Machine Learning Methods in Visualisation for Big Data

Daniel Archambault[1] Ian Nabney[2]
Jaakko Peltonen[3]

1Swansea University

2Aston University

3Aalto University

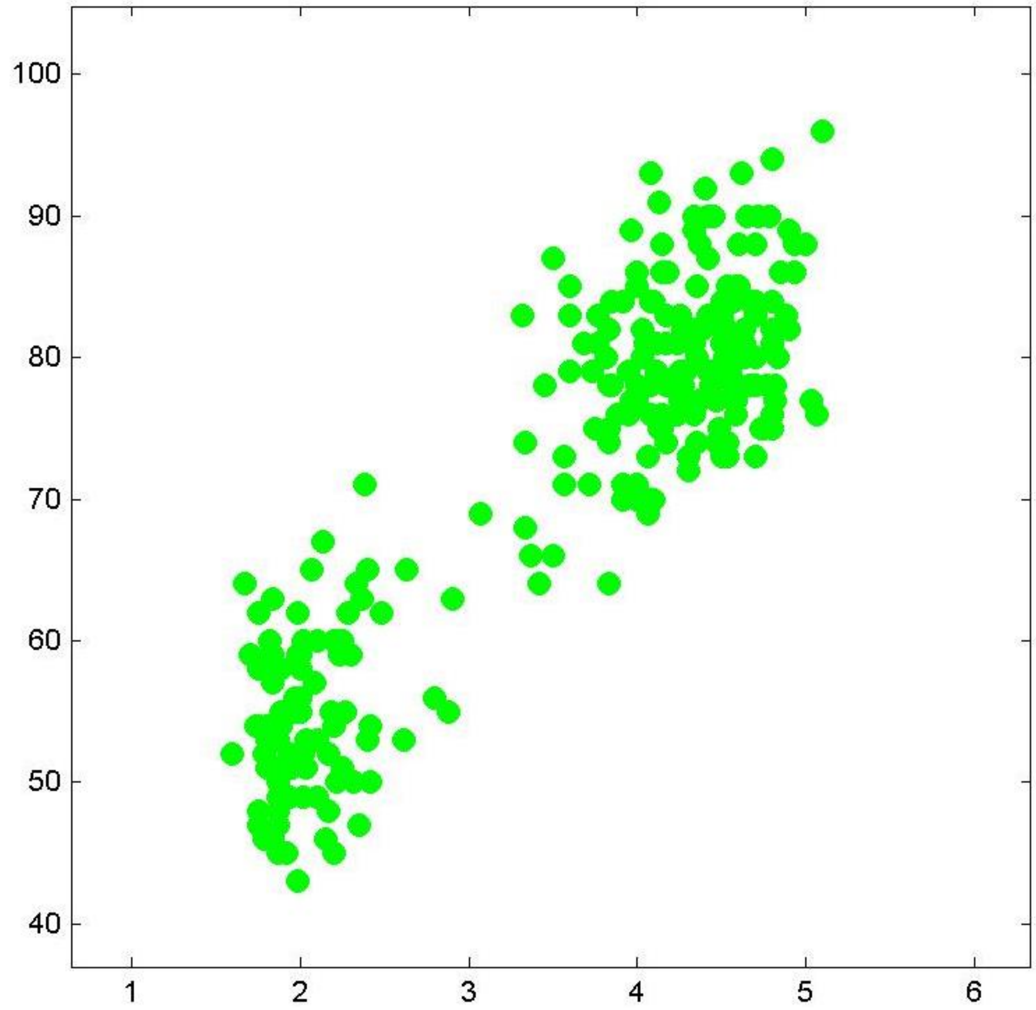
Outline

- K-means clustering
- Gaussian mixtures
- Maximum likelihood and EM
- Latent variables: EM revisited
- Bayesian Mixtures of Gaussians

Old Faithful

Old Faithful Data Set

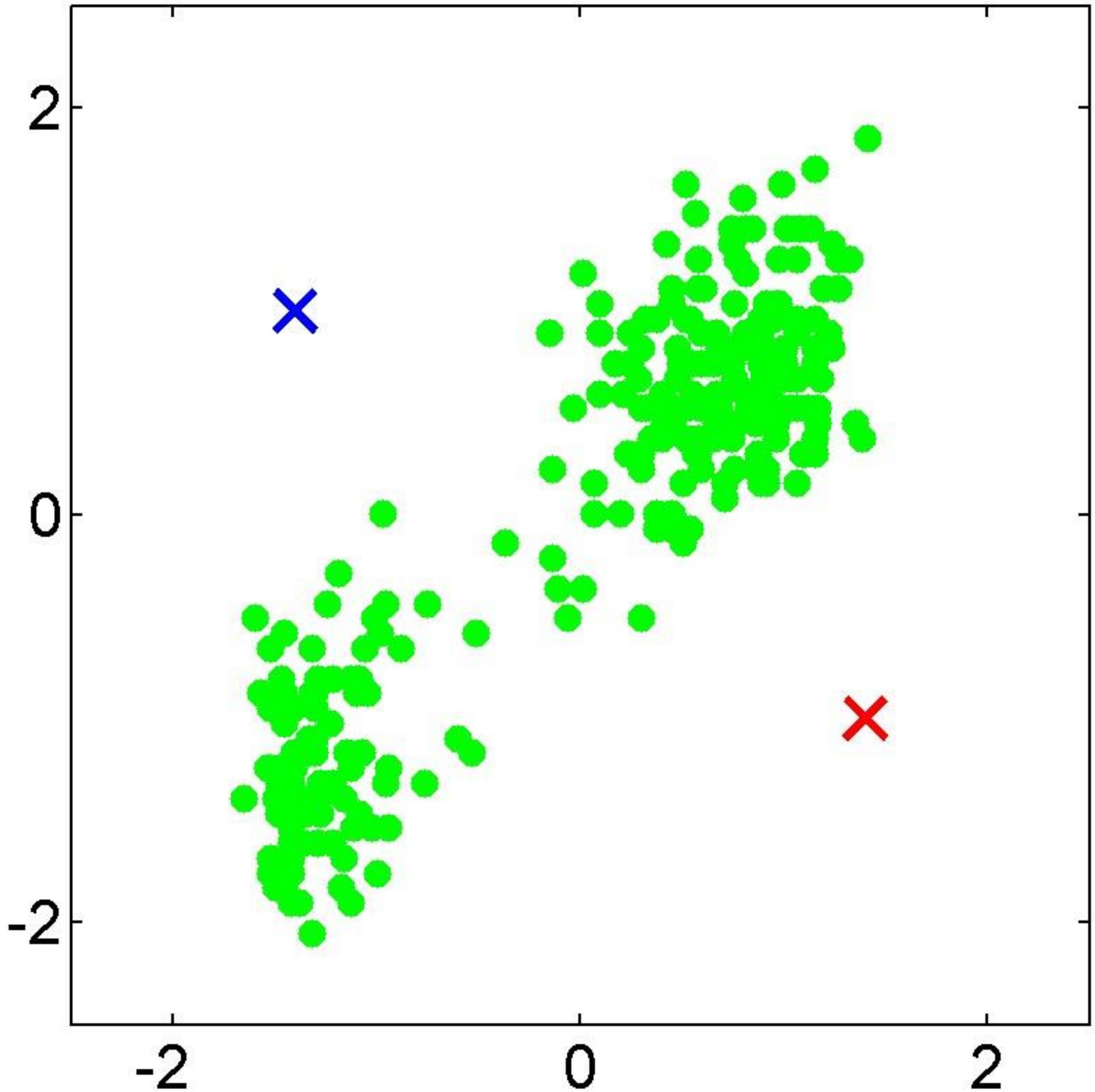
Time between eruptions (minutes)

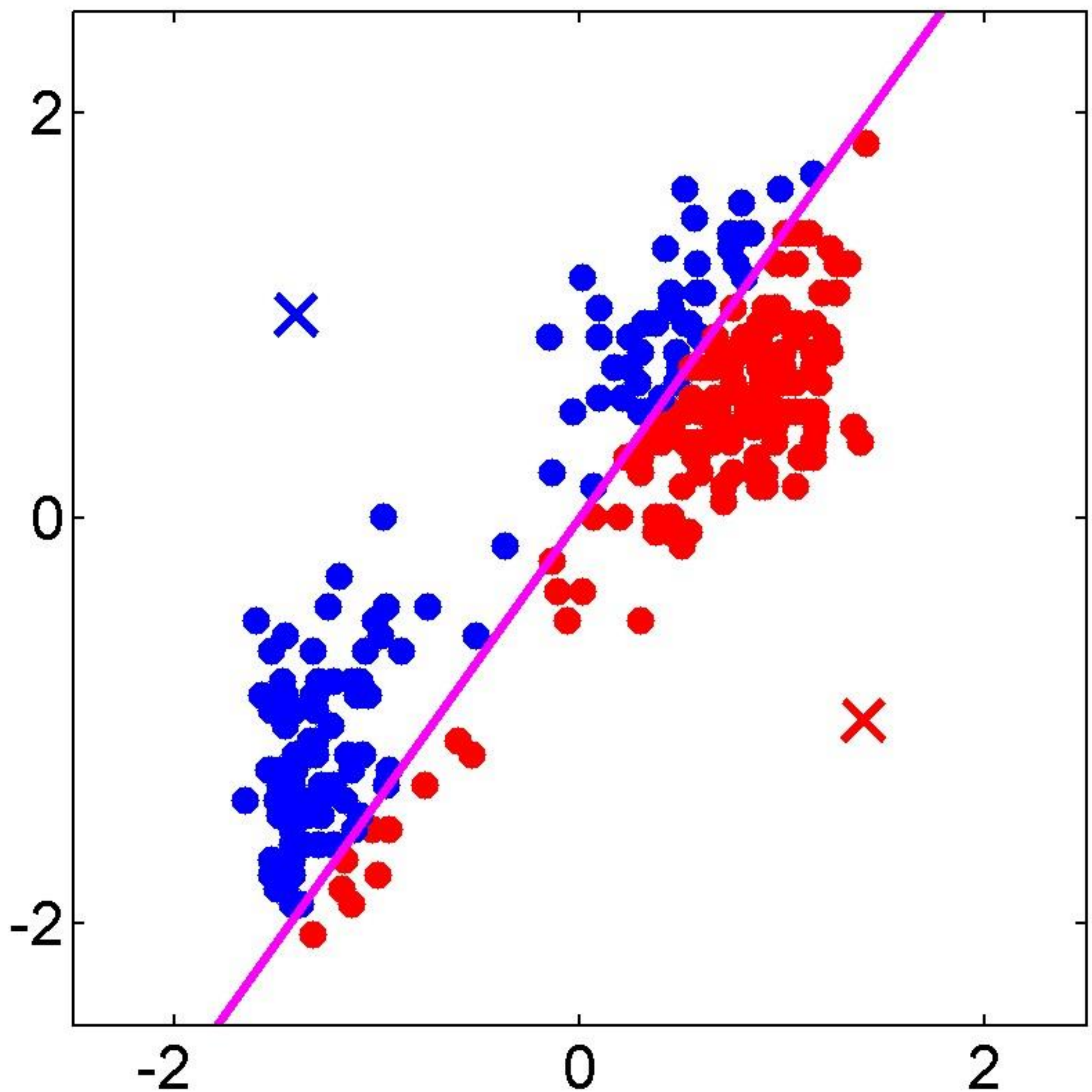


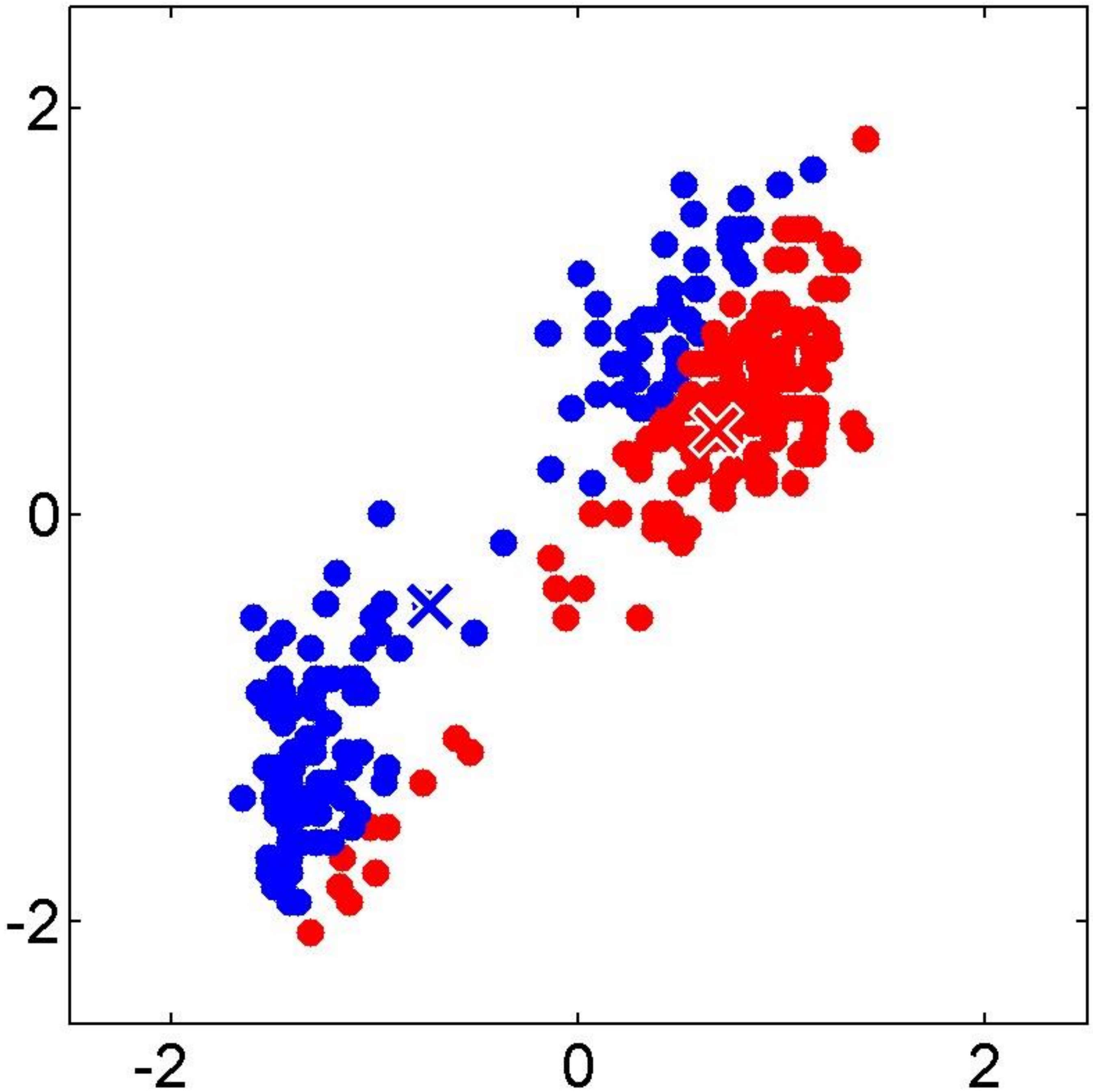
Duration of eruption (minutes)

K-means Algorithm

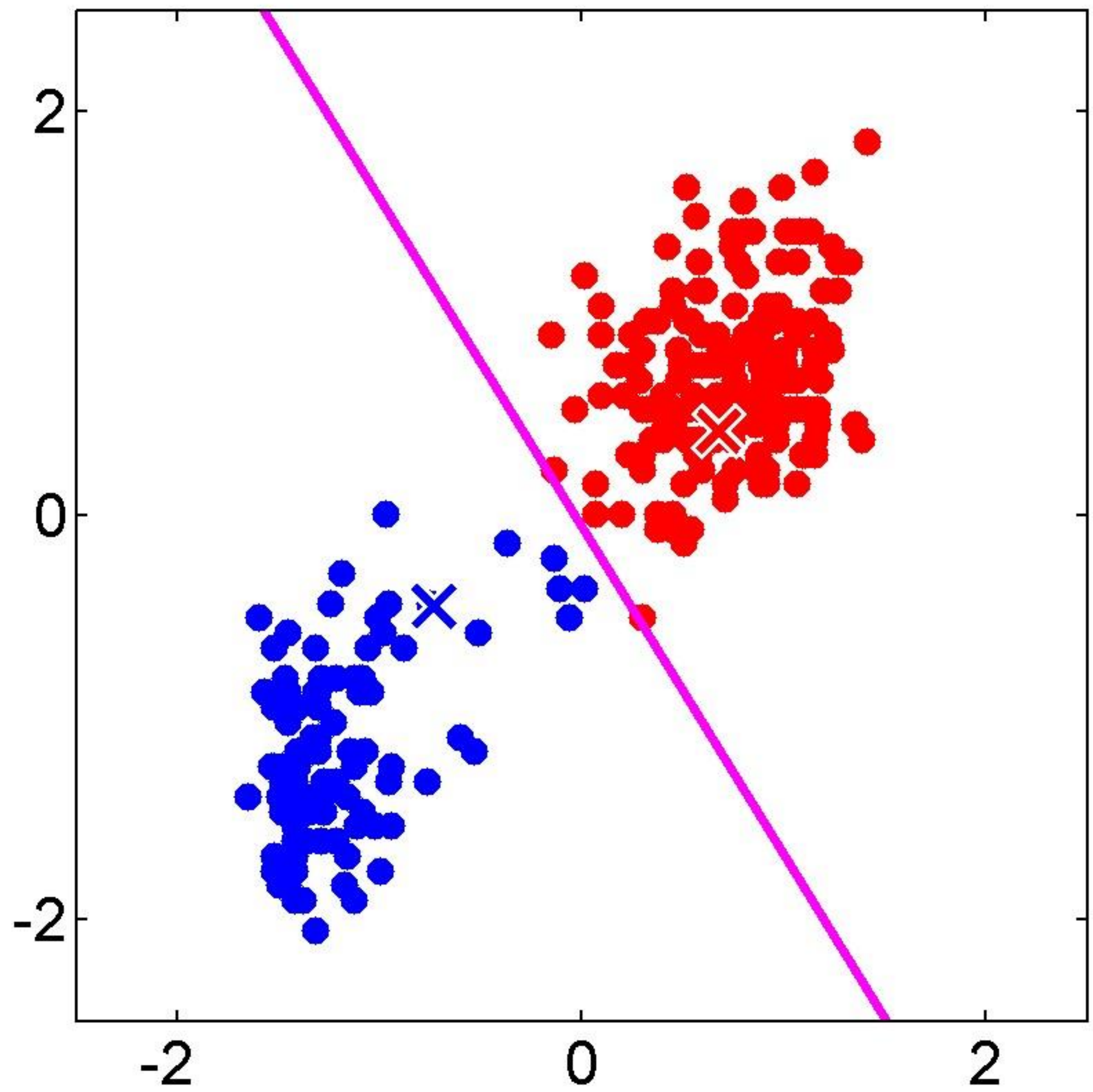
- Goal: represent a data set in terms of K clusters each of which is summarized by a prototype
- Initialize prototypes, then iterate between two phases:
  - E-step: assign each data point to nearest prototype
  - M-step: update prototypes to be the cluster means
- Simplest version is based on Euclidean distance
  - re-scale Old Faithful data

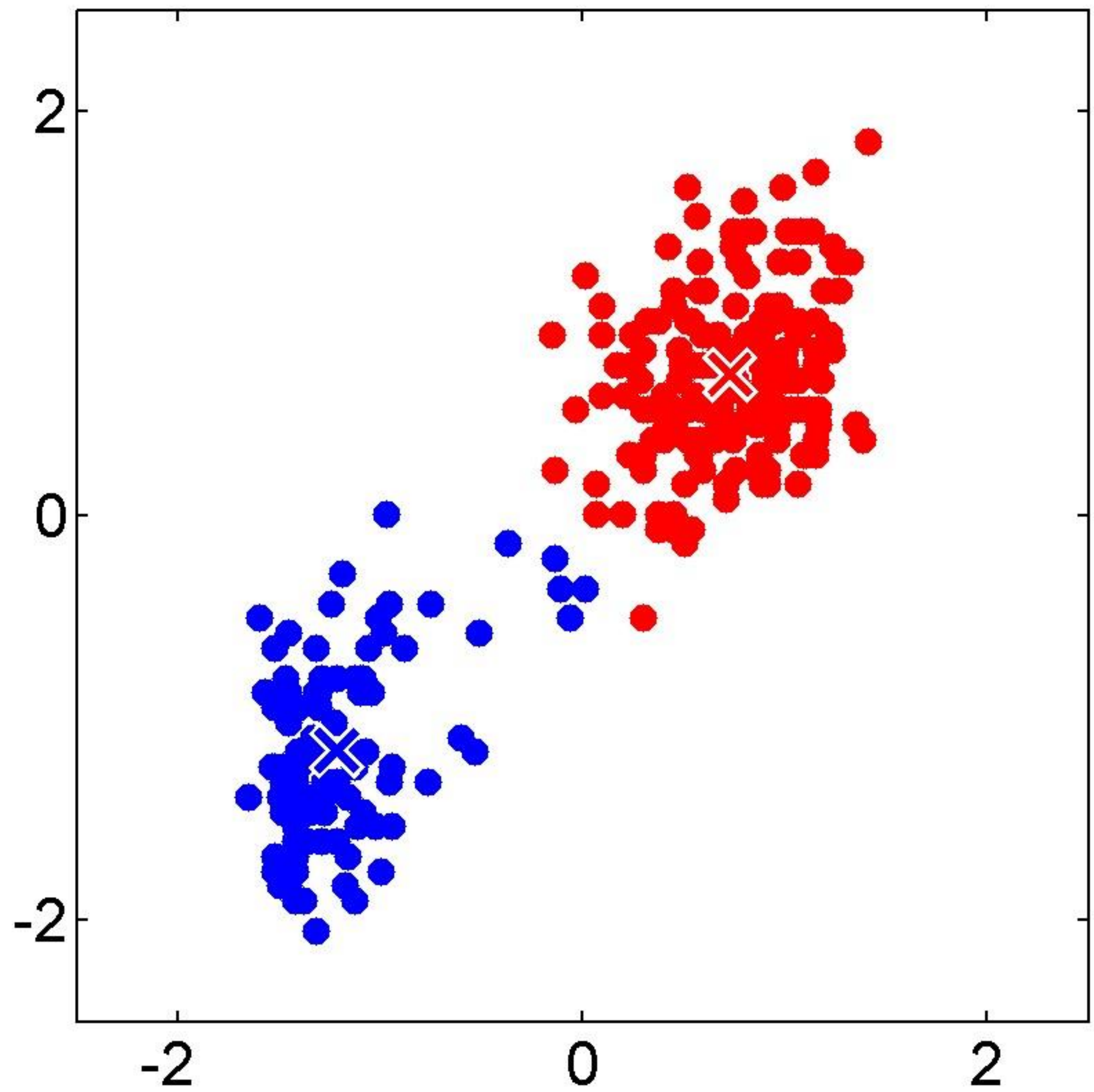

2
0
-2
-2
0
2


2
0
-2
-2
0
2

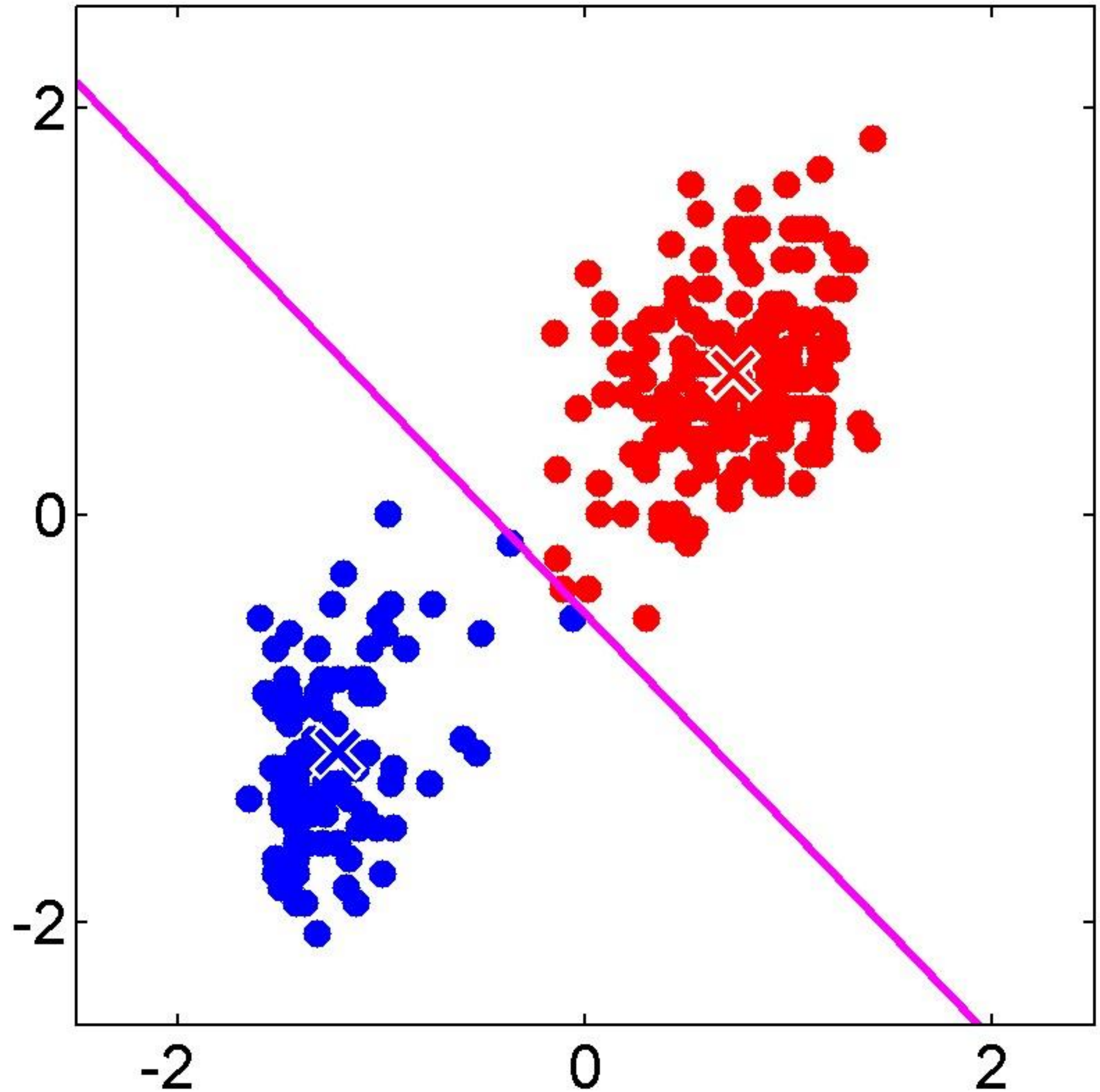


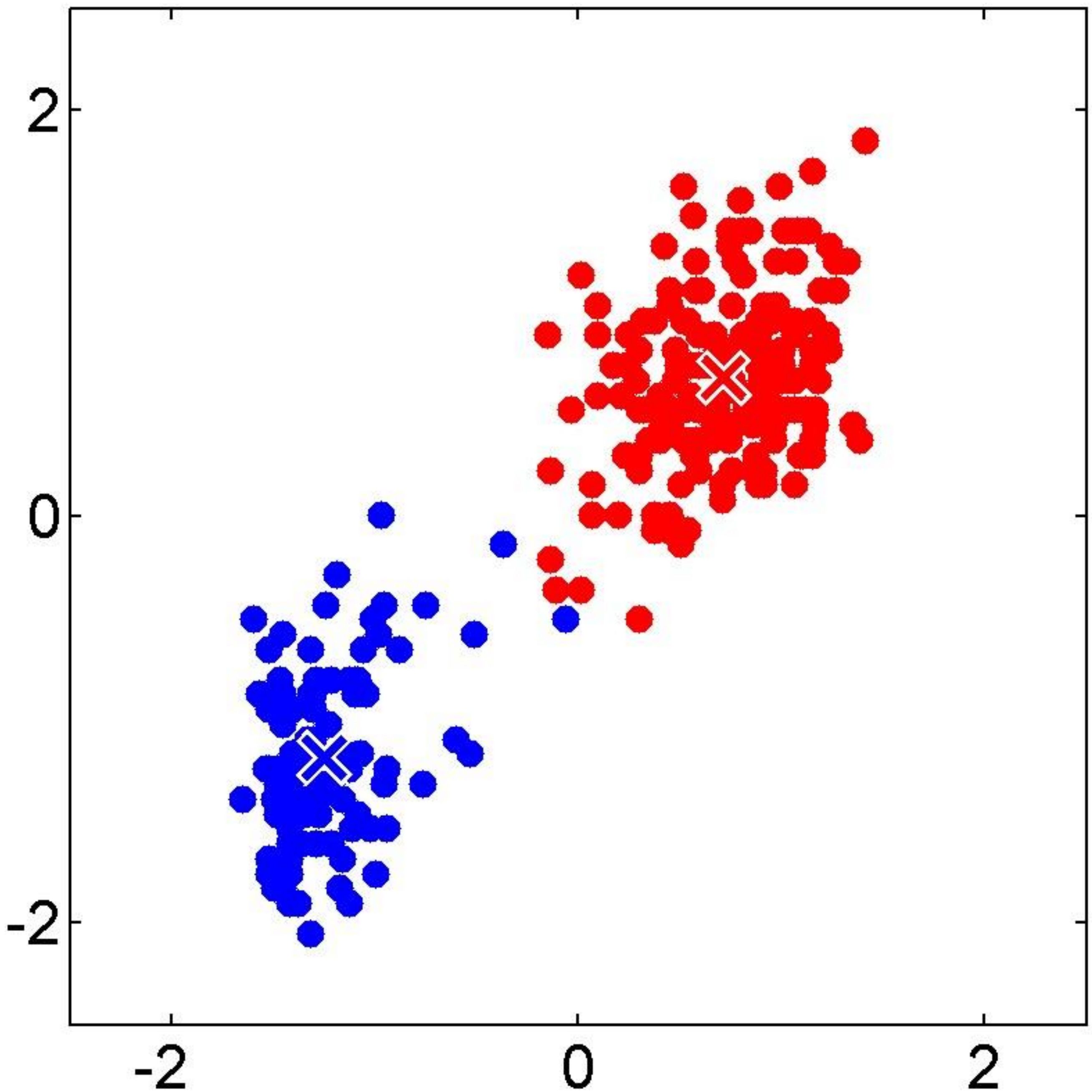
2
0
-2
-2
0
2


2
0
-2
-2
0
2

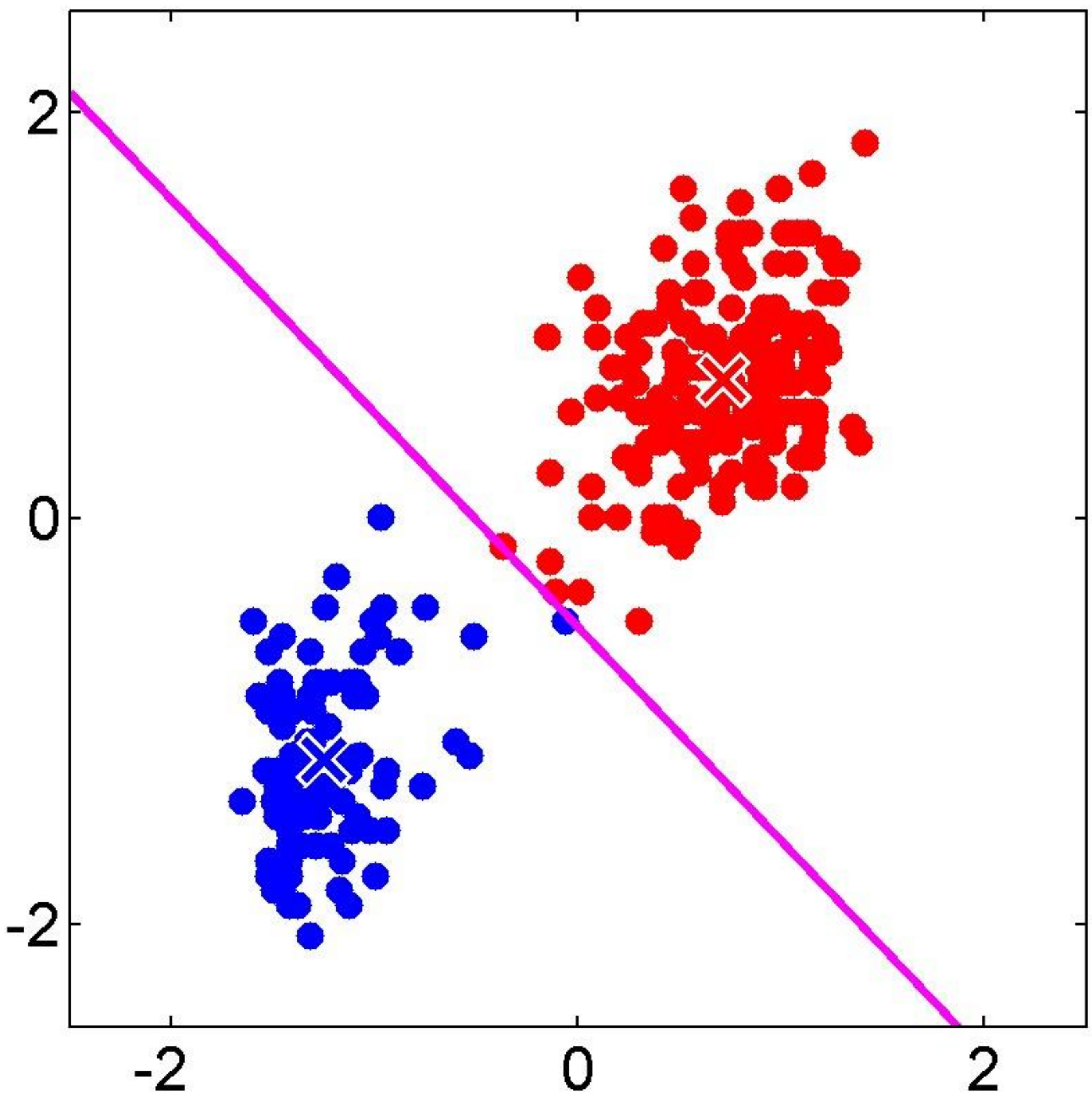


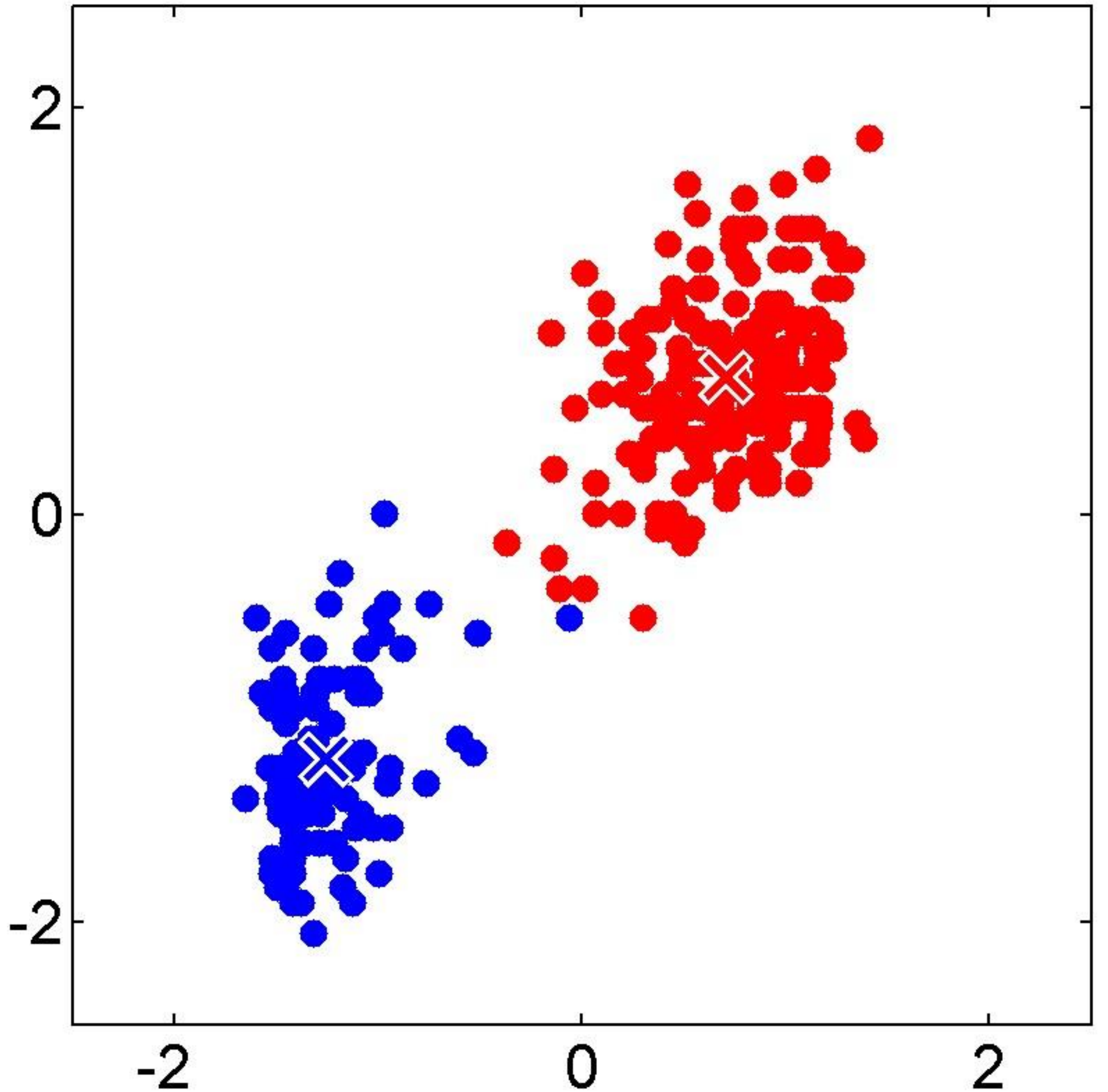
2
0
-2
-2
0
2


2
0
-2
-2
0
2



2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2

Responsibilities

- *Responsibilities* assign data points to clusters

$$r_{nk} \in \{0, 1\}$$

such that $\sum_k r_{nk} = 1$

- Example: 5 data points and 3 clusters

$$(r_{nk}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

K-means Cost Function

data

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

responsibilities

prototypes

Limitations of K-means

- Hard assignments of data points to clusters – small shift of a data point can flip it to a different cluster
- Not clear how to choose the value of K
- Solution: replace ‘hard’ clustering of K-means with ‘soft’ probabilistic assignments
- Represents the probability distribution of the data as a *Gaussian mixture model*

Gaussian Distribution

- Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

mean covariance

- Define precision to be the inverse of the covariance

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$$

- Choice of form of Σ: spherical, diagonal, full, …

Likelihood Function

- Data set

$$D = \{\mathbf{x}_n\} \qquad n = 1, \ldots, N$$

- Assume observed data points generated independently

$$p(D|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Viewed as a function of the parameters, this is known as the *likelihood function*

Maximum Likelihood

- Set the parameters by maximizing the likelihood function
- Equivalently maximize the log likelihood

$$\ln p(D|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Gaussian Mixtures

- Linear super-position of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

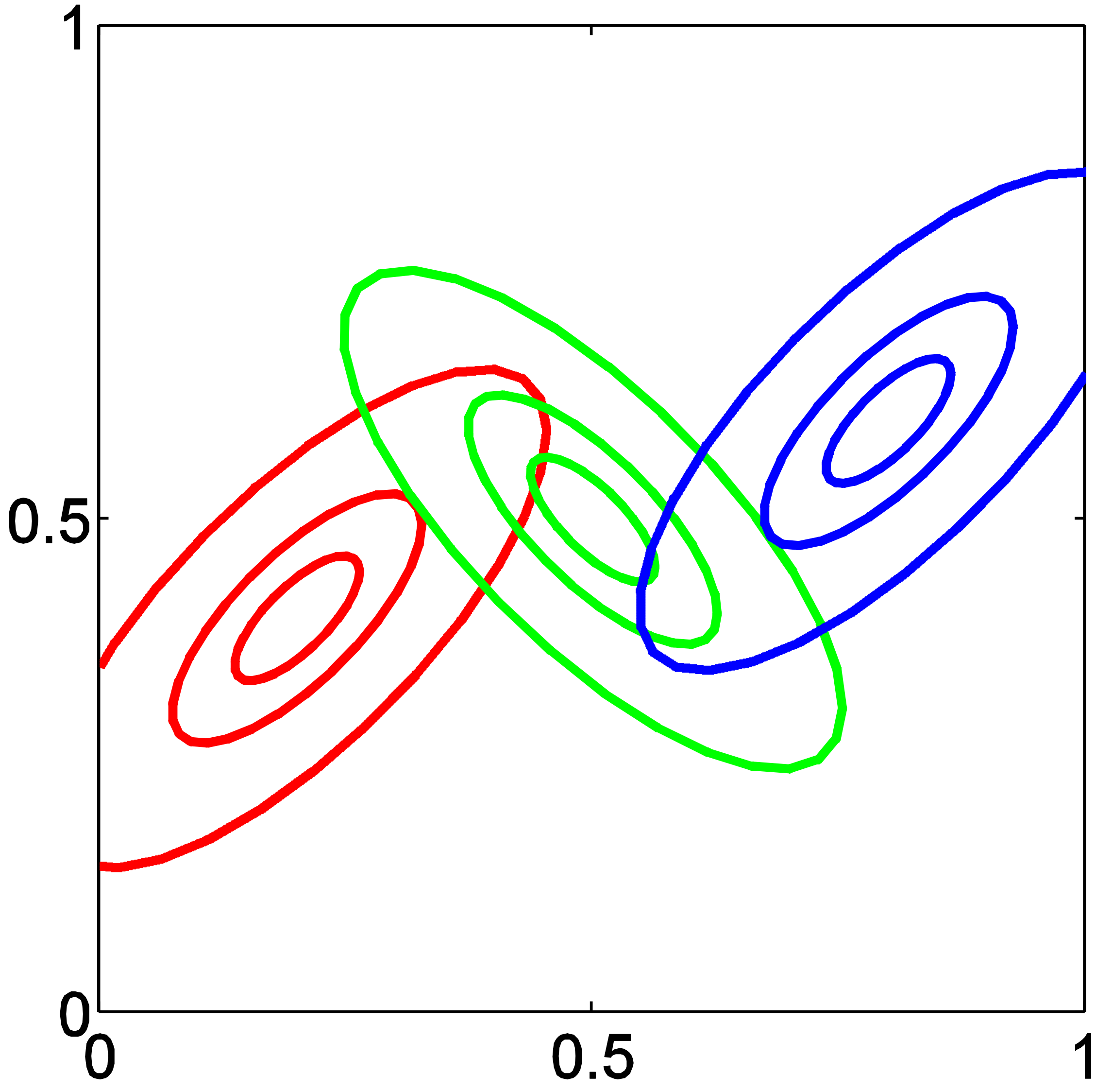
- Normalization and positivity require

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

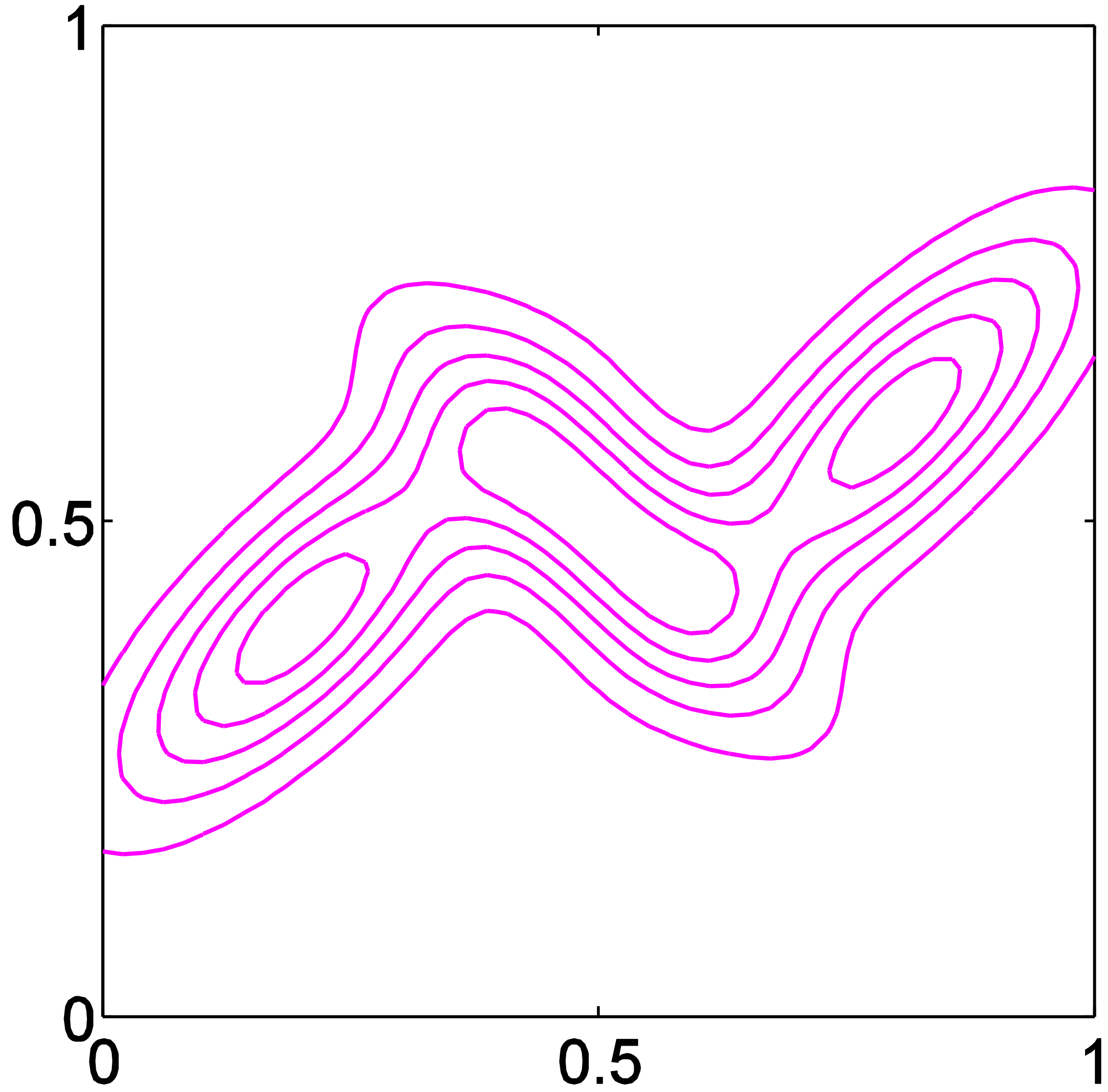
- Can interpret the mixing coefficients as prior probabilities

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

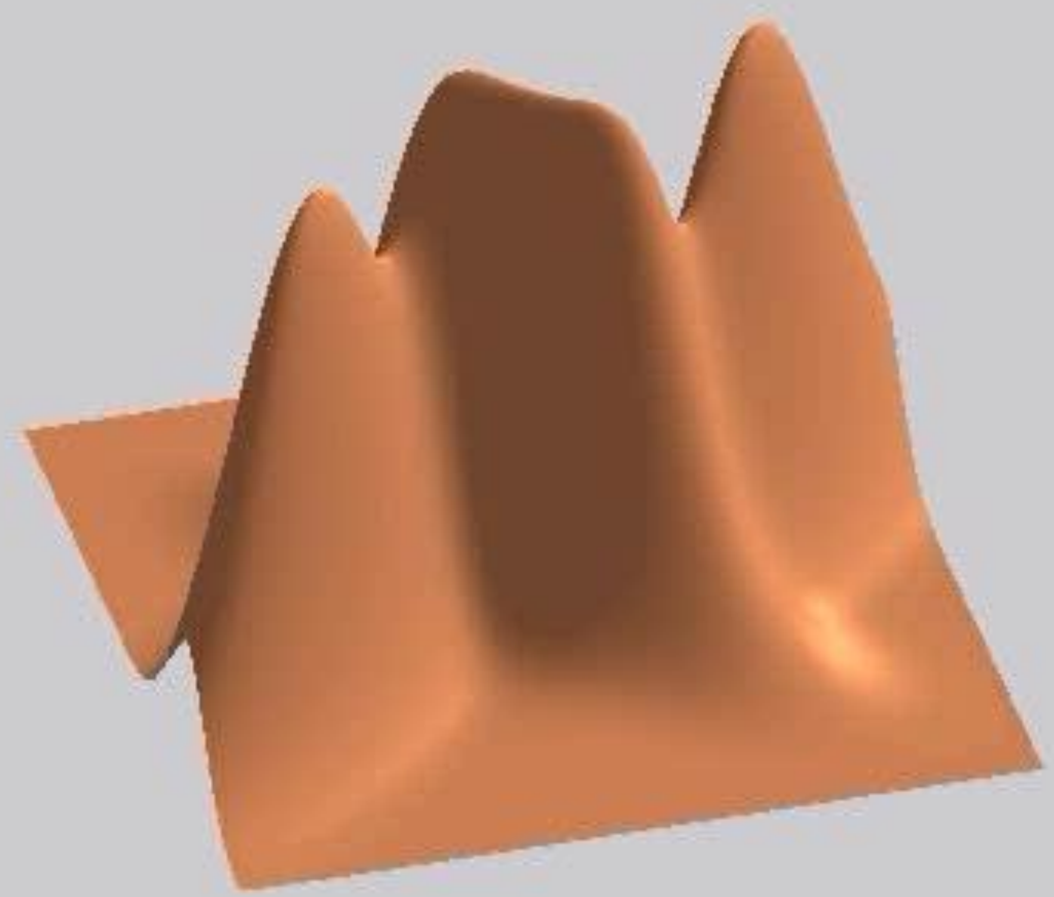
Example: Mixture of 3 Gaussians



Contours of Probability Distribution



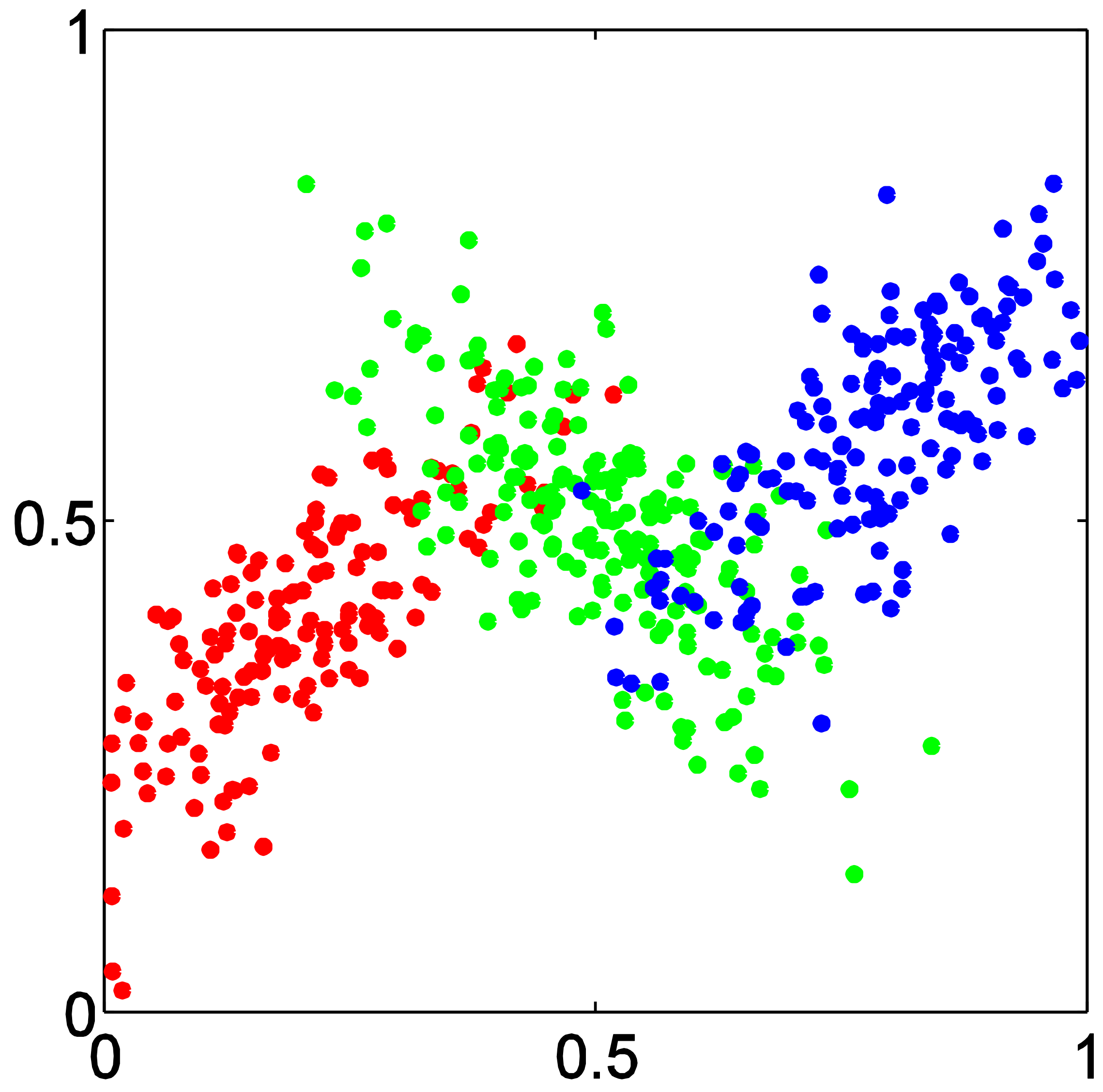
Surface Plot



Generating from the GMM

- To generate a data point:
  - first pick one of the components with probability π_k
  - then draw a sample $\mathbf{x}_n$ from that component
- Repeat these two steps for each new data point

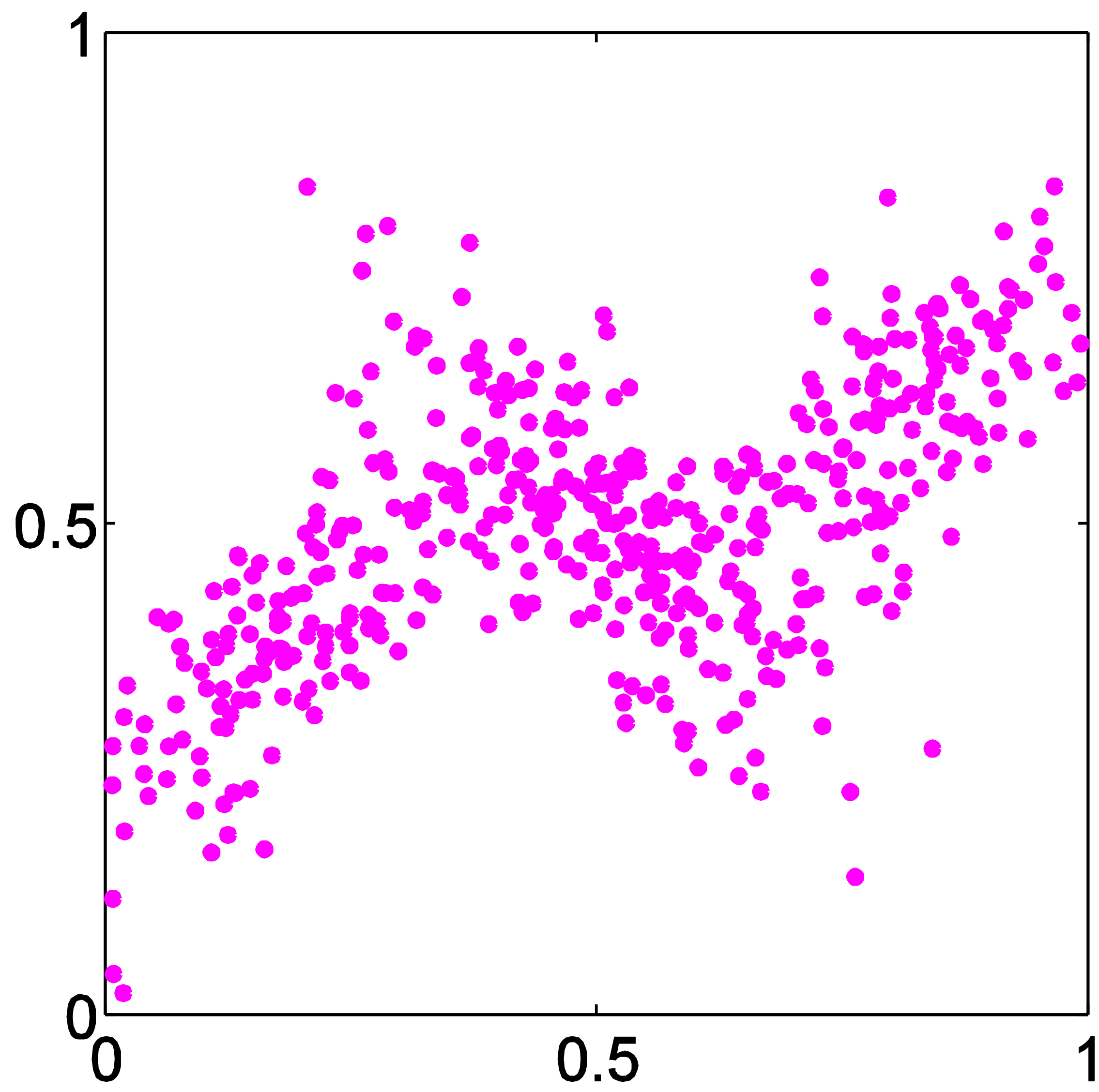
Synthetic Data Set



Fitting the Gaussian Mixture

- We wish to invert this process – given the data set, find the corresponding parameters:
  - mixing coefficients
  - means
  - covariances
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster
- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

Synthetic Data Set Without Labels

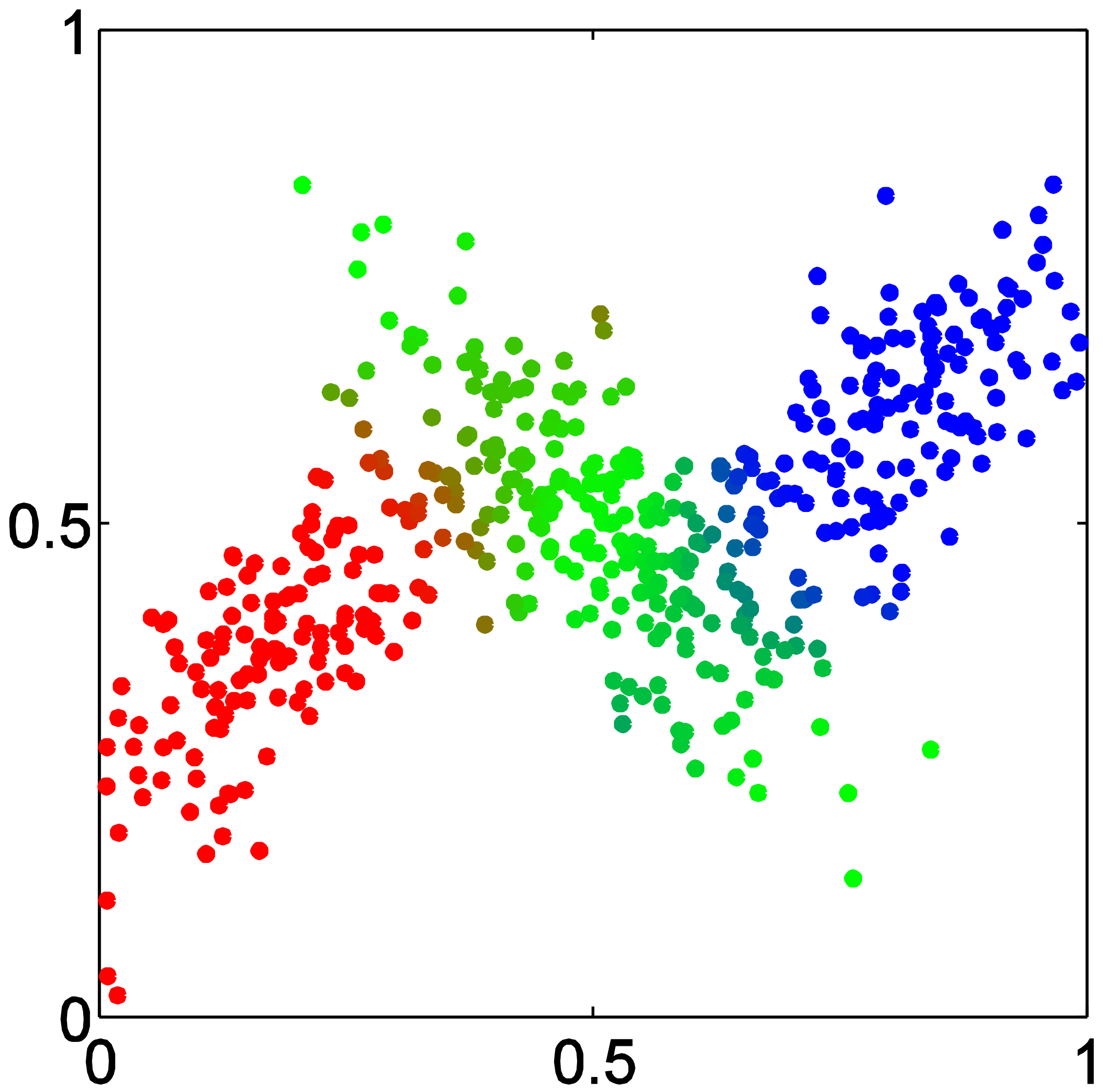


Posterior Probabilities

- We can think of the mixing coefficients as prior probabilities for the components
- For a given value of $\mathbf{x}$ we can evaluate the corresponding posterior probabilities, called *responsibilities*
- These are given from Bayes' theorem by

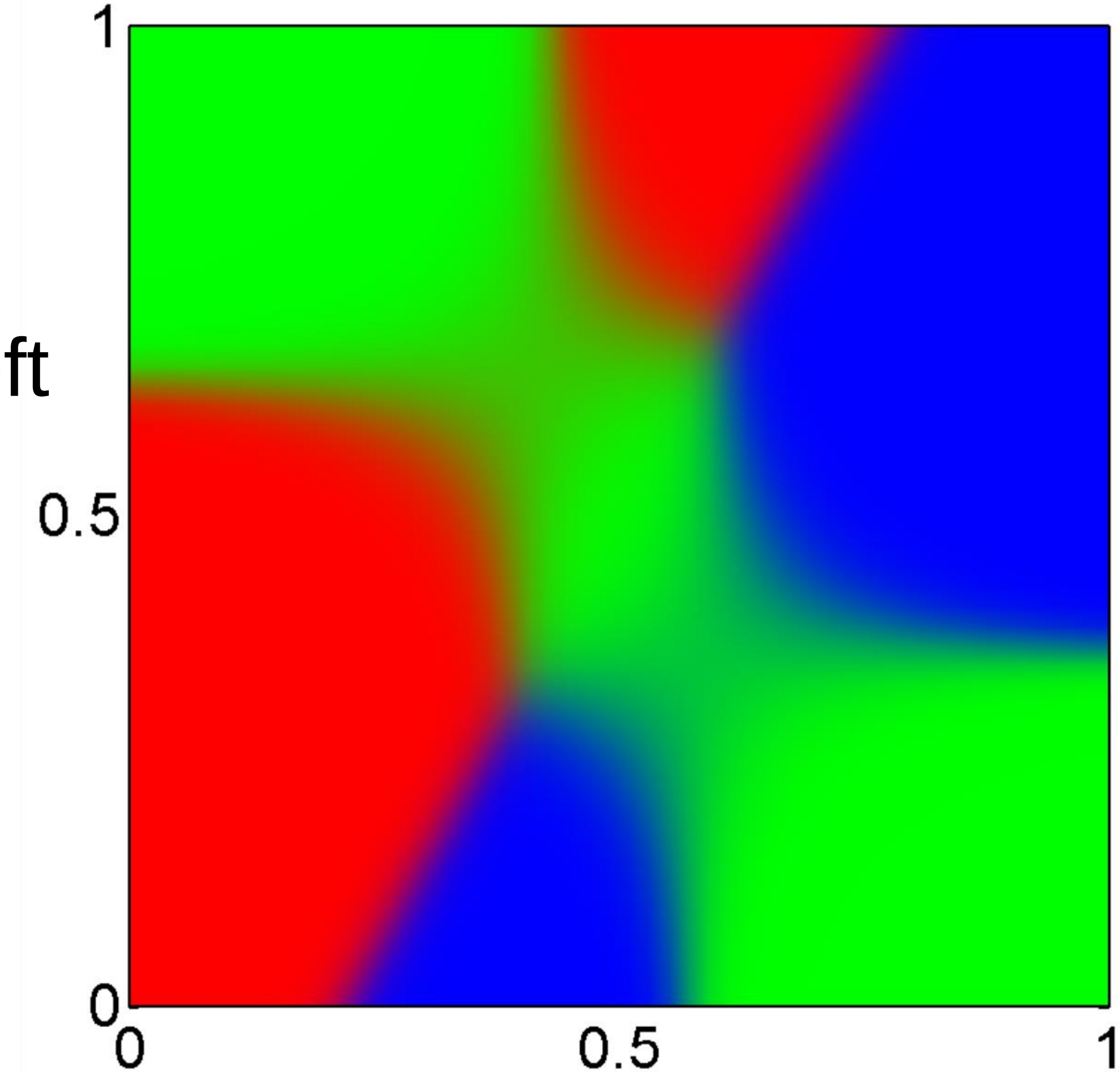
$$\begin{aligned}\gamma_k(\mathbf{x}) \equiv p(k|\mathbf{x}) &= \frac{p(k)p(\mathbf{x}|k)}{p(\mathbf{x})} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\end{aligned}$$

Posterior Probabilities (colour coded)



Posterior Probability Map

Note the soft boundaries



Maximum Likelihood for the GMM

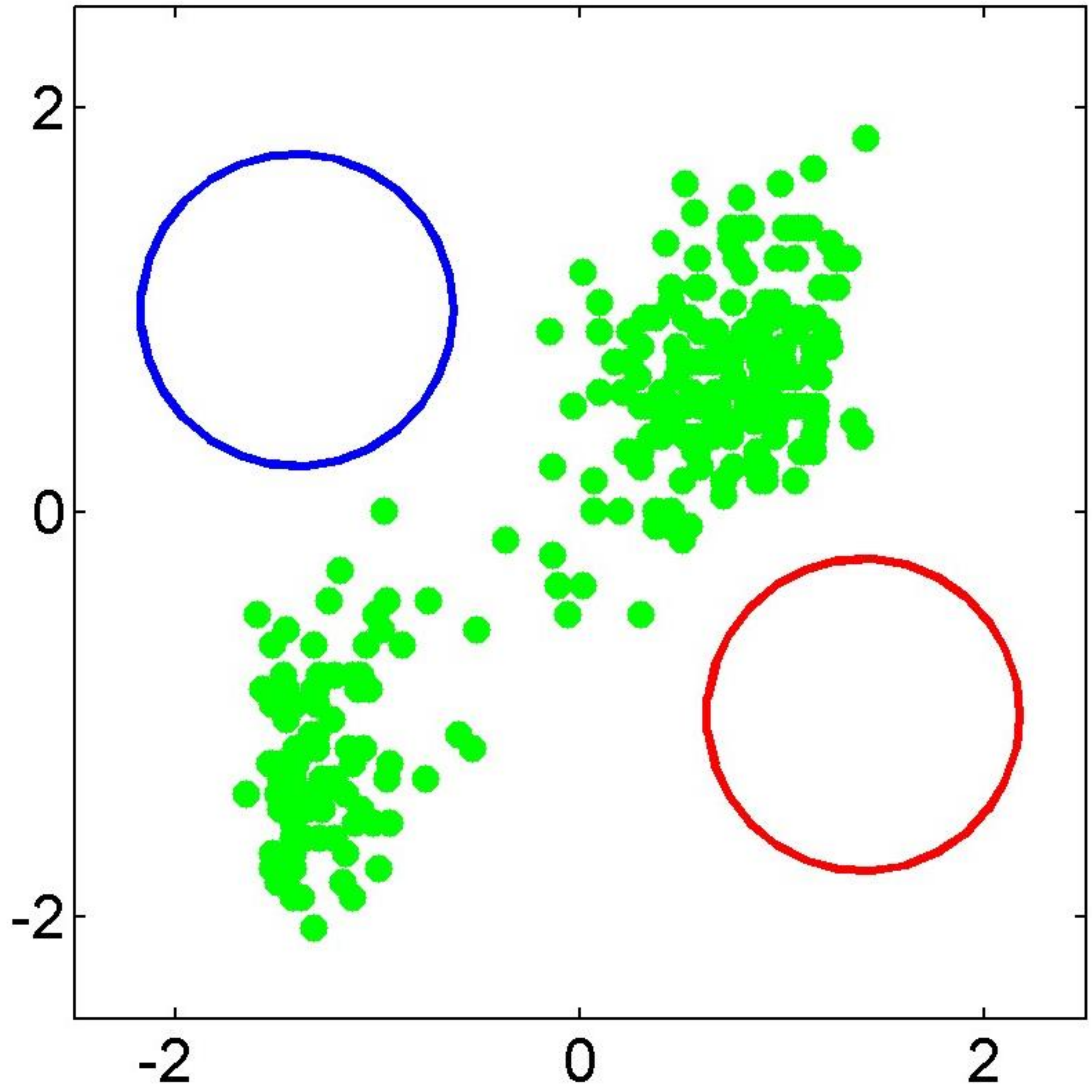
- The log likelihood function takes the form

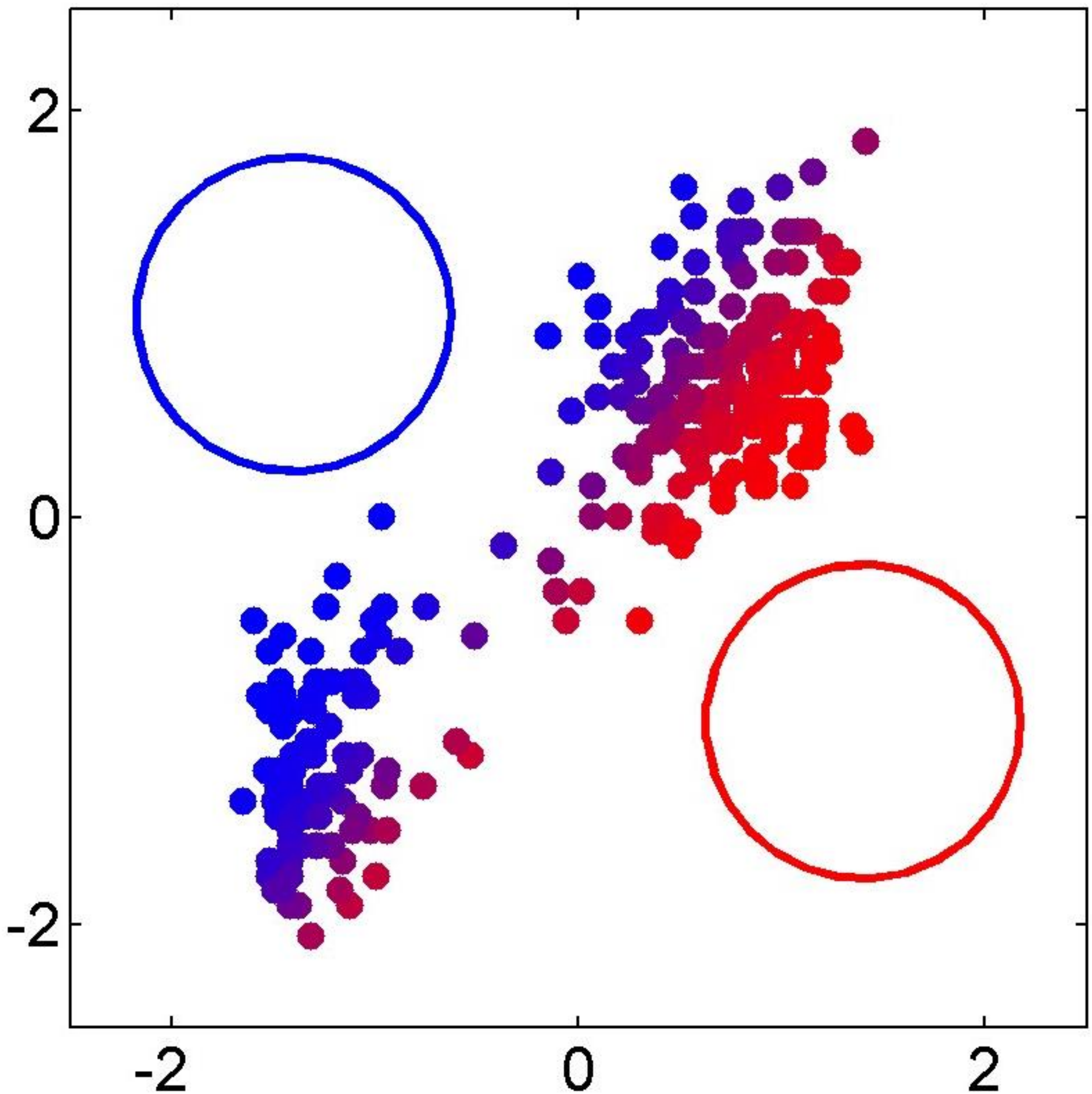
$$\ln p(D|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

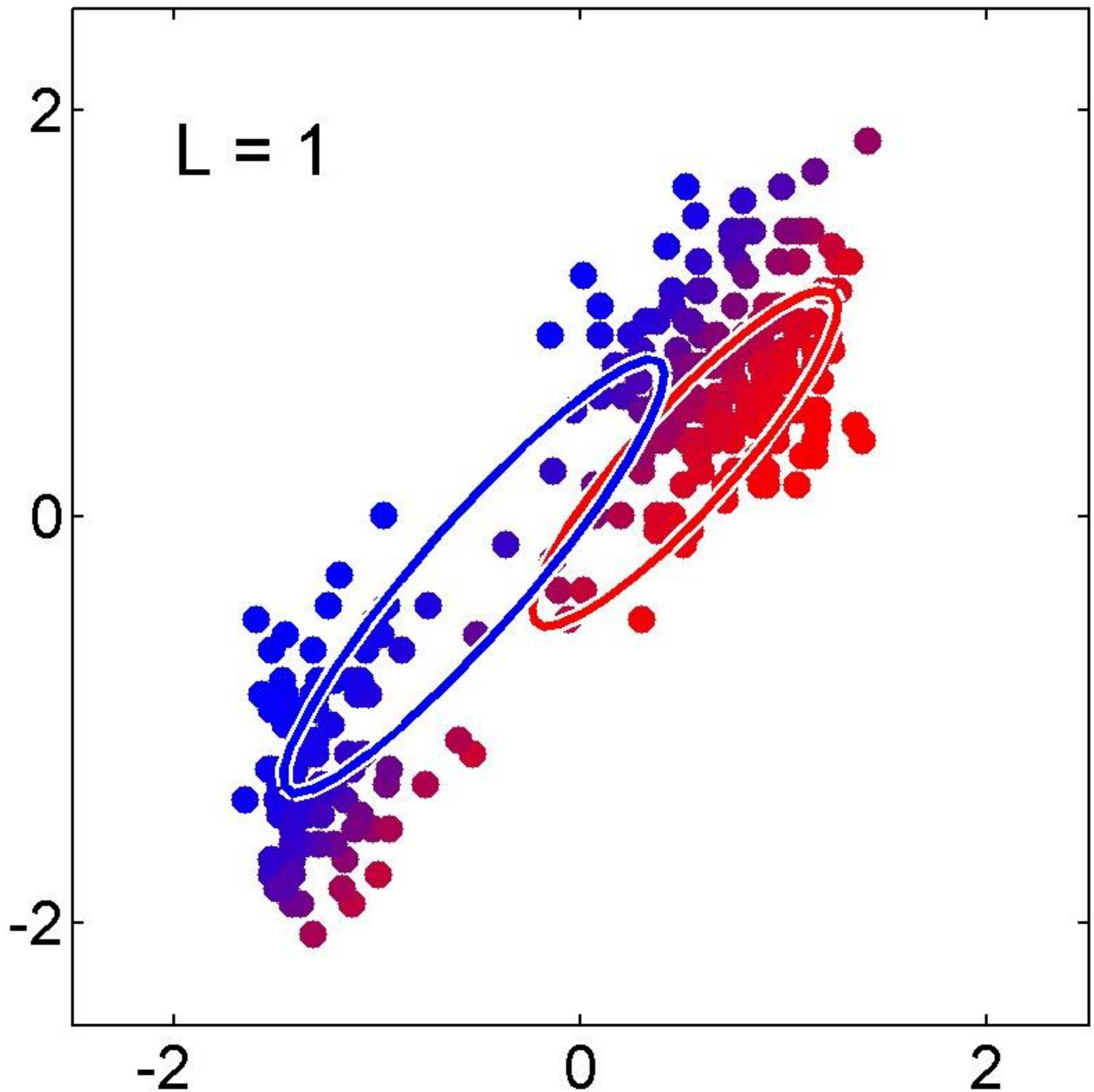
- Note: sum over components appears *inside* the log
- There is no closed form solution for maximum likelihood

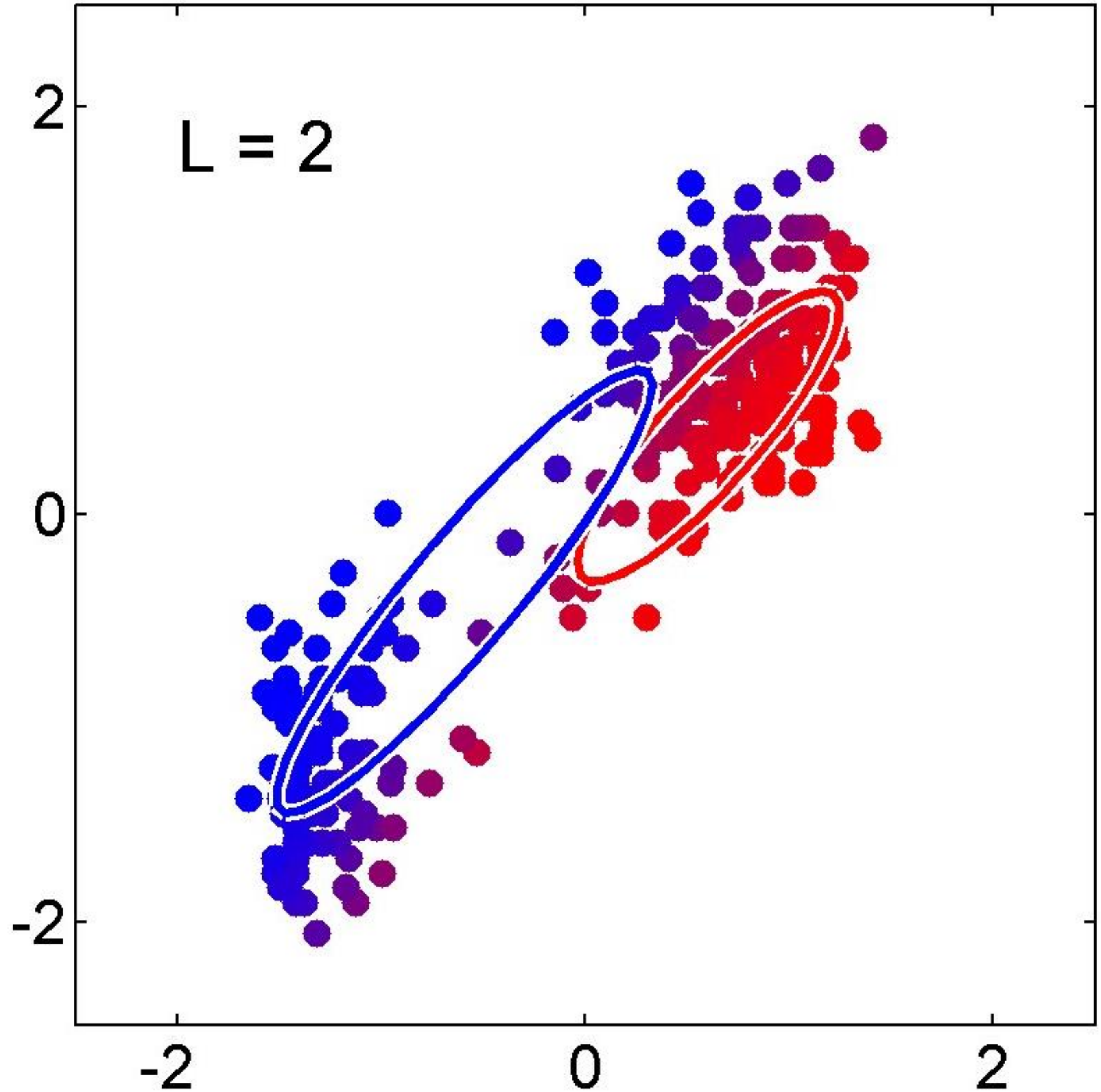
Problems and Solutions

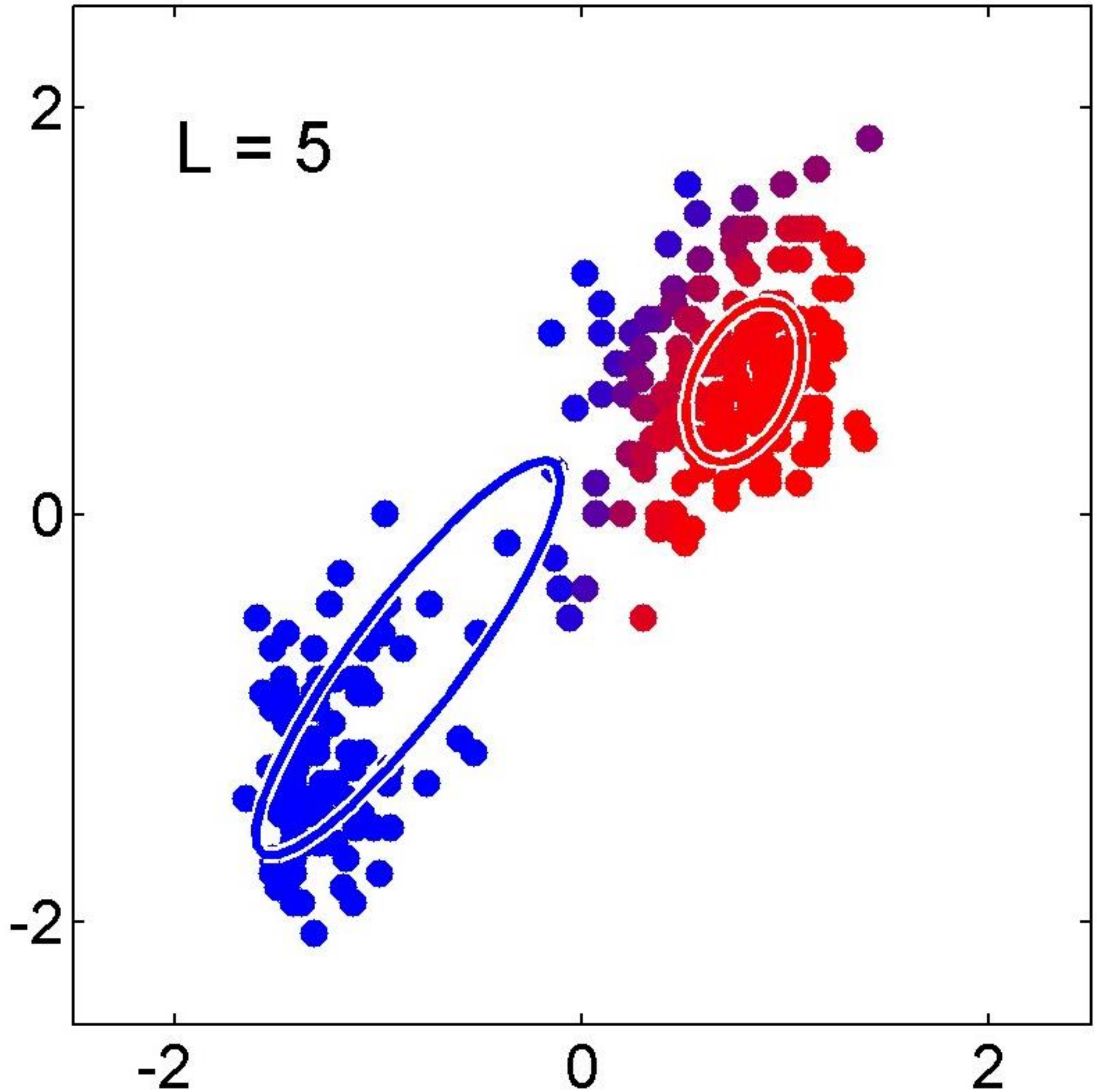
- How to maximize the log likelihood
  - solved by expectation-maximization (EM) algorithm
- How to avoid singularities in the likelihood function
  - solved by a Bayesian treatment
- How to choose number K of components
  - also solved by a Bayesian treatment

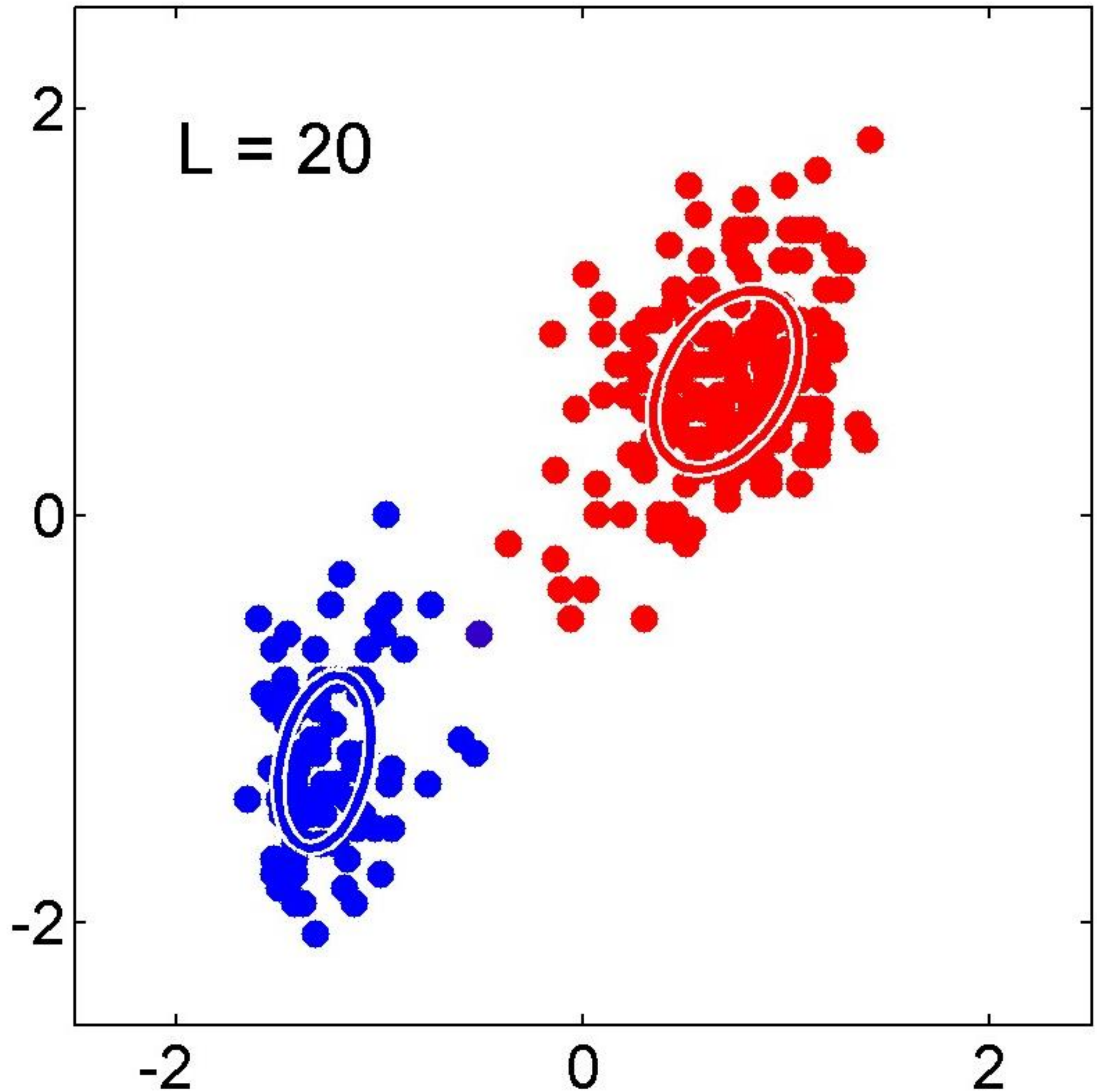

2
0
-2
-2
0
2


2
0
-2
-2
0
2


L = 1
2
0
-2
-2
0
2


L = 2
2
0
-2
-2
0
2


L = 5
2
0
-2
-2
0
2


L = 20
2
0
-2
-2
0
2

Latent Variable Models

- Separate the observed variables and the latent variables. Latent variables generate observations. Use (probabilistic) inference to deduce what is happening in latent variable space.
- Often use Bayes' Theorem:
- $$P(L|O) = \frac{P(O|L)P(L)}{P(O)}$$
- Simplest case is PCA: q latent variables, a linear transformation to observation space and a single Gaussian distribution in latent space.
- Dynamic case:
- Hidden Markov Models: discrete state space. (Speech recognition).
- State-Space Models: continuous state space. (Tracking).

Bayesian Inference

- Include prior distributions over parameters
- Advantages in using *conjugate* priors
- Example: consider a single Gaussian over one variable
  - assume variance is known and mean is unknown
  - likelihood function for the mean

$$p(D|\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

- Choose Gaussian prior for mean

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

Bayesian Inference for a Gaussian

- Posterior (proportional to product of prior and likelihood) will then also be Gaussian

$$p(\mu|D) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

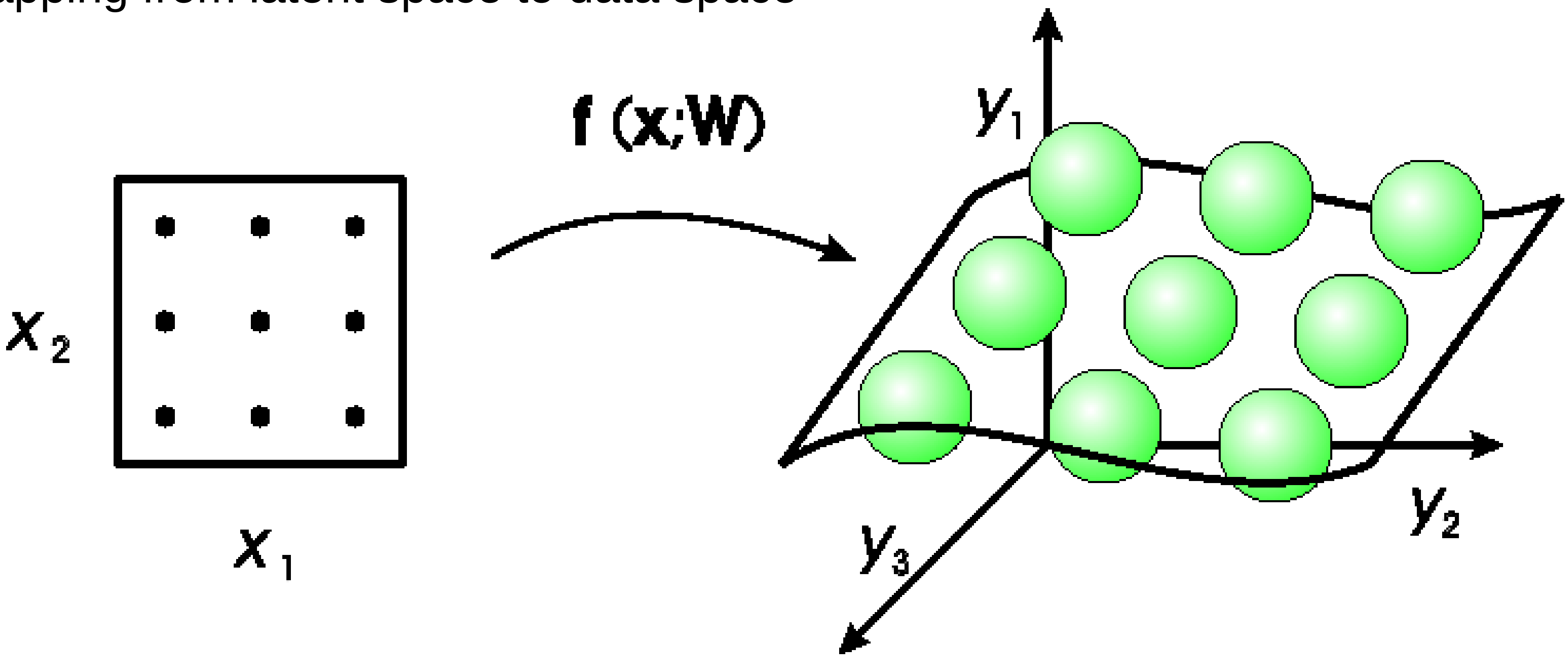
where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

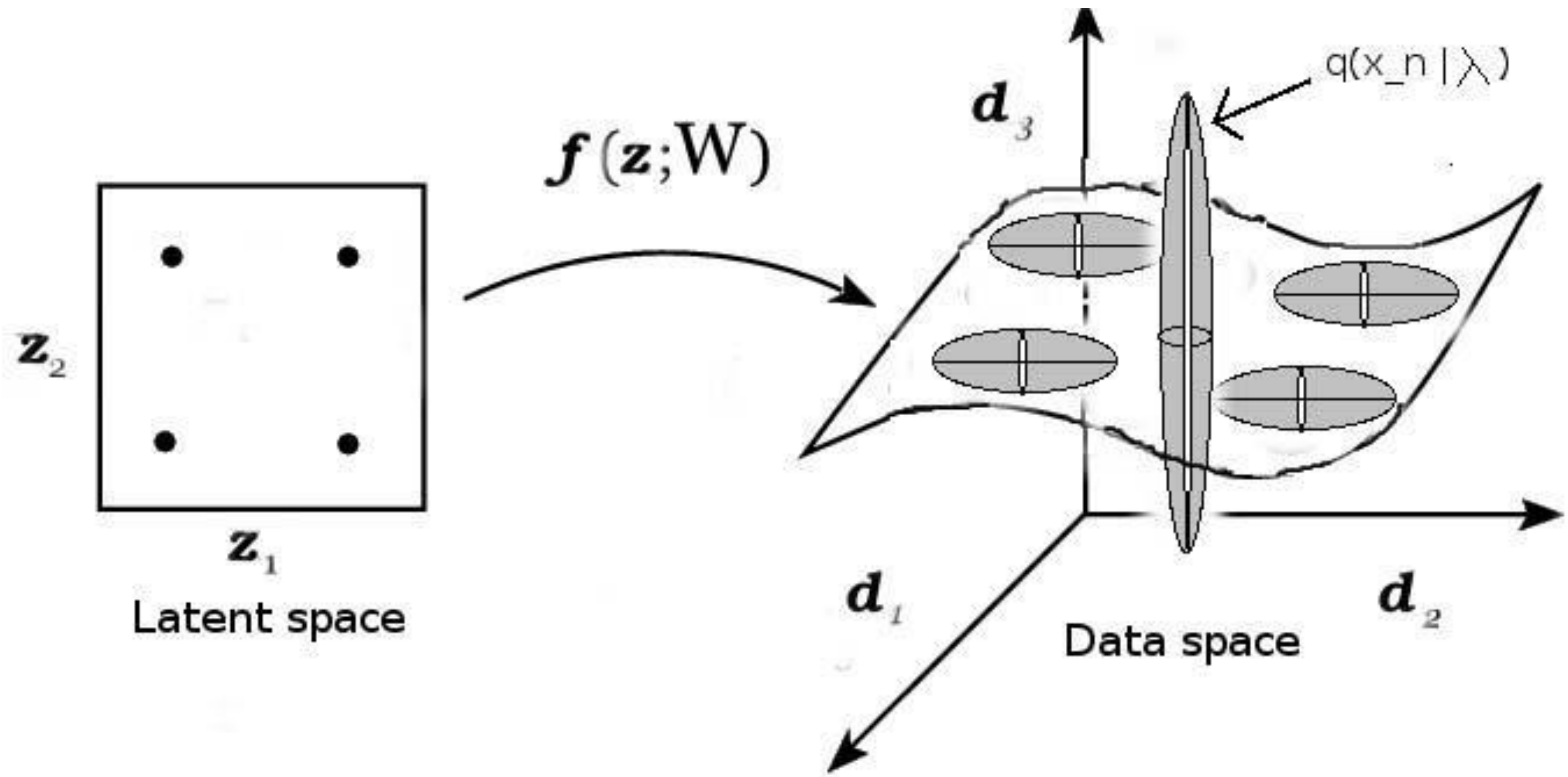
Generative Topographic Mapping

Mapping from latent space to data space



A thick rubber sheet studded with tennis balls. GTM defines p(y|x;W); use Bayes' theorem to compute p(x|y*;W) for a given point y* in data space.

GTM-FS

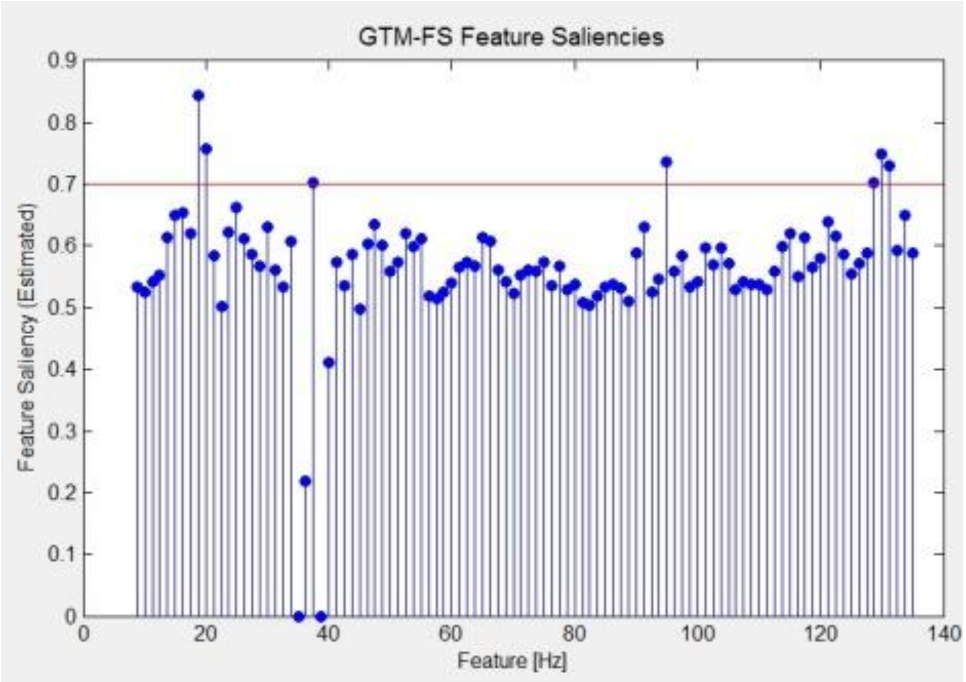


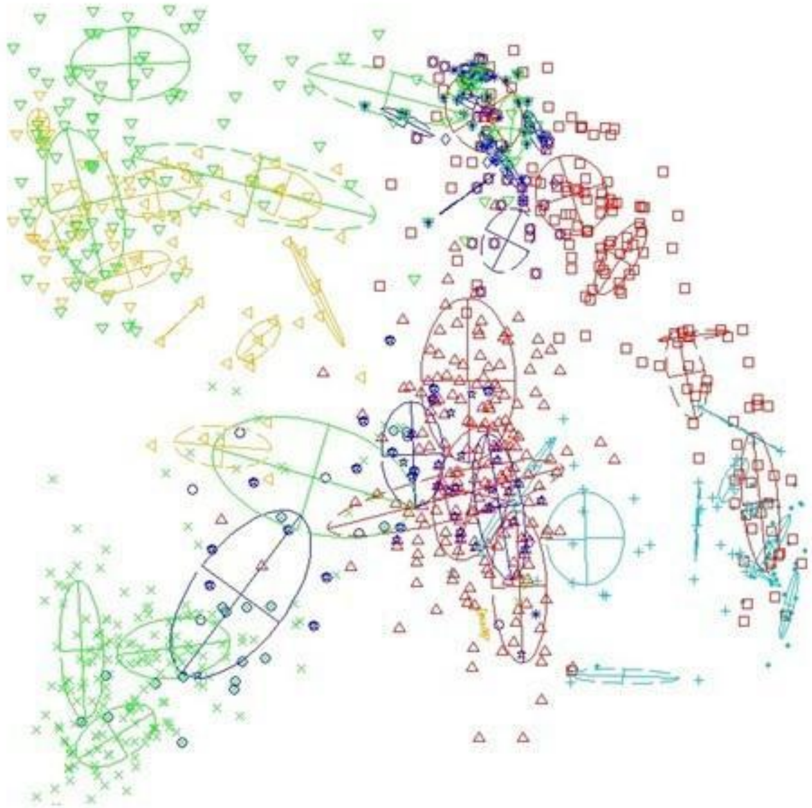
$d_1$and d_2 have high saliency; d_3 has low saliency.

Feature Selection

Features are selected using GTM with Feature Saliencies.

Sensors are selected by comparing inter-class separation in different plots.





Conclusions

- Clustering is an important tool for end users.
- Probabilistic (latent) models allow the user to do clustering+ in a single coherent framework.
- Presenting the data in the right way is key. Feature selection is a very important tool.
- Accounting for known structure (e.g. covariance matrix) improves results.