Speech Recognition

Prof. Dr. Andreas Wendemuth

Lehrstuhl Kognitive Systeme

Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik

Fakultät für Elektrotechnik und Informationstechnik

Otto-von-Guericke Universität Magdeburg

Production Probability modelling & estimating

- Prior / Posterior distributions
- Gaussian mixtures
- Probability densities as generative / production models
- Estimating space partitions: k-means / LBG
- Estimating class distributions: EM-method
- Problems with EM, and example

Preparation: Probability theory

Probability Theory

- X,Y: Elementary event (discrete), (x,y) continuous
- p(x): prob. density function (pdf) = P(X=x...x+dx)/dx
- P(X): discrete (cumulated) probability, $P(X) = \int_{-\infty}^{X} dy\, p(y)$
- Conditional prob. P(Y|X) = P(Y, if X was observed)
- Joint (or compound) prob. P(X,Y) = P(X|Y) P(Y) = P(Y|X) P(X)
  ➔ derive BAYES-Formula P(Y|X) = P(X|Y) P(Y) / P(X)
  for a-posteriori variables
- Independent Events: P(X,Y) = P(X) P(Y) f. all (X,Y)
- Marginalising $P(X) = \sum_i P(X|Y_i)\, P(Y_i)$, {i} complete set
  also: marginal distribution $P(X) = \sum_i P(X,Y_i)$
- BAYES-Decision rule: $Y^* = \text{argmax}_i\, P(Y_i|X)$

Priors, Joint, Posterior Prob.: 2 Examples

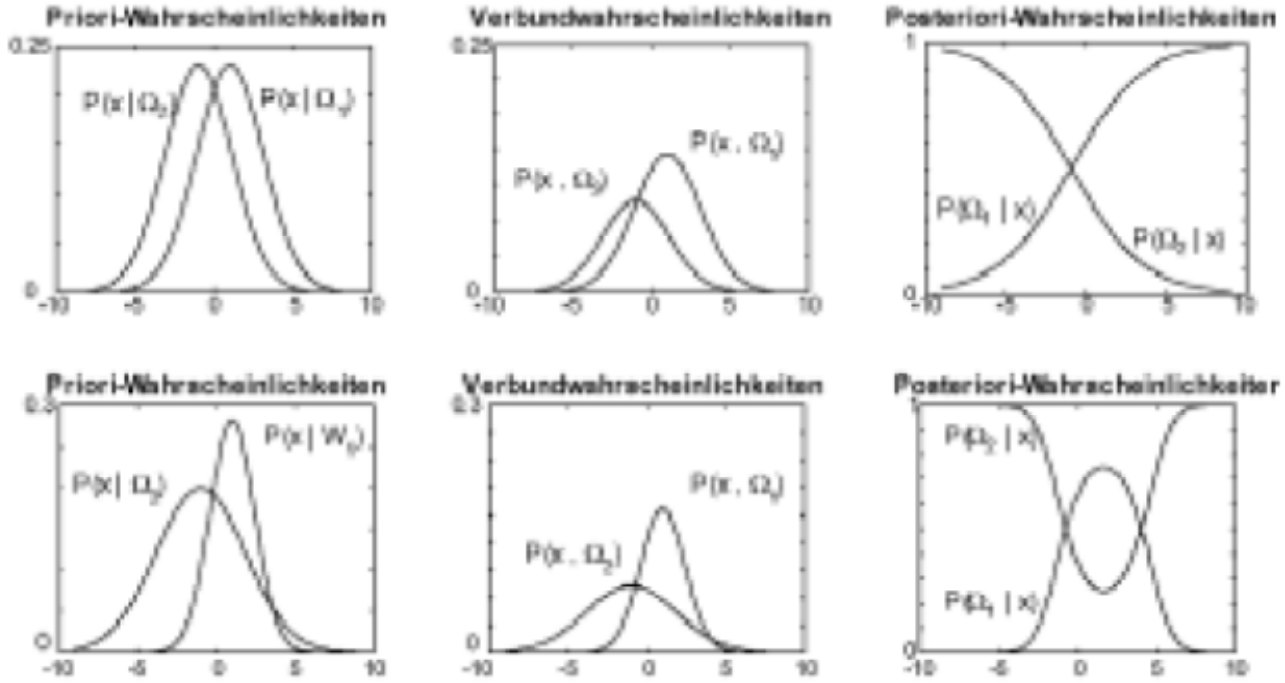


Abbildung 5.1: *Priori-/ Verbund-/ Posteriori Wahrscheinlichkeiten an 2 Beispielen*

Generative / Production Model

- Observation of speech feature in feature space originates from „causes“ (e.g. phonemes)
- *Capture* this structure by „generative (or production) model“ (e.g. normal distribution)
- *Identify* structure (e.g. recognize objects) by classification / decision

Multivariate Distributions

- multivariate distributions with parameters Θ

Multivariate (D-Dimensional) Normal Distributions

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D \det(\mathbf{C})}} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{m})\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}))$$

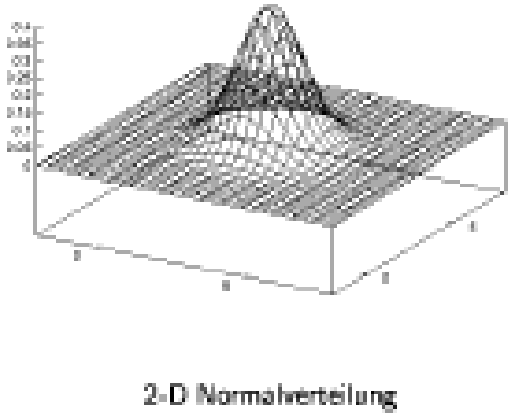
mean **m**, real symmetric matrix **C** modelling covariance of data *production (not: of data observation)*

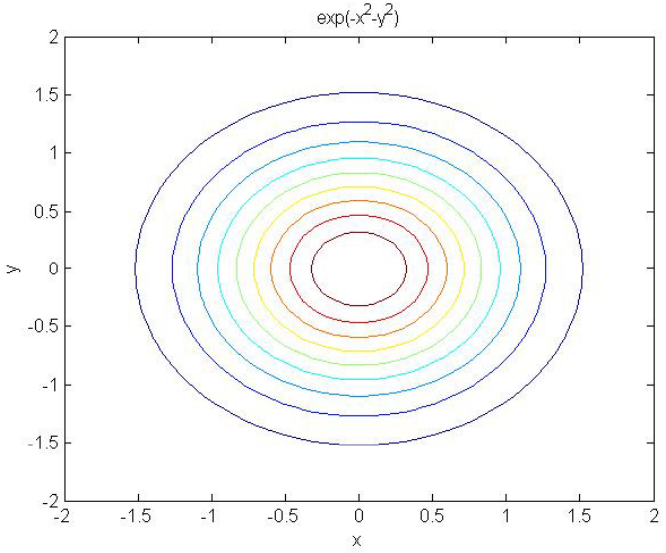
$$u_k(x) = -\log(\det(\mathbf{C}_k)) - (\mathbf{x}-\mathbf{m}_k)\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)$$

decision function $u_k(x)$ for class k, argmax_k leads to Bayes decision k*

2D-normal distribution

- 3d-plot



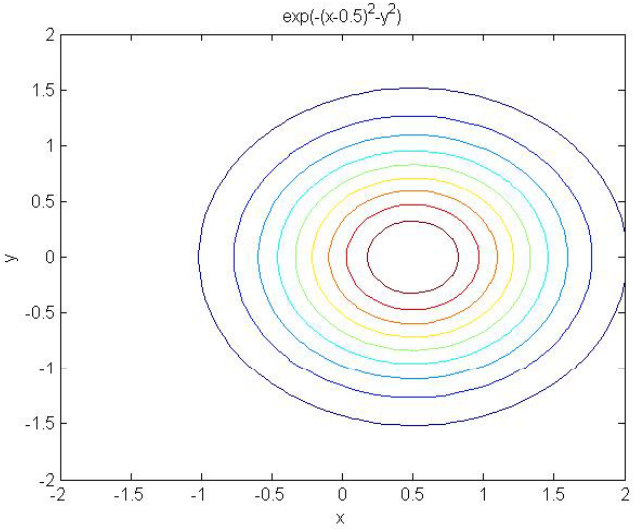


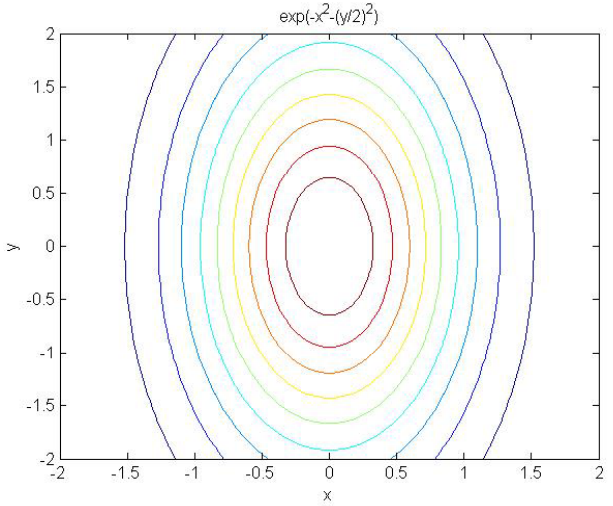
- Contourplot (red = high value)
  Matlab: ezcontour('exp(-x^2-y^2)',[-2,2,-2,2]), contours = circles

Effects on 2d-distributions

- **Shift:** ezcontour('exp(-(x-0.5)^2-y^2)',[-2,2,-2,2])]), contours = circles
- **Axes re-scaling:** ezcontour('exp(-x^2-(y/2)^2)',[-2,2,-2,2])]), contours = ellipses with radii:

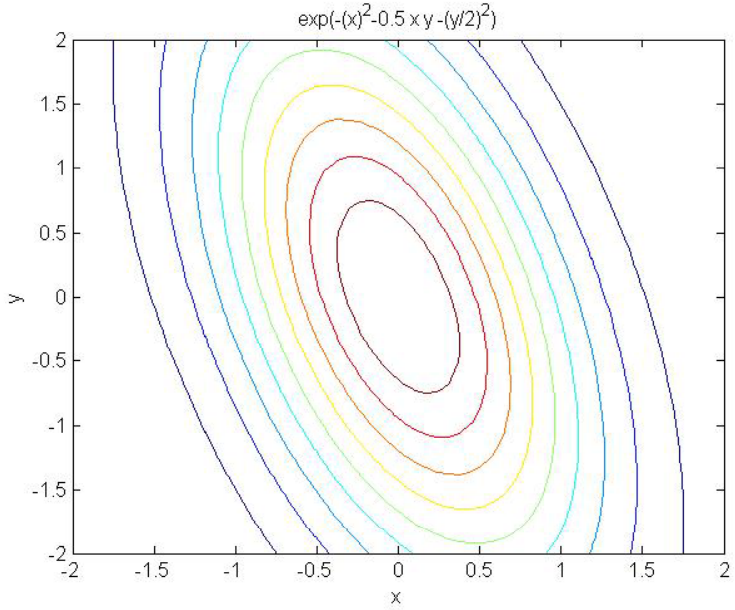
$$(x/r_1)^2 + (y/r_2)^2 = 1$$





Effects on 2d-distributions (2)

- **Rotate:** ezcontour('exp(-x^2-0.5 x y -(y/2)^2)',[-2,2,-2,2]),
  contours = rotated ellipses



$$\begin{pmatrix} x \\ y \end{pmatrix} \begin{pmatrix} ab \\ bc \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1$$

General form (shift x_0,y_0)

$$p(x+x_0, y+y_0) = c * \exp(-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} ab \\ bc \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix})$$

- c such that

$$\int dx \int dy\, p(x,y) = 1$$

Where are the axes of the ellipse?

- Solve eigenvalue problem

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_i = \lambda_i \begin{pmatrix} v \\ w \end{pmatrix}_i$$

- Decompose general vector

$$\begin{pmatrix} x \\ y \end{pmatrix} = \sum_i q_i \begin{pmatrix} v \\ w \end{pmatrix}_i$$

Axes (contd)

- Then

$$A\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ab \\ bc \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \sum_i q_i \begin{pmatrix} ab \\ bc \end{pmatrix}\begin{pmatrix} v \\ w \end{pmatrix}_i = \sum_i q_i \lambda_i \begin{pmatrix} v \\ w \end{pmatrix}_i$$

- and with orthonormalization (A symmetric)

$$\begin{pmatrix} x \\ y \end{pmatrix} A \begin{pmatrix} x \\ y \end{pmatrix} = \sum_{i,j} q_i q_j \lambda_i \begin{pmatrix} v \\ w \end{pmatrix}_i \begin{pmatrix} v \\ w \end{pmatrix}_j = \sum_i q_i^2 \lambda_i$$

Axes (3)

- This means that the axes of the ellipse are the eigenvectors of A, and the radii are $(\lambda^A_i)^{-1/2}$
- Write $C=A^{-1}$, then same axes and radii $(\lambda^C_i)^{1/2}$

$$p(\mathbf{x}+\mathbf{x}_0) = \frac{1}{\sqrt{(2\pi)^D \det(\mathbf{C})}} \exp(-\frac{1}{2}\mathbf{x}\mathbf{C}^{-1}\mathbf{x})$$

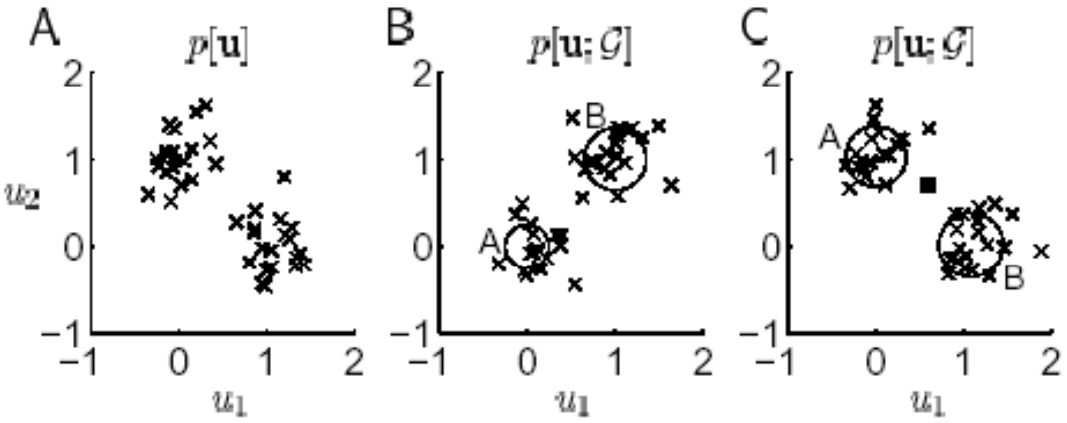
- The best *estimator* for $\mathbf{x}_0$ is the mean of the data, for $\mathbf{C}$ the correlation matrix of the data.

Model hypothesis

- Computing $\mathbf{C}$ and $\mathbf{x}_0$ from the data *assumes* that the data has been generated by $p(\mathbf{x})$.
- If this is not the case, *meaningless distributions* will be estimated (e.g. normal distributions from coin flips)

Generative Model: Gaussian Mixtures

- Let N(**x**|**m**,**S**) be a normal distribution of **x** with mean **m** and Covariance Matrix **S**.
- Then $P(\mathbf{x} \mid \Theta) = \sum_{k=1}^{K} c_k N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{S}_k)$ with $1 = \sum_{k=1}^{K} c_k$ is a Gaussian mixture model. Looks like



Gaussian Mixtures (2)

- Without any further knowledge, phonem class j produces a feature x with *prior probability* or prior distribution over causes $c_j = P(j)$.
- The *generative distribution* is the probability $P(\mathbf{x}|j, \mathbf{m}_j, \mathbf{S}_j)$ that, given phonem class j and parameters $\mathbf{m}_j, \mathbf{S}_j$ the observation was produced by a phoneme of class j.

Gaussian Mixtures (3)

- Hence the Gaussian Mixture model

$$P(\mathbf{x} \mid \boldsymbol{\Theta}) = \sum_{k=1}^{K} c_k N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{S}_k)$$

$$= \sum_{k=1}^{K} p(k) P(\mathbf{x} \mid k, \mathbf{m}_k, \mathbf{S}_k)$$

is a *sum over causes*, in statistics a marginal distribution with parameters $\boldsymbol{\Theta}$.

Gaussian mixtures (4)

- This can be written as a sum of *joint distributions* (joint: causes and generations)

$$P(\mathbf{x}, k \mid \mathbf{\Theta}_k) = p(k) P(\mathbf{x} \mid k, \mathbf{m}_k, \mathbf{S}_k)$$

as

$$P(\mathbf{x} \mid \mathbf{\Theta}) = \sum_{k=1}^{K} P(\mathbf{x}, k \mid \mathbf{\Theta}_k)$$

Recognition/decision

- We would like to know with which probability P(k|x) the feature x was produced by phonem of class k. Apply Bayes Formula:

$$P(\mathbf{x},k)=P(k)P(\mathbf{x}\,|\,k) = P(k\,|\,\mathbf{x})P(\mathbf{x})$$

hence the *posterior probabilities* are

$$P(k\,|\,\mathbf{x})=\frac{P(k)P(\mathbf{x}\,|\,k)}{P(\mathbf{x})} =\frac{P(\mathbf{x},k)}{P(\mathbf{x})}$$

and the maximum is to be selected (B‘s Rule)

Decisions

- In order to make a decision, the denominator P(**x**) is irrelevant. As an *indicator function*, we can use the likelihood (numerator) with $P(\mathbf{x},k) = p(k)P(\mathbf{x}\,|\,k,\mathbf{m}_k,\mathbf{S}_k) = c_k N(\mathbf{x}\,|\,\mathbf{m}_k,\mathbf{S}_k)$ and select the maximum. So Bayes‘ decision is equivalent to Maximum Likelihood decision.

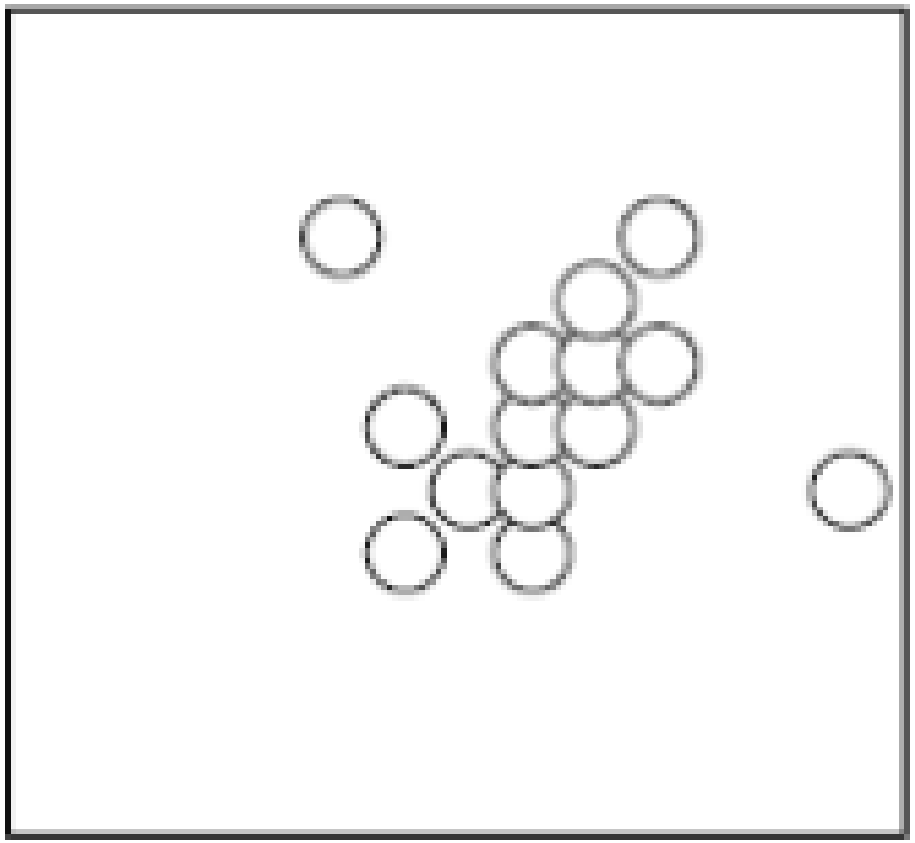
Ex 6/1

- Given are the 14 points of 2-dim. data (see plot later)
  (-2,4),(-1,-1),(-1,1),(0,0),(1,-1),(1,0),(1,1),
  (1,2),(2,1),(2,2),(2,3), (3,2),(3,4),(6,0)
- Draw the data. By inspection, assign 2 mixture densities to the data. What are the means m_k and relative weights c_k of the 2 components?
- With your values of m_k and c_k, compute the covariance matrices.
- Does a coordinate transformation exist (try inspection first!, otherwise calculate) which makes the covariance matrices diagonal? Without computing covariance matrices: what is the geometrical justification for such a transformation?

Ex 6/2

- What is the probability (formula only) that a data point is generated anywhere in $(5 \pm 1, 0 \pm 1)$?
- Approximate that probability, using the Gaussian Mixture probability density at (5,0), and the value of the inspected area.
- What are the probabilities (formula only) that a data point anywhere in $(5 \pm 1, 0 \pm 1)$ is generated by mixture component 1 (2)?
- Approximate these probabilities, acting as above.
- What is the likeliest cause for that data point?

Data plot – watch for Gaussians..



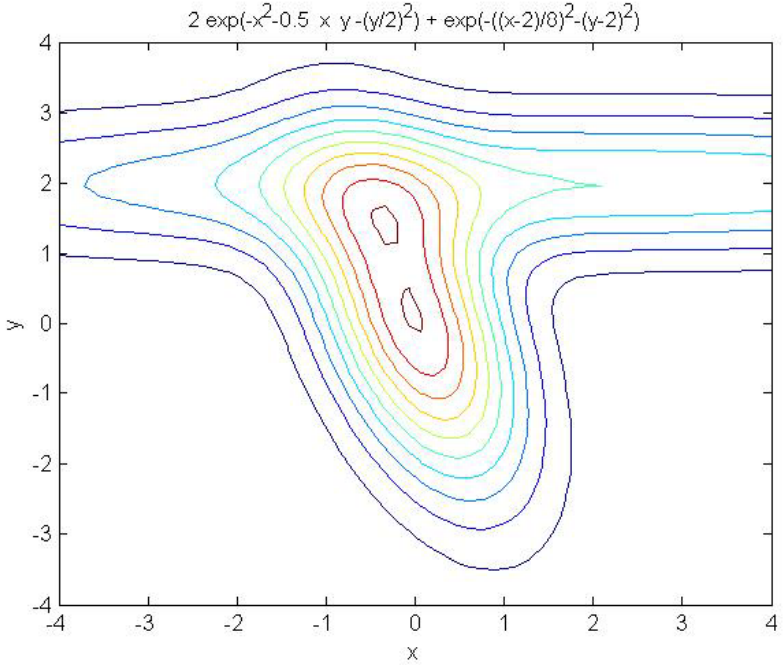
Resumé

- Raw data, if produced by causes, is compactly and meaningfully captured in a generative / production model.
- Gaussian Mixture densities can serve as good pdf‘s for generative models.
- If classification only is required, space partitioning methods such as k-means are useful.

Gaussian mixtures

- Recall $$P(\mathbf{x} \mid \Theta) = \sum_{k=1}^{K} c_k N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{C}_k)$$

- ezcontour('2*exp(-x^2-0.5* x *y -(y/2)^2) + exp(-((x-2)/8)^2-(y-2)^2)', [-4,4,-4,4])



Clustering

- Gaussian Mixtures effectively provide a clustering of data into the mixture components.
- To find the parameters c_k, $\mathbf{m_k}$, $\mathbf{C_k}$, we need some training algorithm.
- First, we look at a simplified version called the K-means algorithm

K-means

- Aka: Linde-Buzo-Gray Algorithm / Vector Quantization
- Want to find *regions* (not: probability distributions) for K clusters in an unsupervised fashion.
  Algorithm:
- Randomly arrange K means about global mean
- Repeat (until fluctuations become low):
  - Define regions by smallest distance to means (M)
  - Compute new means from all points in region (E)

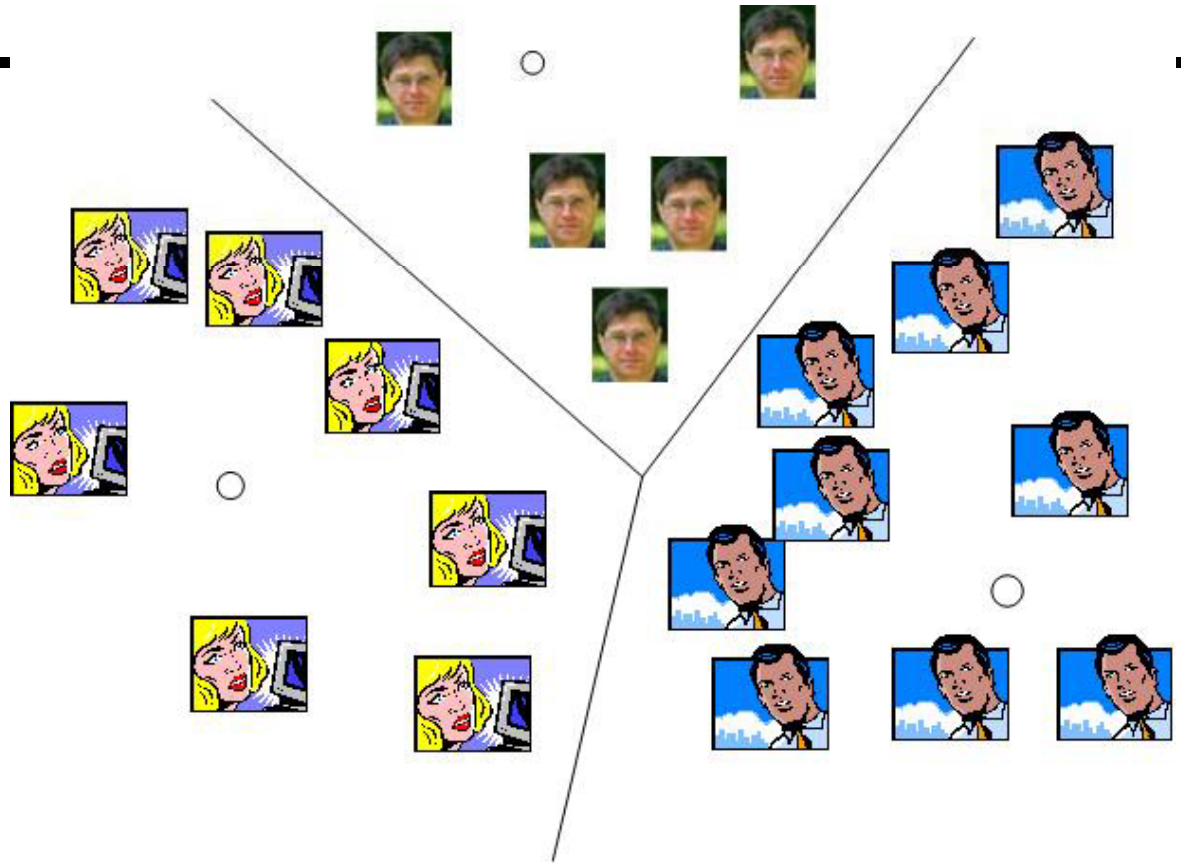
Expectation – Maximization (EM-Algorithmus)

- K-Means is the simplest variant for EM
  - (E)stimate means with *known region labels for data*
  - (M)aximise overall likelihood of algorithm by assigning data to regions (posterior) with *known means*
- Decision boundaries are bisections between means

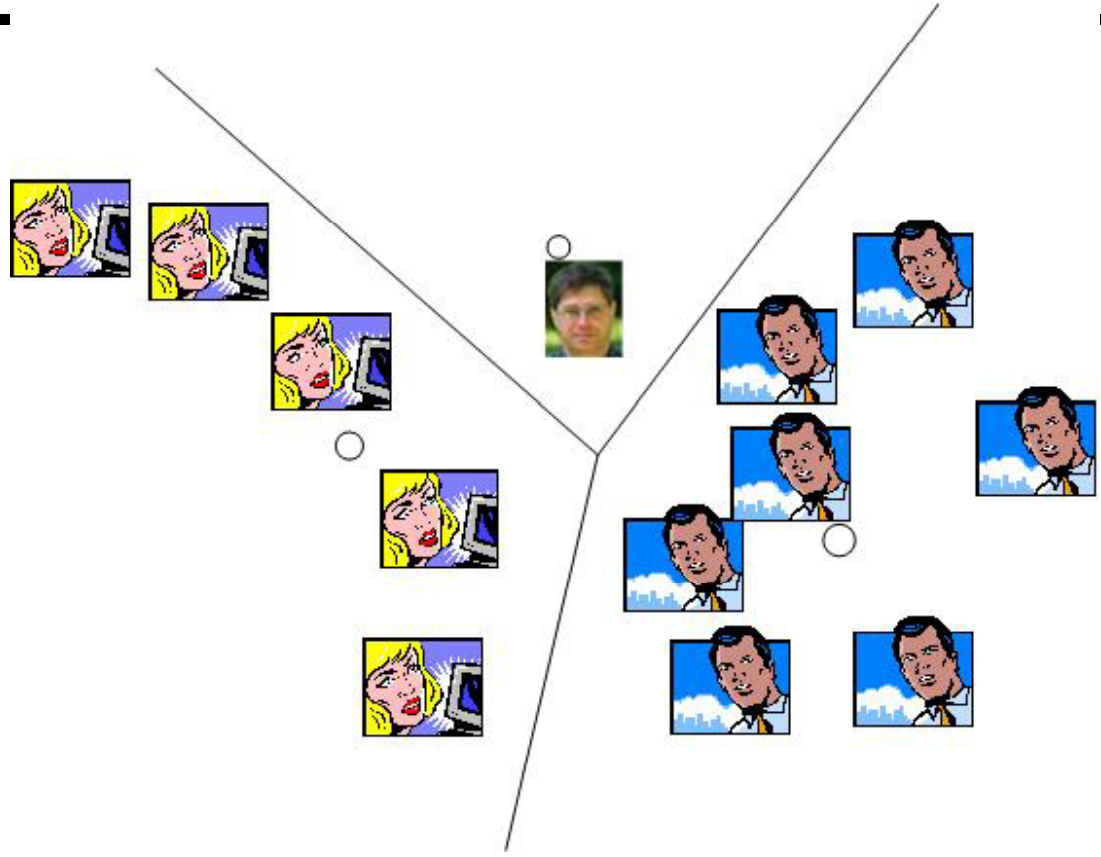
Example distributions

- In the following 2-dimensional example distributions, the class label, in other words the cause for the data, is given visually.
- Note that the observers (and the algorithms) have no idea about these causes.
- [If they had, this would be supervised learning.]
- We can use the causes to see whether our (the algorithms) decision was correct.

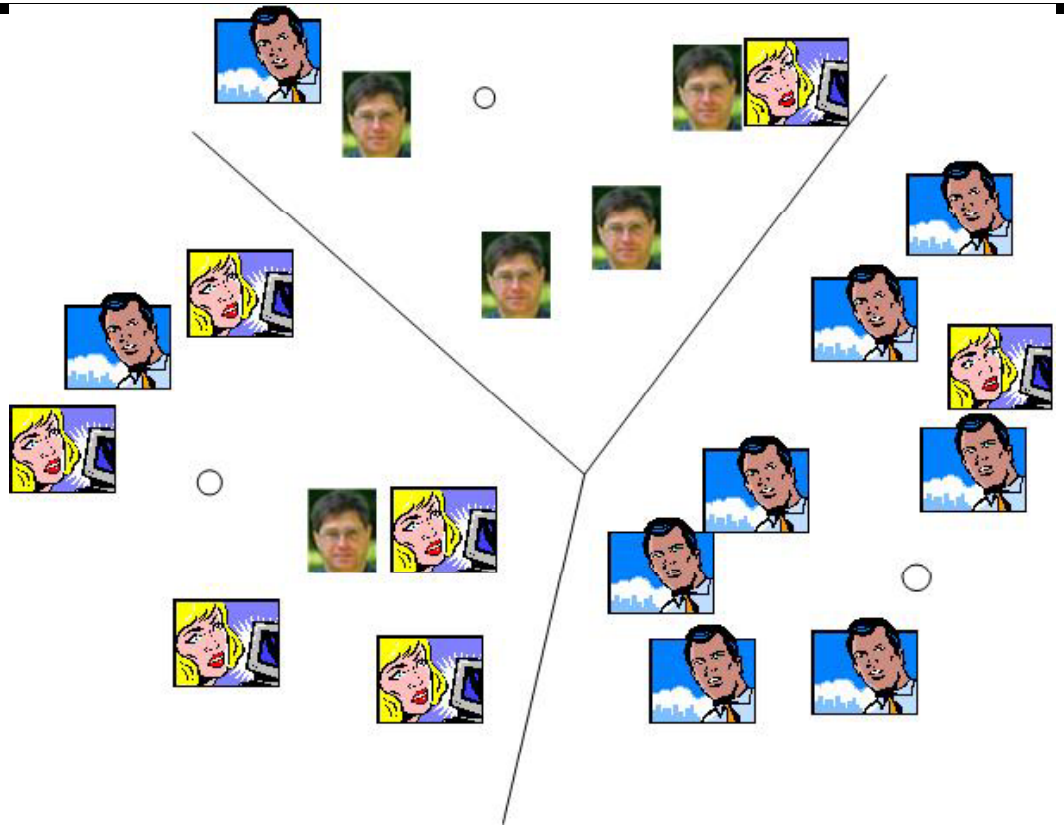
End result of k-means (dots = means)



End result of k-means with unequal priors



End result of k-means with overlapping distributions



Problems with k-means

- Strict decision is enforced, i.e. data i was created by cause k or not, i.e. with probability 1 or 0.
- K-means does not distinguish between tightly clustered and widespread data.
- K-means does not distinguish between „strong“ causes which produce many data and weak ones.
- Hence, we require a method where all data i „belong“ to all causes k with probability γ_{ik}, and where the strength of causes and data spread will be identified.

Expand K-means to finding normal distributions

Instead of regions, estimate p(k) $N(x|m_k,C_k)$ for classes k:

- (E)stimate priors, means and covariance matrices with *known posteriors* γ_{ik} *for data:*

$$p(k)=\frac{\sum_{i=1}^{N}\gamma_{ik}}{N},\mathbf{m}^k=\frac{\sum_{i=1}^{N}\gamma_{ik}\mathbf{x}^i}{\sum_{i=1}^{N}\gamma_{ik}},C_{pq}^k=\frac{\sum_{i=1}^{N}\gamma_{ik}(x_p^i-m_p^k)(x_q^i-m_q^k)}{\sum_{i=1}^{N}\gamma_{ik}}$$

- (M)aximise overall likelihood of algorithm by assigning data to regions (posteriors) with *known* m_k,C_k:

$$\gamma_{ik}=p(k\mid\mathbf{x}^i)=p(k)N(\mathbf{x}^i\mid\mathbf{m}_k,\mathbf{C}_k)/p(\mathbf{x}^i)$$

... finding normal distributions

- All data i „belong“ to all classes k with γ_{ik}
- Extreme case (decision-controlled / supervised): γ_{ik} =0,1
- Class priors p(k) do matter (other than in LBG), see

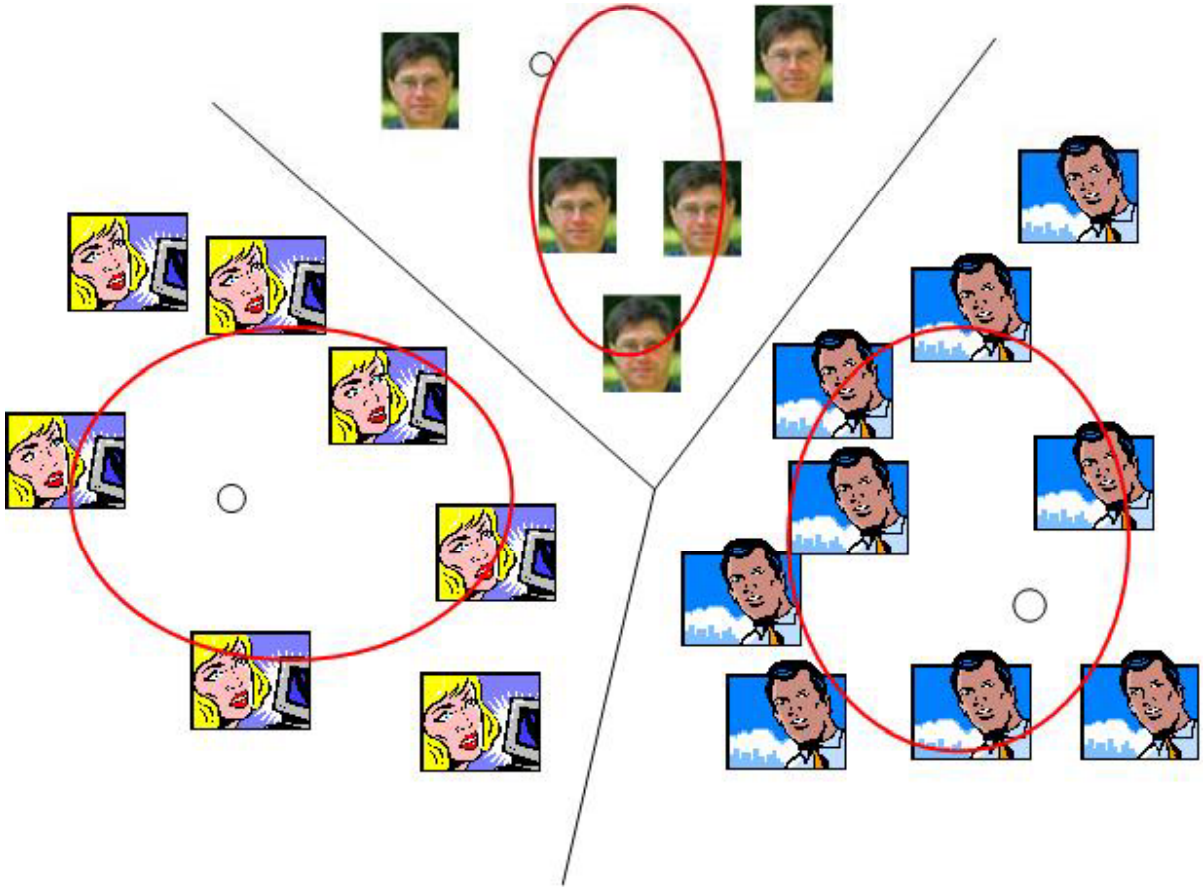
$$\gamma_{ik} = p(k \mid \mathbf{x}^i) = p(k) N(\mathbf{x}^i \mid \mathbf{m}_k, \mathbf{C}_k) / p(\mathbf{x}^i)$$

- $p(\mathbf{x}^i)$ is just normalisaton and independent of k.
- this results in computing the *likelihood*

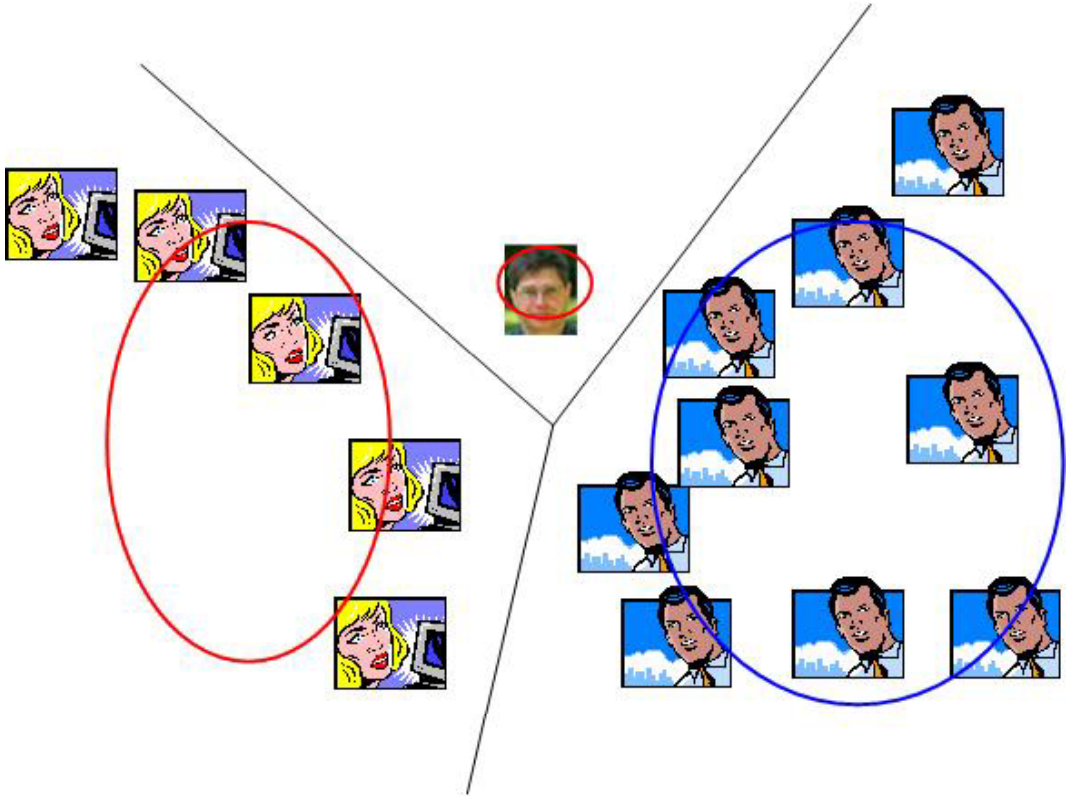
$$p(k) N(\mathbf{x}^i \mid \mathbf{m}_k, \mathbf{C}_k)$$

Final: normal distributions

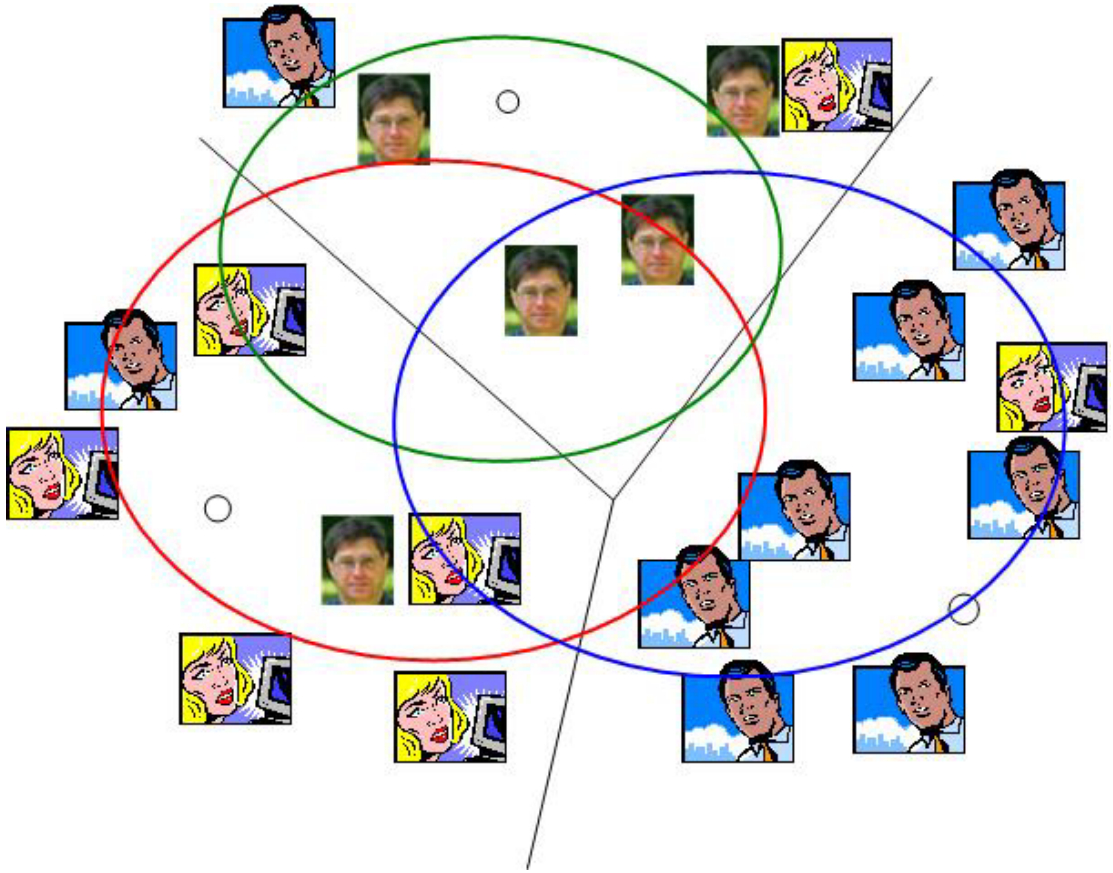
for comparison: LBG-limits



... Unequal priors



Overlapping distributions



Decision surfaces {**x**}(as in LBG) ...

... when class posteriors p(k|x) = p(l|x) > p(q|x) (q ≠k,l):

$$p(k)N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{C}_k) = p(l)N(\mathbf{x} \mid \mathbf{m}_l, \mathbf{C}_l) > p(q)N(\mathbf{x} \mid \mathbf{m}_q, \mathbf{C}_q)$$

- Take logarithms: quadratic decision surface {**x**}:

$$\log(p(k)/\sqrt{(\det(\mathbf{C}_k))} - 0.5(\mathbf{x}-\mathbf{m}_k)\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k) =$$

$$\log(p(l)/\sqrt{(\det(\mathbf{C}_l))} - 0.5(\mathbf{x}-\mathbf{m}_l)\mathbf{C}_l^{-1}(\mathbf{x}-\mathbf{m}_l)$$

- When Covariances equal, linear decision surf. {**x**}:

$$\log(p(k)/p(l)) + 0.5(\mathbf{m}_k^2 - \mathbf{m}_l^2) + (\mathbf{m}_k - \mathbf{m}_l)\mathbf{C}^{-1}\mathbf{x} = 0$$

Properties of decision surfaces {**x**} ...

- quadratic decision surface {**x**} can be ellipsoid (cigar), paraboloid (vase), hyperboloid (bowl):

$$\log(p(k)/\sqrt{(\det(\mathbf{C}_k))} - 0.5(\mathbf{x}-\mathbf{m}_k)\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k) =$$

$$\log(p(l)/\sqrt{(\det(\mathbf{C}_l))} - 0.5(\mathbf{x}-\mathbf{m}_l)\mathbf{C}_l^{-1}(\mathbf{x}-\mathbf{m}_l)$$

- Priors and Covariances matter, as they should (previous slides): cf. LBG, where only means matter!
- If distributions overlap (previous slides), still decision surfaces are safely found

Special decision surfaces {**x**} ...

- When Covariances equal, linear decision surf. {**x**}:

$$\log(p(k)/p(l)) + 0.5(\mathbf{m}_k^2 - \mathbf{m}_l^2) + (\mathbf{m}_k - \mathbf{m}_l)\mathbf{C}^{-1}\mathbf{x} = 0$$

- When also p(k) = p(l):

$$\mathbf{C}^{-1}\mathbf{x} = 0.5(\mathbf{m}_k + \mathbf{m}_l) + \text{component orthogonal to } (\mathbf{m}_k - \mathbf{m}_l)$$

- This is *equivalent* to LBG (bisection of means) with a metric (distance measure) **C**
- If **C=I** (Euclidian metric), then *equal* to LBG.

$$\mathbf{x} = 0.5(\mathbf{m}_k + \mathbf{m}_l) + \text{component orthogonal to } (\mathbf{m}_k - \mathbf{m}_l)$$

Convergence problems with EM

- Only locally optimal Parameters θ
- Cyclically-alternating Likelihood L_{EM}
- L_{EM} not bounded (nullvariate Gaussian pdfs) :

$$L_{EM} \sim \log P(y|\theta) = \log [1/\sigma \exp ((y-\mu)/\sigma)^2]$$
$$= - \log \sigma + ((y-\mu)/\sigma)^2 \dashrightarrow \infty \mid \sigma \to 0, y \to \mu$$

Convergence problems: solutions

- Minimal Number of Observations per Mixtures Density
- „Flooring“ = Minimum for Variance
- „Tying“ = common covariance matrices of the following type(leads to higher number of observations per density = robust estimation):

  Covariance matrix for all classes: $C^k = \sum_a c_a^k v_a \cdot v_a^{'T}$

  with Scalars c_a^k and Vectors (class independent!) v_a
  i.e. only estimate the scalars!
- Interpretation of observations as (uncertain) distributions, allows incorporation of a-priori-knowledge about uncertainty/confidence of observations [Wendemuth 2001]

Reference [Wendemuth 2001]

- A. Wendemuth: "Modeling Uncertainty of Data Observation". Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, Utah, USA , vol. 1, session Speech P12.1, pp. 296-299 (2001).

EM* (decision controlled)

- (1) Calculate most probable values for hidden variables u:

  $u^{(\text{Step i})} = \text{argmax}(u)\ \ P(x,u|\ \theta^{(\text{Step i-1})})$

- (2) Compute ML-estimates for complete data $(x, u^{(\text{Step i})})$:

  $\theta^{(\text{Step i})} = \text{argmax}(\theta)\ \ P(x,u^{(\text{Step i})}|\ \theta)$

  L^*_{EM} increases monotonically!

Applications

- Suppose we are at position x in a dark room where now k light bulbs at positions $\mathbf{m}_k$ are being switched on, each with power p(k) and light emitting spacial characteristics $\mathbf{C}_k$.
- p(k) is the prior, p(x|k) the light emitted from bulb k to position x, e.g. $p(x \mid k) = N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{C}_k)$
- All we know is k, the number of bulbs. We do not know their positions and characteristics. Our information is sensory, i.e. we cannot go to the bulbs and simply „look up“.

Applications (ctd...)

- We observe at our position x the sum of light from all bulbs, i.e. $P(\mathbf{x} \mid \Theta) = \sum_{k=1}^{K} p(k) N(\mathbf{x} \mid \mathbf{m}_k, \mathbf{C}_k)$
- How do we find the parameters $\Theta = \{p(k), \mathbf{m}_k, \mathbf{C}_k\}$?
- Applying LBG k-means will not help since Priors and Covariances cannot be found.
- Solution: measure at many positions x_i and apply EM for normal distributions!
- Now we know the positions of the light bulbs, their power and their emitting characteristics!

Ex 7

- Given the data points in ex 6/1 (previous lecture), perform a K-means clustering with 2 classes. You can do this by drawing and/or by computing.
- Does that clustering give the same most likely cause as obtained in ex 6/2?
- On the same data, perform EM for Gaussian mixtures with 2 components.
- Explain the difference between discriminative clustering (k-means) and maximum likelihood clustering (using a pdf).

Resumé

- Gaussian Mixture densities can serve as good pdf‘s for generative models.
- LBG / K-means is useful for unsupervised clustering, however metric **C** and priors p(k) ignored
- Gaussian Mixture densities can be estimated by EM methods
- EM will converge locally, maximizing a bound of the maximum likelihood.