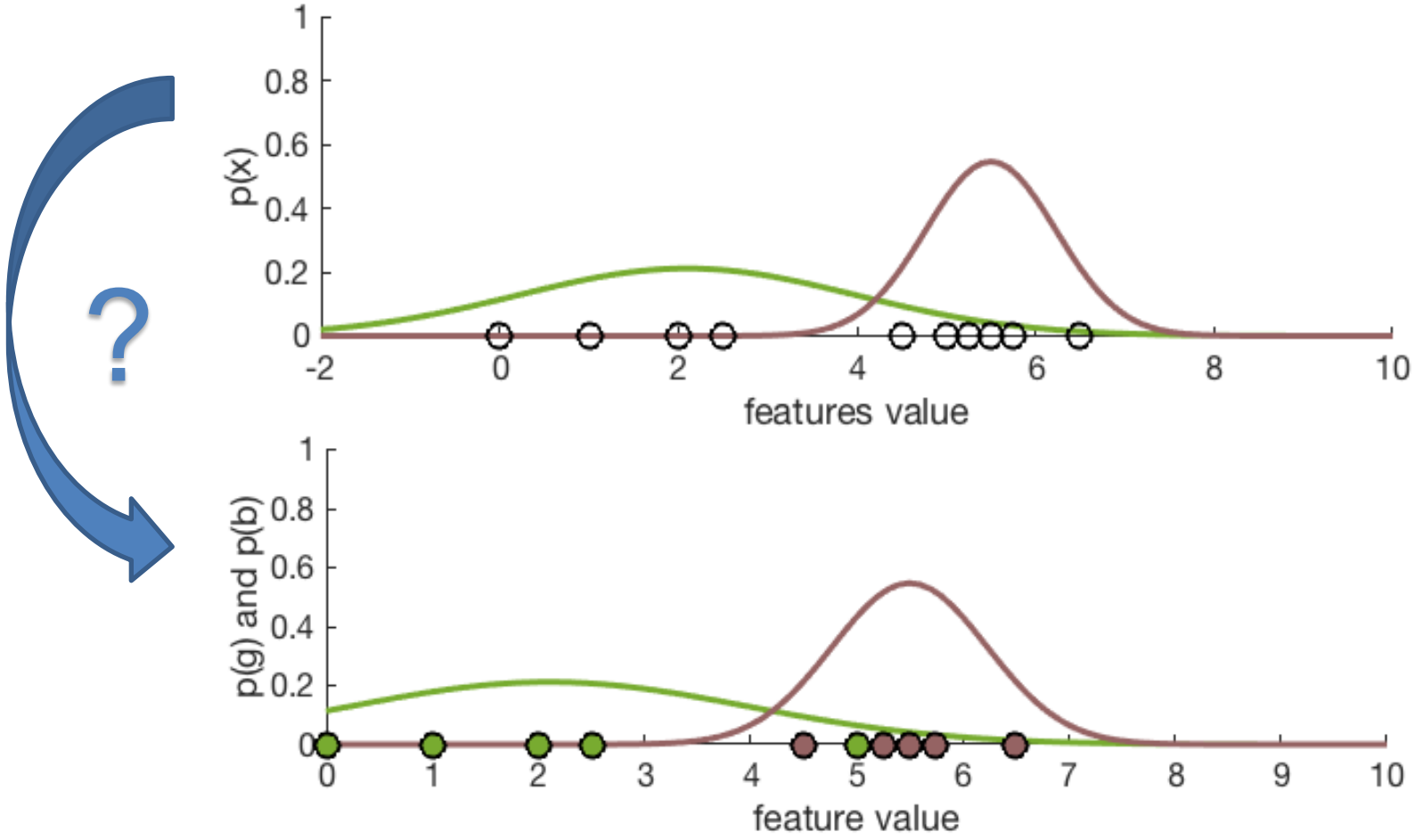
The final result we get is:

$$p(x) = 0.4578 * \mathcal{N}(x|2.1,2.84) + 0.5422 * \mathcal{N}(x|5.5,0.425)$$



3. Example 2 – GMM in 1-D

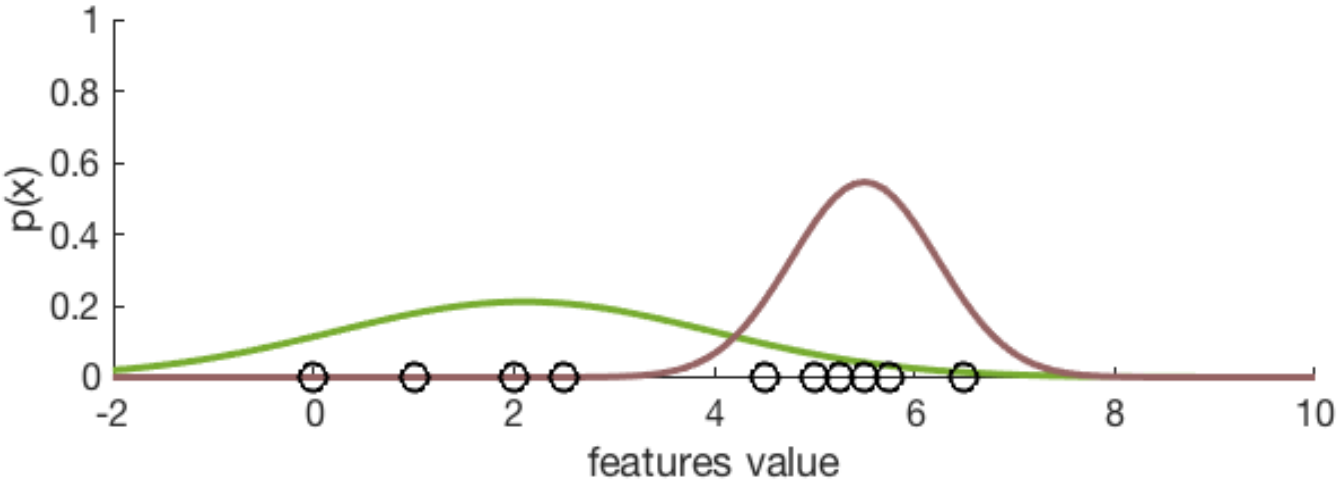
What if we don‘t know the data set?



3. Example 2 - GMM in 1-D

What if we don‘t know the data set?

If we knew parameters of the Gaussians μ_g, σ_g^2 and μ_b, σ_b^2, we can calculate whether the data point is more likely to be green or brown.



(Victor Lavrenko [2])

3. Example 2 - GMM in 1-D

ComTec

How typical is x under source green?

$$p(x|g) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)}$$

How likely that x comes from green?

$$p(g|x) = \frac{p(x|g)p(g)}{p(x|g)p(g) + p(x|b)p(b)}$$

Assumption $x = 6.5$ and $p(g) = p(b) = 0.5$

$\mu_b = 5.5$	$\sigma_b^2 = 0.425$	$\sigma_b = 0.652$
$\mu_g = 2.1$	$\sigma_g^2 = 2.84$	$\sigma_g = 1.685$

(Victor Lavrenko [2])

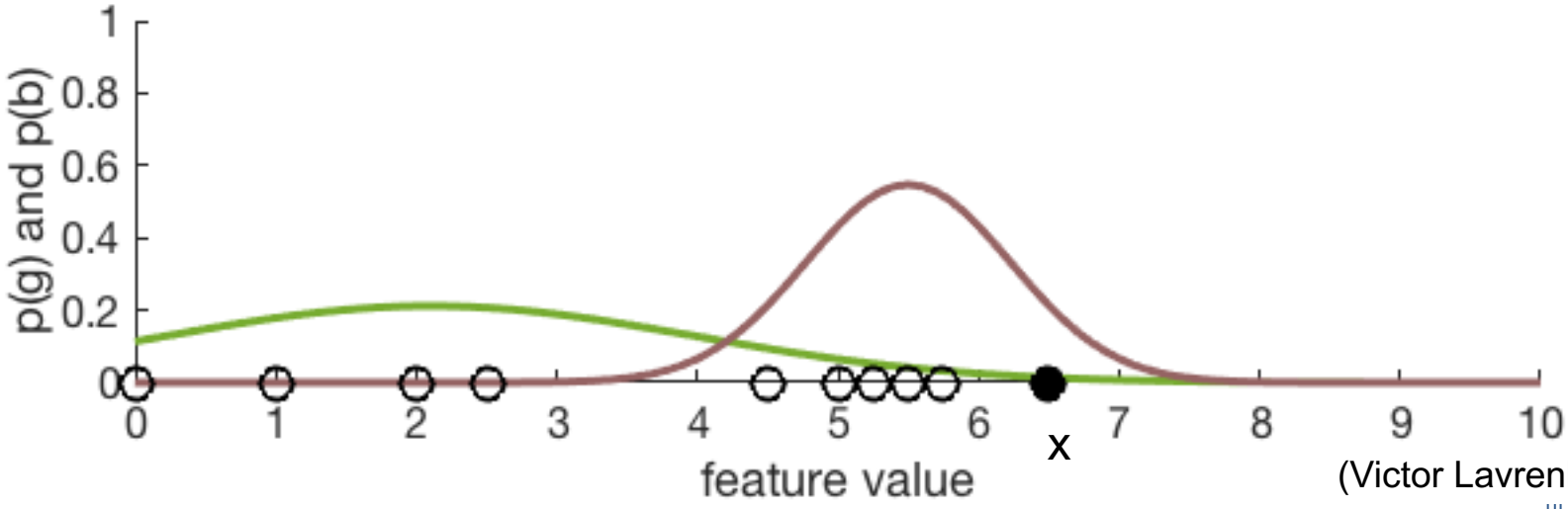
3. Example 2 - GMM in 1-D

How typical is $x = 6.5$ under source green/brown?

$$p(x|g) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)}$$

$$p(x = 6.5\,|g) = \frac{1}{\sqrt{2\pi * 2.84}} e^{\left(-\frac{(6.5\,-2.1)^2}{2*2.84}\right)} = 0.00783$$

$$p(x = 6.5\,|b) = \frac{1}{\sqrt{2\pi * 0.425}} e^{\left(-\frac{(6.5\,-5.5)^2}{2*0.425}\right)} = 0.18870$$



(Victor Lavrenko [2])

3. Example 2 - GMM in 1-D

How likely that x comes from green?

$$p(g|x) = \frac{p(x|g)p(g)}{p(x|g)p(g) + p(x|b)p(b)}$$

Assumption $x = 6.5$ and $p(g) = p(b) = 0.5$

$\mu_b = 5.5$ $\quad \sigma_b^2 = 0.425$ $\quad \sigma_b = 0.652$

$\mu_g = 2.1$ $\quad \sigma_g^2 = 2.84$ $\quad \sigma_g = 1.685$

$p(x = 6.5\,|g) = 0.00783$

$p(x = 6.5\,|b) = 0.18870$

(Victor Lavrenko [2])

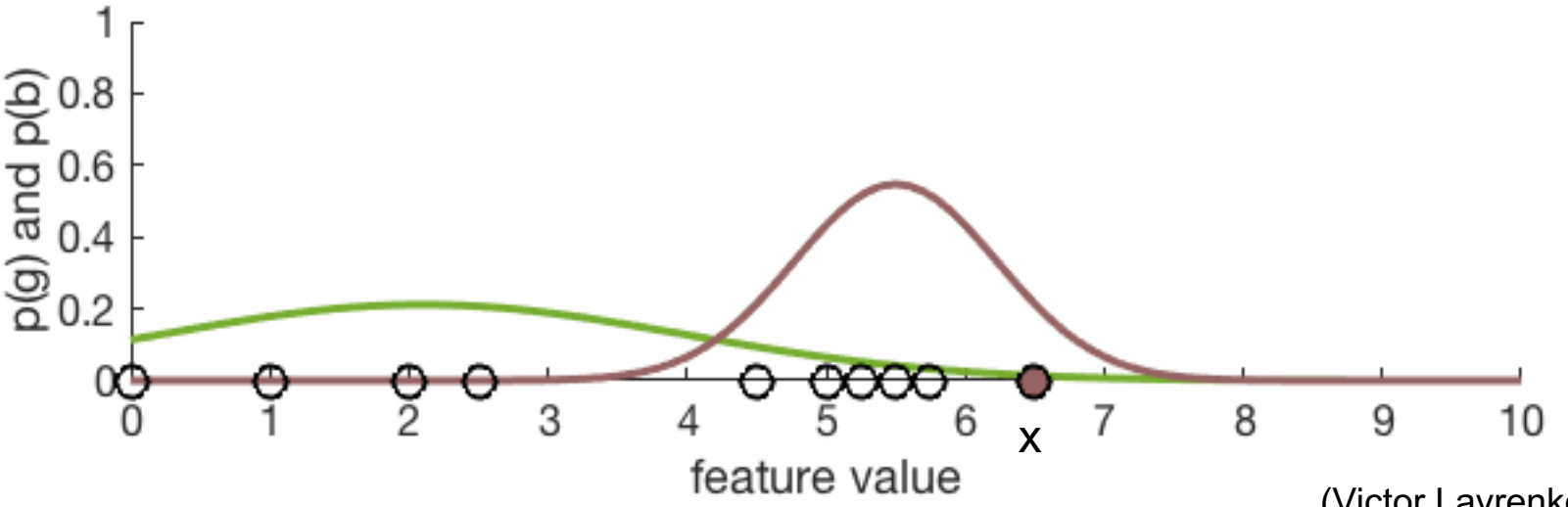
3. Example 2 - GMM in 1-D

How likely that $x = 6.5$ comes from green/brown?

$$p(g|x) = \frac{p(x|g)p(g)}{p(x|g)p(g) + p(x|b)p(b)}$$

$$p(g|x = 6.5) = \frac{0.00783 * 0.5}{0.00783 * 0.5 + 0.1887 * 0.5} = 0.03984$$

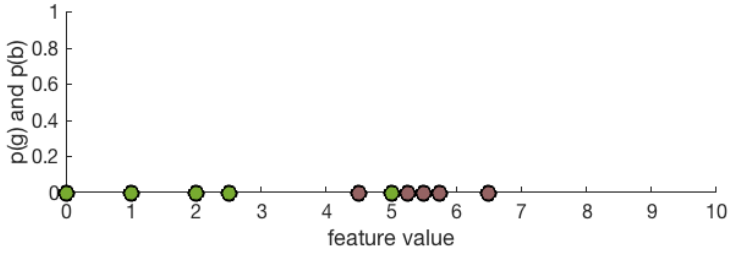
$$p(b|x = 6.5) = \frac{0.1887 * 0.5}{0.00783 * 0.5 + 0.1887 * 0.5} = 0.96016$$

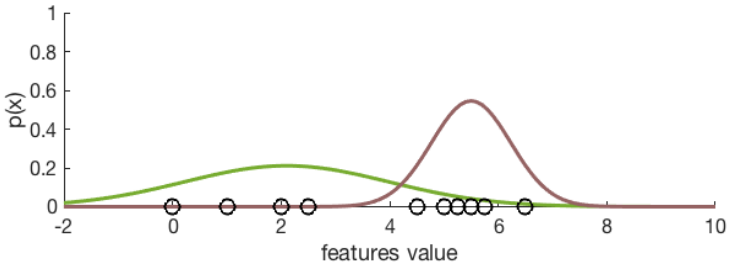


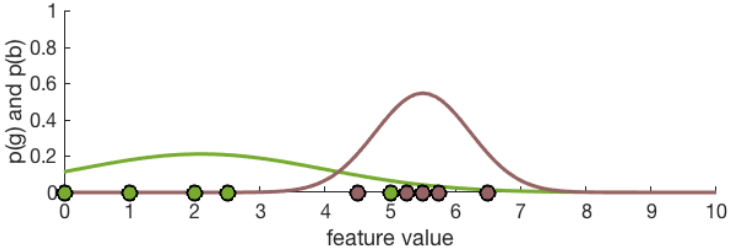
(Victor Lavrenko [2])

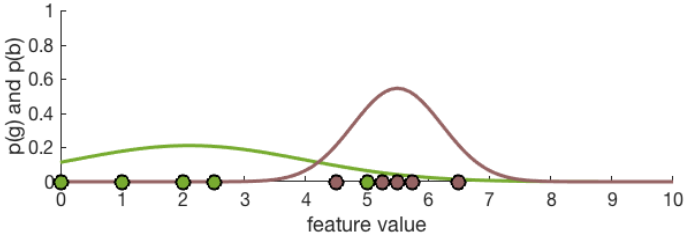
3. Example – GMM Summary

ComTec



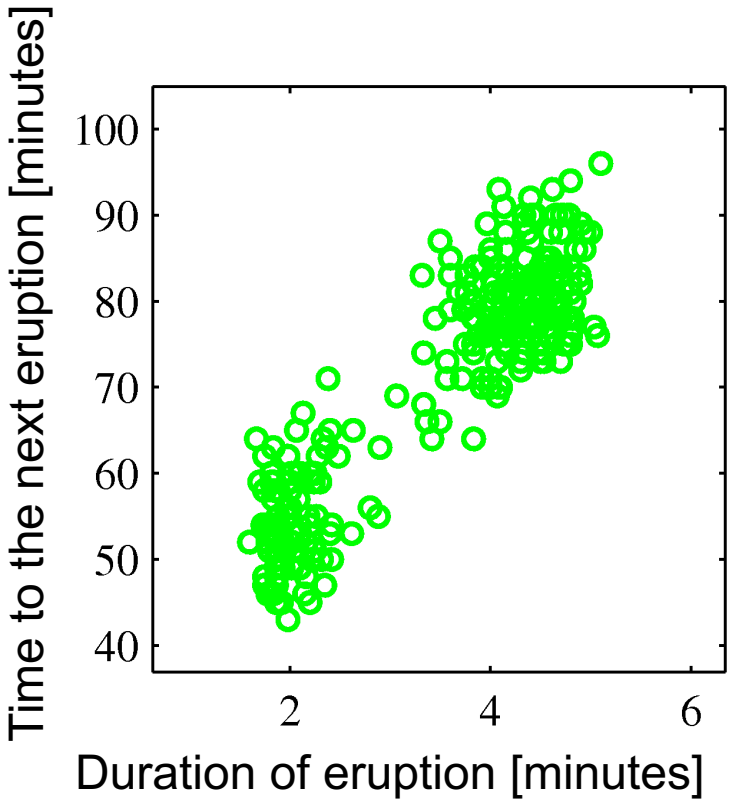




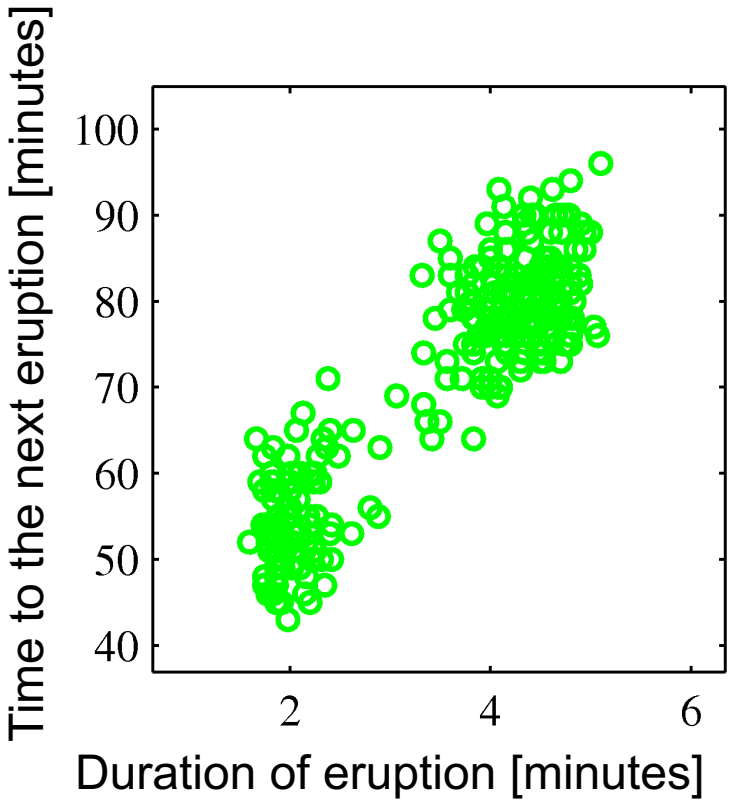


3. Example – GMM Summary

ComTec



3. Example – GMM Summary

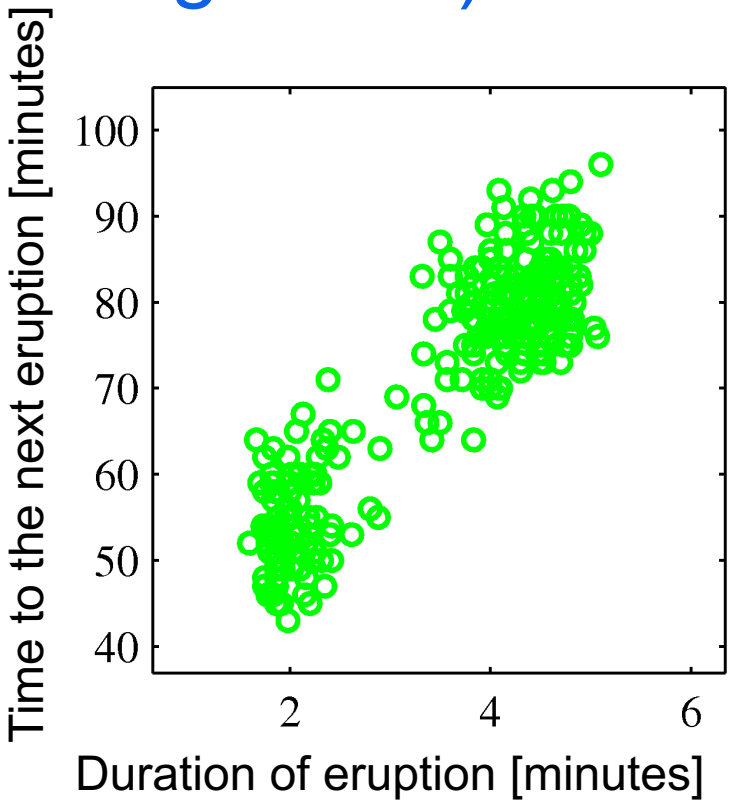


Problem:

- To guess source of data points we need μ_c and σ_c^2
- To estimate μ_c and σ_c^2 we need to know the source of the data points

4. Expectation Maximization algorithm (EM algorithm)

ComTec



Solution: **Expectation Maximization algorithm**

'[…] the goal is to maximize the likelihood function with respect to the parameters.' (Bishop [1])

4. Likelihood Function

ComTec

Bayes Theorem:

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

with $D = \{x_1, x_2, \ldots, x_N\}$ containing the observed data and $\vec{w} = \{w_1, w_2, \ldots, w_k\}$ the model parameters.

4. Likelihood Function

ComTec

Bayes Theorem:

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

Posterior Probability
Given the data D, probability for the model parameters $\vec{w}$ (or evaluates the uncertainty in the parameters after observing the data)

with $D = \{x_1, x_2, \ldots, x_N\}$ containing the observed data and $\vec{w} = \{w_1, w_2, \ldots, w_k\}$ the model parameters.

4. Likelihood Function

ComTec

Bayes Theorem:

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

Posterior Probability
Given the data D, probability for the model parameters $\vec{w}$ (or evaluates the uncertainty in the parameters after observing the data)

Prior Probability
for the model parameters $\vec{w}$ (knowledge we have before performing EM)

with $D = \{x_1, x_2, \ldots, x_N\}$ containing the observed data and $\vec{w} = \{w_1, w_2, \ldots, w_k\}$ the model parameters.

4. Likelihood Function

Bayes Theorem:

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

Posterior Probability
Given the data D, probability for the model parameters $\vec{w}$ (or evaluates the uncertainty in the parameters after observing the data)

Normalization constant

Prior Probability
for the model parameters $\vec{w}$ (knowledge we have before performing EM)

with $D = \{x_1, x_2, \ldots, x_N\}$ containing the observed data and $\vec{w} = \{w_1, w_2, \ldots, w_k\}$ the model parameters.

4. Likelihood Function

Bayes Theorem:

Likelihood Function
Assuming to know the model parameters $\vec{w}$, the probability for the data D (or influence of the data)

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

Posterior Probability
Given the data D, probability for the model parameters $\vec{w}$ (or evaluates the uncertainty in the parameters after observing the data)

Normalization constant

Prior Probability
for the model parameters $\vec{w}$ (knowledge we have before performing EM)

with $D = \{x_1, x_2, \dots, x_N\}$ containing the observed data and $\vec{w} = \{w_1, w_2, \dots, w_k\}$ the model parameters.

4. Likelihood Function

ComTec

Bayes Theorem:

Likelihood Function
Assuming to know the model parameters $\vec{w}$, the probability for the data D (or influence of the data)

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

Posterior Probability
Given the data D, probability for the model parameters $\vec{w}$ (or evaluates the uncertainty in the parameters after observing the data)

Normalization constant

Prior Probability
for the model parameters $\vec{w}$ (knowledge we have before performing EM)

Maximum Likelihood method serves as estimation of model parameters. For this the likelihood function is maximized.

4. Likelihood Function

ComTec

Likelihood Function $p(D|\vec{w})$

Assuming to know the model parameters $\vec{w}$, the probability for the data D (or influence of the data)

Given D-dimensional data points $\overrightarrow{x_n}$. Let $X = \left\{\overrightarrow{x_1}^T, \dots, \overrightarrow{x_N}^T\right\}$ be a matrix which contains the D-dimensional data points $\overrightarrow{x_n}$. $\boldsymbol{\mu} \equiv \{\overrightarrow{\mu_1}, \overrightarrow{\mu_2}, \dots, \overrightarrow{\mu_K}\}$ containing all mean values, $\vec{\omega} \equiv \{\omega_1, \omega_2, \dots, \omega_k\}$ containing all components and $\boldsymbol{\Sigma} \equiv \{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$ containing all covariance matrices.

The log likelihood function for a Gaussian mixture model is given by:

$$\ln p(X|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \vec{\omega}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \omega_k \mathcal{N}(\overrightarrow{x_n}|\overrightarrow{\mu_k}, \Sigma_k) \right\}$$

4. Expectation Maximization Algorithm (EM- Algorithm)

ComTec

The idea of the EM algorithm is to **maximize** the probability of a stochastically process with a given model. For this, we iterate between classification (**Expectation**) and adaption of the model's parameters (**Maximization**), alternately.

[3]

4. EM for GMM

ComTec

1. Initialize all mean values $\overrightarrow{\mu_k}$, covariances Σ_k and mixing coefficients ω_k. Given are D-dimensional data points $\overrightarrow{x_n}$. Let $X = \left\{\overrightarrow{x_1}^T, \ldots, \overrightarrow{x_N}^T\right\}$ be a matrix which represents the data points. Calculate the initial value of the log likelihood function:

$$\ln p(X|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \vec{\omega}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \omega_k \mathcal{N}(\overrightarrow{x_n}|\overrightarrow{\mu_k}, \Sigma_k) \right\}$$

2. Expectation step:

   Evaluate the probability $p(k|\overrightarrow{x_n})$ that x_n belongs to component k:

$$p(k|\overrightarrow{x_n}) = \frac{\omega_k \mathcal{N}(\overrightarrow{x_n}|\overrightarrow{\mu_k}, \Sigma_k)}{\sum_{j=1}^{K} \omega_j \mathcal{N}(\overrightarrow{x_n}|\overrightarrow{\mu_j}, \Sigma_j)}$$

(Nach Bishop [1])

4. Expectation Maximization algorithm

ComTec

3. Maximization step:

   For each component k re-estimate the parameters using the current probabilities $p(k|\overrightarrow{x_n})$:

$$\omega_k^{new} = \frac{1}{N}\sum_{n=1}^{N} p(k|\overrightarrow{x_n})$$

$$\overrightarrow{\mu_k^{new}} = \frac{1}{N \cdot \omega_k^{new}}\sum_{n=1}^{N} p(k|\overrightarrow{x_n})\overrightarrow{x_n}$$

$$\Sigma_k^{new} = \frac{1}{N \cdot \omega_k^{new}}\sum_{n=1}^{N} p(k|\overrightarrow{x_n})\left(\overrightarrow{x_n} - \overrightarrow{\mu_k^{new}}\right)\left(\overrightarrow{x_n} - \overrightarrow{\mu_k^{new}}\right)^T$$

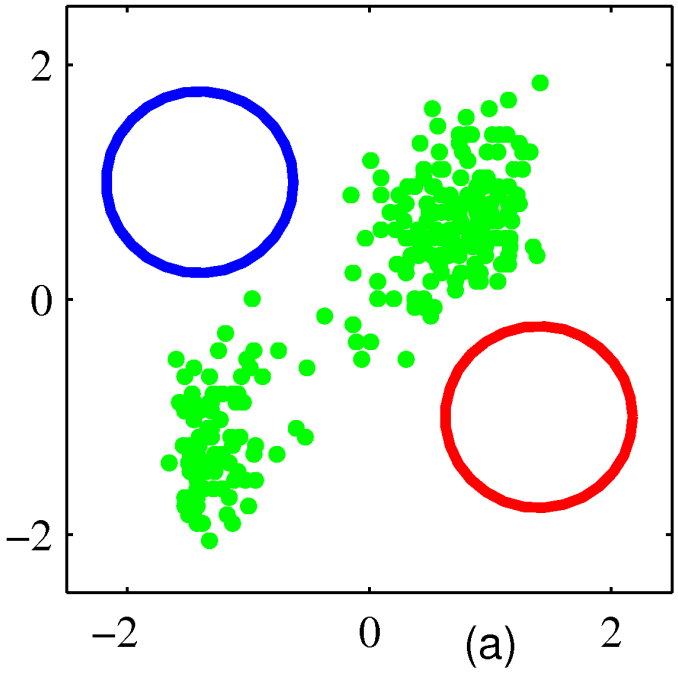
4. Evaluate the log likelihood function. Return to step 2 until the function satisfies the convergence criterion.

(Bishop [1])

4. Example - Clustering with GMM

ComTec

1. Initialize all mean vectors $\overrightarrow{\mu_k}$, covariances Σ_k and mixing coefficients ω_k. Calculate the initial value of the log likelihood function.



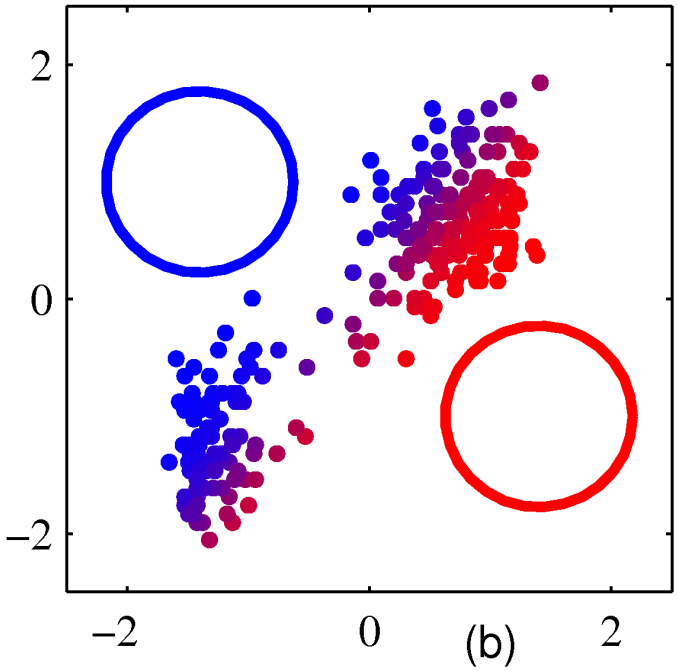
[1]

Old Faithful data set in a two-dimensional Euclidean space.

4. Example - Clustering with GMM

2. Expectation step:

Evaluate the probability $p(k|\overrightarrow{x_n})$ that x_n belongs to component k:



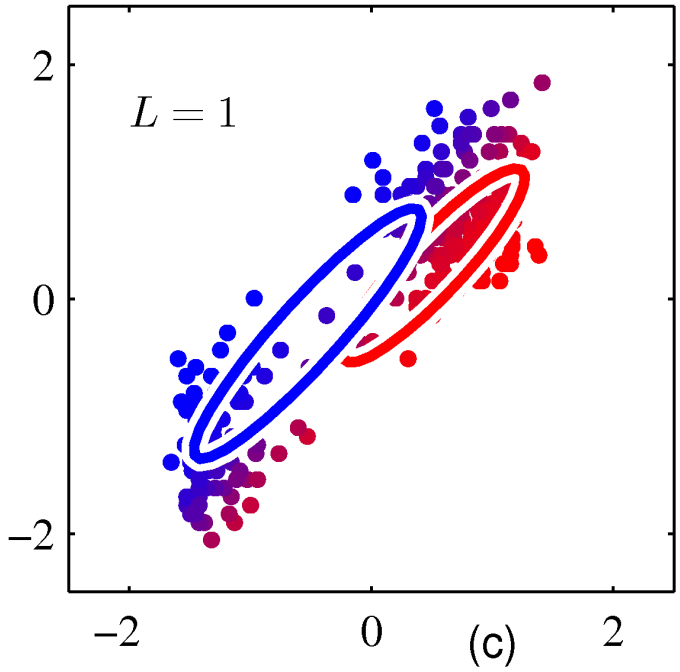
[1]

4. Example - Clustering with GMM

ComTec

3. Maximization step:

For each component k re-estimate the parameters using the current probabilities $p(k|\overrightarrow{x_n})$:

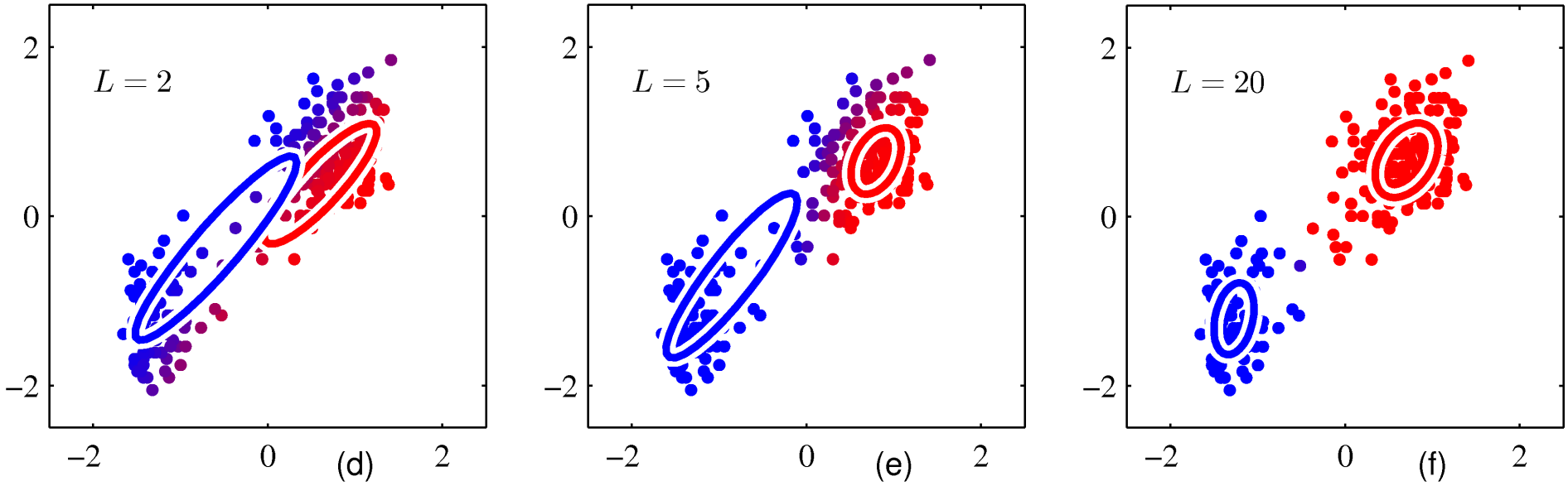


[1]

Iteration count: L

4. Example - Clustering with GMM

Results after 2 (d), 5 (e) and 20 (f) iterations of EM, respectively.



[1]

5. GMM Advantages / Disadvantages

ComTec

Advantages

- Soft clustering: soft allocation of data-points to clusters
- recognition of overlapped clusters
- depiction of complex probability density functions
- Smooth functions (filtering)
- Good representation of data

Disadvantages

- Number of clusters needs to be known
- Result of clustering is influenced by initialization of model parameters
- Problem with singularities

Referenzen

ComTec

[1] C. J. Bishop, *Pattern recognition and machine learning*, 8. corr. print. ed. (Information science and statistics). New York: Springer-Verlag, 2009.
[2] V. Lavrenko, "IAML: Mixture models and EM," S. o. I. University of Edinburgh, Ed., ed. http://www.inf.ed.ac.uk/teaching/courses/iaml/2011/slides/em.pdf, 2011.
[3] M. Behnisch, *Urban data mining: Operationalisierung der Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt*. Univ-Verlag Karlsruhe, 2008, p. 303.
[4] DIW Berlin. *Verteilung der Körpergrößen nach Geschlecht im Jahr 2006*. https://de.statista.com/statistik/daten/studie/1825/umfrage/koerpergroesse-nach-geschlecht/ (zugegriffen am 20.11.18 14:36).