# Data fitting stochastic models

## Explaining data by stochastic models

**Given** set of points $X \subset \mathbb{R}^d, |X| < \infty$.
**Find** a stochastic distribution (model, process) that explains the data well.

## Explaining data by stochastic models

**Given** set of points $X \subset \mathbb{R}^d, |X| < \infty$.
**Find** a stochastic distribution (model, process) that explains the data well.

- Impossible to solve if we do not restrict the distributions that have to be considered.

## Explaining data by stochastic models

**Given** set of points $X \subset \mathbb{R}^d, |X| < \infty$.
**Find** a stochastic distribution (model, process) that explains the data well.

- Impossible to solve if we do not restrict the distributions that have to be considered.

$\Rightarrow$ Need to fix a family of distribution in advance.

## Explaining data by stochastic models

**Given** set of points $X \subset \mathbb{R}^d, |X| < \infty$.
**Find** a stochastic distribution (model, process) that explains the data well.

- Impossible to solve if we do not restrict the distributions that have to be considered.

$\Rightarrow$ Need to fix a family of distribution in advance.

- Find a good or even best distribution from that family.

## Explaining data by stochastic models

**Given** set of points $X \subset \mathbb{R}^d, |X| < \infty$.
**Find** a stochastic distribution (model, process) that explains the data well.

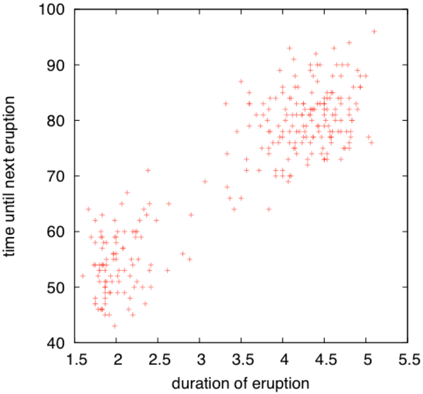
- Impossible to solve if we do not restrict the distributions that have to be considered.

$\Rightarrow$ Need to fix a family of distribution in advance.

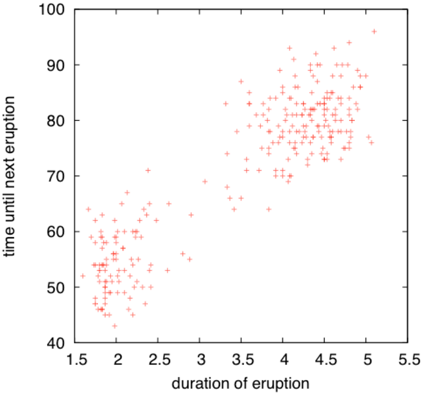
- Find a good or even best distribution from that family.
- When does a distribution explain data well?

# The Old Faithful data set

# The Old Faithful data set



# The Old Faithful data set




Is there a distribution that explains the apparent dependency between duration and time until next eruption?

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

## Families of continuous distributions

- $d \in \mathbb{N}, S \subseteq \mathbb{R}^s$ for some $s \in \mathbb{N}, |S| = \infty$
- for each $\Theta \in S$ a density function $p(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, i.e.

$$\int_{\mathbb{R}^d} p(x|\Theta)\mathrm{d}x = 1.$$

- denote family of distributions or density functions by $\{p(\cdot|\Theta)\}_{\Theta \in S}$ or simply $\{p(\cdot|\Theta)\}_\Theta$

# Families of distributions

## Families of continuous distributions

- $d \in \mathbb{N}, S \subseteq \mathbb{R}^s$ for some $s \in \mathbb{N}, |S| = \infty$
- for each $\Theta \in S$ a density function $p(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, i.e.

$$\int_{\mathbb{R}^d} p(x|\Theta)\mathrm{d}x = 1.$$

- denote family of distributions or density functions by $\{p(\cdot|\Theta)\}_{\Theta \in S}$ or simply $\{p(\cdot|\Theta)\}_{\Theta}$

## Example - univariate Gaussian distributions

- $d = 1, S = \mathbb{R} \times \mathbb{R}_{>0}, \Theta = (\mu, \sigma)$
- $p(\cdot|\Theta) = p(\cdot|\mu, \sigma) = \mathcal{N}(\cdot|\mu, \sigma^2)$

## Definition 6.1

*Let* $X \subset \mathbb{R}^d, |X| < \infty$ *and* $\{p(\cdot|\Theta)\}_{\Theta \in S}$ *a family of density functions.*

1. $p(X|\Theta) := \prod_{y \in X} p(y|\Theta)$ *is called the likelihood of* $X$ *with respect to* $p(\cdot|\Theta)$ *or simply with respect to* $\Theta$.
2. $\mathcal{L}_X(\Theta) = -\ln(p(X|\Theta)) = -\sum_{y \in X} \ln(p(y|\Theta)$ *called negative log-likelihood of* $X$ *with respect to* $\Theta$.

## Problem 6.2 (Maximum likelihood estimation)

*Given a family* $\{p(\cdot|\Theta\}_{\Theta \in S}$ *of distributions on* $\mathbb{R}^d$ *and a finite set* $X \subset \mathbb{R}^d$, *find* $\Theta_0 \in S$ *that minimizes the negative log-likelihood* $\mathcal{L}_X(\Theta)$.

## Problem 6.2 (Maximum likelihood estimation)

*Given a family* $\{p(\cdot|\Theta\}_{\Theta \in S}$ *of distributions on* $\mathbb{R}^d$ *and a finite set* $X \subset \mathbb{R}^d$, *find* $\Theta_0 \in S$ *that minimizes the negative log-likelihood* $\mathcal{L}_X(\Theta)$.

**Remarks**

- Depending on the definition of $S$ the maximum likelihood estimation problem is not well defined.

## Problem 6.2 (Maximum likelihood estimation)

*Given a family* $\{p(\cdot|\Theta\}_{\Theta \in S}$ *of distributions on* $\mathbb{R}^d$ *and a finite set* $X \subset \mathbb{R}^d$*, find* $\Theta_0 \in S$ *that minimizes the negative log-likelihood* $\mathcal{L}_X(\Theta)$*.*

### Remarks

- Depending on the definition of $S$ the maximum likelihood estimation problem is not well defined.
- In other cases, the parameters $\Theta$ that have the minimal negative log-likelihood are not very useful.

## Problem 6.2 (Maximum likelihood estimation)

*Given a family* $\{p(\cdot|\Theta\}_{\Theta \in S}$ *of distributions on* $\mathbb{R}^d$ *and a finite set* $X \subset \mathbb{R}^d$, *find* $\Theta_0 \in S$ *that minimizes the negative log-likelihood* $\mathcal{L}_X(\Theta)$.

**Remarks**

- Depending on the definition of $S$ the maximum likelihood estimation problem is not well defined.
- In other cases, the parameters $\Theta$ that have the minimal negative log-likelihood are not very useful.
- In this case, the goal is to find "useful" or "relevant" parameters $\Theta$ that model the point set $X$.

# Log-likelihood estimation for spherical Gaussians

## Theorem 6.3

*Let* $S = \mathbb{R} \times \mathbb{R}_{>0}$ *and* $p(\cdot|\mu,\sigma) = \mathcal{N}(\cdot|\mu,\sigma^2)$ *for all* $(\mu,\sigma) \in S$. *For a finite point set* $X \subset \mathbb{R}, |X| \geq 2$,

1. *for fixed* $\mu$ *the value for* $\sigma^2$ *minimizing* $\mathcal{L}_X(\mu,\sigma)$ *is given by*

$$\sigma^2 = \frac{1}{|X|} \sum_{y \in X} (y-\mu)^2,$$

2. *the parameters* $\Theta = (\mu,\sigma)$ *minimizing* $\mathcal{L}_X(\mu,\sigma)$ *are given by*

$$\mu = \frac{1}{|X|} \sum_{y \in X} y \quad \text{and} \quad \sigma^2 = \frac{1}{|X|} \sum_{y \in X} (y-\mu)^2.$$

*Consequently, given* $X$ *the optimal values for* $\mu$ *and* $\sigma$ *can be computed in time* $\mathcal{O}(|X|)$.

# Multivariate Gaussians

## Spherical Gaussian distributions

- $d$ arbitrary, fixed, $S = \mathbb{R}^d \times \mathbb{R}_{>0}, \Theta = (\mu, \sigma)$
- $\mathcal{N}(\cdot|\mu, \sigma^2) : \mathbb{R}^d \to \mathbb{R}_{>0}$

$$x \mapsto \frac{1}{(2\pi\sigma^2)^{d/2}} \cdot \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$$

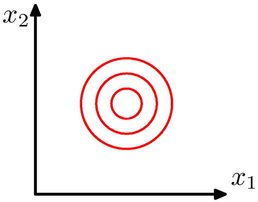
UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

## Spherical Gaussian distributions

- $d$ arbitrary, fixed, $S = \mathbb{R}^d \times \mathbb{R}_{>0}, \Theta = (\mu, \sigma)$
- $\mathcal{N}(\cdot|\mu, \sigma^2) : \mathbb{R}^d \to \mathbb{R}_{>0}$

$$x \mapsto \frac{1}{(2\pi\sigma^2)^{d/2}} \cdot \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$$

Contours of constant probability density for spherical Gaussians




(c)

## Axis-aligned Gaussian distributions

$d$ arbitrary, fixed, $S \subset \mathbb{R}^d \times \mathbb{R}^d_{>0}, \Theta = (\mu, \sigma_1, \ldots, \sigma_d)$

$$\mathcal{N}(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{>0}$$
$$x \mapsto \frac{1}{(2\pi)^{d/2}(\prod \sigma_i^2)^{1/2}} \cdot \exp\left(-\sum \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

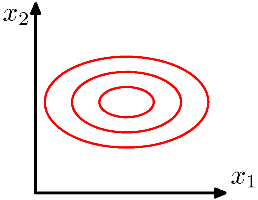
## Axis-aligned Gaussian distributions

$d$ arbitrary, fixed, $S \subset \mathbb{R}^d \times \mathbb{R}^d_{>0}, \Theta = (\mu, \sigma_1, \dots, \sigma_d)$

$$\mathcal{N}(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{>0}$$

$$x \mapsto \frac{1}{(2\pi)^{d/2}(\prod \sigma_i^2)^{1/2}} \cdot \exp\left(-\sum \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

Contours of constant probability density for axis-aligned Gaussians




(b)

## (General) Gaussian distributions

$d$ arbitrary, fixed, $S \subset \mathbb{R}^d \times \mathbb{R}^{d \times d}, \Theta = (\mu, \Sigma), \Sigma$ positive definite

$$\mathcal{N}(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{>0}$$

$$x \mapsto \frac{1}{(2\pi)^{d/2}(\det(\Sigma))^{1/2}} \cdot \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

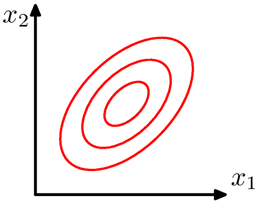
## (General) Gaussian distributions

$d$ arbitrary, fixed, $S \subset \mathbb{R}^d \times \mathbb{R}^{d \times d}, \Theta = (\mu, \Sigma), \Sigma$ positive definite

$$\mathcal{N}(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{>0}$$

$$x \mapsto \frac{1}{(2\pi)^{d/2}(\det(\Sigma))^{1/2}} \cdot \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

Contours of constant probability density for general Gaussians




(a)

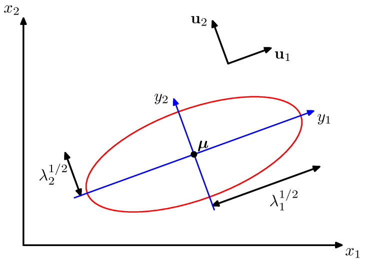
## (General) Gaussian distributions

$d$ arbitrary, fixed, $S \subset \mathbb{R}^d \times \mathbb{R}^{d \times d}, \Theta = (\mu, \Sigma), \Sigma$ positive definite

$$\mathcal{N}(\cdot|\Theta) : \mathbb{R}^d \to \mathbb{R}_{>0}$$

$$x \mapsto \frac{1}{(2\pi)^{d/2}(\det(\Sigma))^{1/2}} \cdot \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

Contour in terms of eigenvalues and eigenvectors of $\Sigma$



# Log-likelihood estimation for spherical multivariate Gaussi

## Theorem 6.4

*Let* $S = \mathbb{R}^d \times \mathbb{R}_{>0}$ *and* $p(\cdot|\mu,\sigma) = \mathcal{N}(\cdot|\mu,\sigma^2)$ *for all* $(\mu,\sigma) \in S$. *For a finite point set* $X \subset \mathbb{R}^d, |X| \geq 2$,

1. *for fixed* $\mu$ *the value for* $\sigma^2$ *minimizing* $\mathcal{L}_X(\mu,\sigma)$ *is given by*

$$\sigma^2 = \frac{1}{d|X|} \sum_{y \in X} \|y - \mu\|^2,$$

2. *the parameters* $\Theta = (\mu,\sigma)$ *minimizing* $\mathcal{L}_X(\mu,\sigma)$ *are given by*

$$\mu = \frac{1}{|X|} \sum_{y \in X} y \quad \text{and} \quad \sigma^2 = \frac{1}{d|X|} \sum_{y \in X} \|y - \mu\|^2.$$

*Consequently, given* $X$ *the optimal values for* $\mu$ *and* $\sigma$ *can be computed in time* $\mathcal{O}(|X|)$.

## Log-likelihood estimation for multivariate Gaussians

### Theorem 6.5

*Let* $d \in \mathbb{N}, S \subset \mathbb{R}^d \times \mathbb{R}_{d\times d}$, $p(\cdot|\Theta) = \mathcal{N}(\cdot|\Theta), \Theta = (\mu, \Sigma)$, $\Sigma \in \mathbb{R}^{d\times d}$ *positive definite. For a finite point set* $X \subset \mathbb{R}^d, |X| \geq 2$,

1. *for fixed* $\mu$ *the value for* $\sigma^2$ *minimizing* $\mathcal{L}_X(\mu, \sigma)$ *is given by*

$$\Sigma = \frac{1}{|X|} \sum_{y\in X} (y-\mu)\cdot(y-\mu)^T,$$

2. *the parameters* $\Theta = (\mu, \sigma)$ *minimizing* $\mathcal{L}_X(\mu, \sigma)$ *are given by*

$$\mu = \frac{1}{|X|} \sum_{y\in X} y \quad \textit{and} \quad \Sigma = \frac{1}{|X|} \sum_{y\in X} (y-\mu)\cdot(y-\mu)^T.$$

# Mixtures of Gaussians

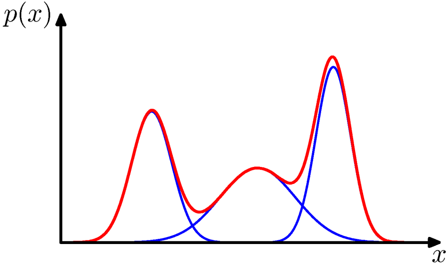
## Gaussian mixture distributions

- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k$ models for $d$-variate Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

## Gaussian mixture distributions

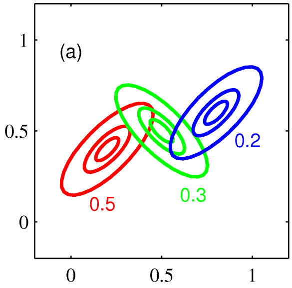
- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k$ models for $d$-variate Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

Mixture of three univariate Gaussian distributions



# Mixtures of Gaussians

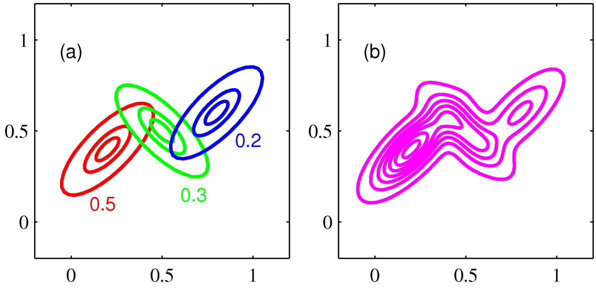
Contours of constant probability densities for three Gaussians



# Mixtures of Gaussians

Contours of constant probability densities for three Gaussians

Contours of constant probability densities for mixture of three Gaussians

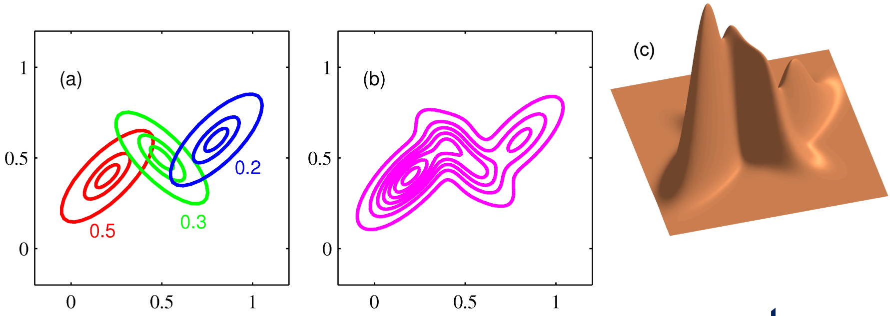


# Mixtures of Gaussians

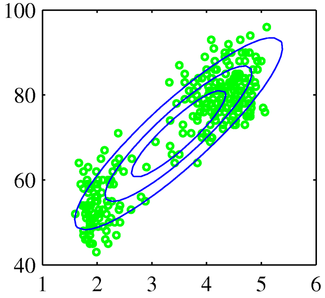
Contours of constant probability densities for three Gaussians

Contours of constant probability densities for mixture of three Gaussians

Surface plot for mixture of three Gaussians

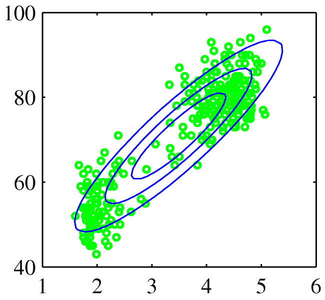


Explaining Old Faithful with a single multivariate Gaussian

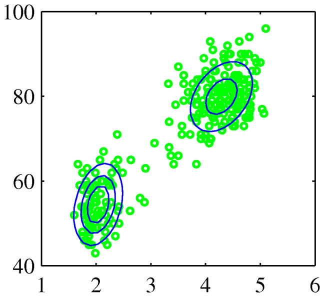


# Old Faithful and mixtures of Gaussians

Explaining Old Faithful with a single multivariate Gaussian

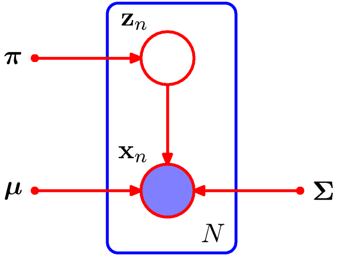



Explaining Old Faithful with a mixture of two multivariate Gaussians




UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

# Graphical representation of Gaussian mixtures




To generate a point distributed according to a mixture of Gaussians:

1. choose an index $k$ according to the distribution $\pi = (\pi_1, \ldots, \pi_K)$
2. choose a point $x$ according to the distribution $\mathcal{N}(\cdot|\Theta_k)$.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

# Likelihood for mixtures of Gaussians

## Gaussian mixture distributions

- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k$ models for $d$-variate Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

## Likelihoods

$X \subset \mathbb{R}^d, |X| = N, X = \{x_1, \ldots, x_N\}$

- $p(X|\Theta) = \prod_{n=1}^N p(x_n|\Theta) = \prod_{n=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(x_n|\Theta_k)\right)$
- $\mathcal{L}_X(\Theta) = -\ln(p(X|\Theta)) = -\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\Theta_k)\right)$

# Likelihood for mixtures of Gaussians

## Gaussian mixture distributions

- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k$ models for $d$-variate Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

## Likelihoods

$X \subset \mathbb{R}^d, |X| = N, X = \{x_1, \ldots, x_N\}$

- $p(X|\Theta) = \prod_{n=1}^N p(x_n|\Theta) = \prod_{n=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(x_n|\Theta_k)\right)$
- $\mathcal{L}_X(\Theta) = -\ln(p(X|\Theta)) = -\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\Theta_k)\right)$

# Likelihood for mixtures of spherical Gaussians

## Gaussian mixture distributions

- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k = (\mu_k, \sigma_k)$, models for $d$-variate spherical Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

# Likelihood for mixtures of spherical Gaussians

## Gaussian mixture distributions

- $d, K$ arbitrary, fixed, $\Theta = (\Theta_1, \ldots, \Theta_K, \pi), \Theta_k = (\mu_k, \sigma_k)$, models for $d$-variate spherical Gaussian distributions, $\pi \in \mathbb{R}^k_{\geq 0}, \|\pi\|_1 = 1$
- $x \mapsto \sum_k \pi_k \mathcal{N}(x|\Theta_k)$

## Likelihoods

$X \subset \mathbb{R}^d, |X| = N, X = \{x_1, \ldots, x_N\}$. Set $\mu_1 = x_1, \pi_1 \neq 0$. Then

$$\lim_{\sigma_1 \to 0} \mathcal{L}_X(\Theta) = -\infty,$$

i.e. negative log-likelihood not well-defined.

# Optimality conditions for $d = 1$

No closed formula for

$$\text{argmin}_{\Theta}\mathcal{L}_X(\theta) = \text{argmin}_{\Theta} - \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\Theta_k)\right)$$

# Optimality conditions for $d = 1$

No closed formula for

$$\text{argmin}_{\Theta}\mathcal{L}_X(\theta) = \text{argmin}_{\Theta} - \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\Theta_k)\right)$$

Taking derivatives (with Lagrange multipliers) yields

$$\mu_k = \frac{1}{R_k}\sum_{n=1}^{N} \gamma_{nk} x_n, k = 1, \ldots, K, \quad (1)$$

$$\sigma_k^2 = \frac{1}{R_k}\sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)^2, k = 1, \ldots, K, \quad (2)$$

$$\pi_k = \frac{R_k}{N}, k = 1, \ldots, K, \quad (3)$$

where $R_k = \sum_{n=1}^{N} \gamma_{nk}$, and $\gamma_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k,\sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\sigma_j)}$.

# The EM algorithm

$\mathrm{EM}(X), X = \{x_1, \ldots, x_n\}$

choose $K$ initial means, variances, and mixing coefficients $\mu_k, \sigma_k^2, \pi_k, i = 1, \ldots, K$;

**repeat**

for all $n = 1, \ldots, N, k = 1 \ldots, K$ set $\gamma_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k,\sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\sigma_j)}$;

for $k = 1, \ldots K$ set $\mu_k^{new} := \frac{1}{R_k} \sum_n \gamma_{nk} x_n$,
$\sigma_k^{2new} := \frac{1}{R_k} \sum_n \gamma_{nk} (x_n - \mu_k^{new})^2, R_k := \sum_n \gamma_{nk}, \pi_k^{new} := \frac{R_k}{N}$;

**until** *convergence*;

**return** $\mu_k, \sigma_k^2, \pi_k, k = 1, \ldots, K$

$\mathrm{EM}(X), X = \{x_1, \dots, x_n\}$

choose $K$ initial means, variances, and mixing coefficients $\mu_k, \sigma_k^2, \pi_k, i = 1, \dots, K$;

**repeat**

for all $n = 1, \dots, N, k = 1 \dots, K$ set $\gamma_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k,\sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\sigma_j)}$;

for $k = 1, \dots K$ set $\mu_k^{new} := \frac{1}{R_k} \sum_n \gamma_{nk} x_n$, $\sigma_k^{2new} := \frac{1}{R_k} \sum_n \gamma_{nk}(x_n - \mu_k^{new})^2, R_k := \sum_n \gamma_{nk}, \pi_k^{new} := \frac{R_k}{N}$;

**until** *convergence*;

**return** $\mu_k, \sigma_k^2, \pi_k, k = 1, \dots, K$

*convergence:* quality of solution no longer improves

# The EM algorithm

$\mathrm{EM}(X), X = \{x_1, \ldots, x_n\}$

choose $K$ initial means, variances, and mixing coefficients $\mu_k, \sigma_k^2, \pi_k, i = 1, \ldots, K$;

**repeat**

```
/* expectation step */
```

for all $n = 1, \ldots, N, k = 1 \ldots, K$ set $\gamma_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k,\sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\sigma_j)}$;

for $k = 1, \ldots K$ set $\mu_k^{new} := \frac{1}{R_k} \sum_n \gamma_{nk} x_n$, $\sigma_k^{2new} := \frac{1}{R_k} \sum_n \gamma_{nk}(x_n - \mu_k^{new})^2, R_k := \sum_n \gamma_{nk}, \pi_k^{new} := \frac{R_k}{N}$;

**until** *convergence*;

**return** $\mu_k, \sigma_k^2, \pi_k, k = 1, \ldots, K$

*convergence:* quality of solution no longer improves

UNIVERSITÄT PADERBORN
*Die Universität der Informationsgesellschaft*

# The EM algorithm

$\mathrm{EM}(X), X = \{x_1, \ldots, x_n\}$

choose $K$ initial means, variances, and mixing coefficients $\mu_k, \sigma_k^2, \pi_k, i = 1, \ldots, K$;

**repeat**

```
/* expectation step */
```

for all $n = 1, \ldots, N, k = 1 \ldots, K$ set $\gamma_{nk} := \frac{\pi_k \mathcal{N}(x_n | \mu_k, \sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n | \mu_j, \sigma_j)}$;

```
/* maximization step */
```

for $k = 1, \ldots K$ set $\mu_k^{new} := \frac{1}{R_k} \sum_n \gamma_{nk} x_n$, $\sigma_k^{2new} := \frac{1}{R_k} \sum_n \gamma_{nk} (x_n - \mu_k^{new})^2, R_k := \sum_n \gamma_{nk}, \pi_k^{new} := \frac{R_k}{N}$;

**until** *convergence*;

**return** $\mu_k, \sigma_k^2, \pi_k, k = 1, \ldots, K$

*convergence:* quality of solution no longer improves k-MEANS($P$)

```
choose k initial centroids c_1, ..., c_k;
repeat
    /* assignment step                                        */
    for i = 1, ..., k do
        C_i := set of points in P closest to c_i;
    end
    /* estimation step                                        */
    for i = 1, ..., k do
        c_i := c(C_i) = 1/|C_i| ∑_{p∈C_i} p;
    end
until convergence;
return c_1, ..., c_k and C_1, ..., C_k
```

Where: $C_i :=$ set of points in $P$ closest to $c_i$; $c_i := c(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} p$

# Properties of EM

- EM very popular in practice

# Properties of EM

- EM very popular in practice
- EM is reasonably efficient

# Properties of EM

- EM very popular in practice
- EM is reasonably efficient
- EM usually finds good solutions

# Properties of EM

- EM very popular in practice
- EM is reasonably efficient
- EM usually finds good solutions
- Quality of solutions depends crucially on initial solution