Maximum Likelihood Estimators, Maximum a-Posteriori, Expectation Maximization

Maksim Bolonkin

Moscow State University

In the previous episode...

- Underconstrained LSP can produce infinitely many solutions, we want "good enough" solutions
- Common way of obtaining good solutions is to add regularization term to objective function

$$\min_{\theta} ||y - A\theta||^2 + \lambda||\theta||^2$$

- Choise of penalty hyperparameter λ is important - risk of over-regularizing

In the previous episode...

- Nonlinear Least Squares - parameters do not vanish after getting the derivative $\longrightarrow$ need to solve non-linear equation
- Variations of Newton's method: iterative approximation of a root of equation
- Update step is a solution to an ordinary linear Least Squares problem

$$\min_{\Delta b_n} ||J_r(b_n)\Delta b_n + r(b_n)||^2$$

- Improvement of Gauss-Newton method is the line search: find step size such that objedtive function is not increasing
- Levenberg-Marquardt method - further improvement: restrict update vector length to some small values

Example: Poisson distribution

Suppose that $X_1, \ldots, X_N$ are i.i.d. discrete random variables, such that $X_i \sim Pois(\theta)$ with probability mass function defined as

$$Pr(X_i = x_i) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

where θ is some unknown parameter.

Example: Poisson distribution

Suppose that $X_1, \ldots, X_N$ are i.i.d. discrete random variables, such that $X_i \sim Pois(\theta)$ with probability mass function defined as

$$Pr(X_i = x_i) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

where θ is some unknown parameter.

Question: What is the probability of observing **particular sample** $\{x_1, \ldots, x_N\}$, assuming that a Poisson distribution with parameter θ (yet unknown) generated the data?

Example: Poisson distribution

Suppose that $X_1, \ldots, X_N$ are i.i.d. discrete random variables, such that $X_i \sim Pois(\theta)$ with probability mass function defined as

$$Pr(X_i = x_i) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

where θ is some unknown parameter.

Question: What is the probability of observing **particular sample** $\{x_1, \ldots, x_N\}$, assuming that a Poisson distribution with parameter θ (yet unknown) generated the data?

$$Pr((X_1 = x_1) \cap \ldots \cap (X_N = x_N)) = \prod_{i=1}^{N} Pr(X_i = x_i)$$

Example: Poisson distribution

We know pmf of the Poisson distribution. Therefore

$$Pr((X_1 = x_1) \cap \ldots \cap (X_N = x_N)) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

$$= e^{-\theta N} \frac{\theta^{\sum_{i=1}^{N}}}{\prod_{i=1}^{N} x_i!}$$

Example: Poisson distribution

We know pmf of the Poisson distribution. Therefore

$$Pr((X_1 = x_1) \cap \ldots \cap (X_N = x_N)) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

$$= e^{-\theta N} \frac{\theta^{\sum_{i=1}^{N}}}{\prod_{i=1}^{N} x_i!}$$

This joint probability is a function of θ and corresponds to the likelihood of the sample $\{x_1, \ldots, x_N\}$

$$L(\theta; x_1, \ldots, x_N) = Pr((X_1 = x_1) \cap \ldots \cap (X_N = x_N))$$

$$L(\theta; x_1, \ldots, x_N) = e^{-\theta N} \times \theta^{\sum_{i=1}^{N}} \times \frac{1}{\prod_{i=1}^{N} x_i!}$$

Example: Poisson distribution

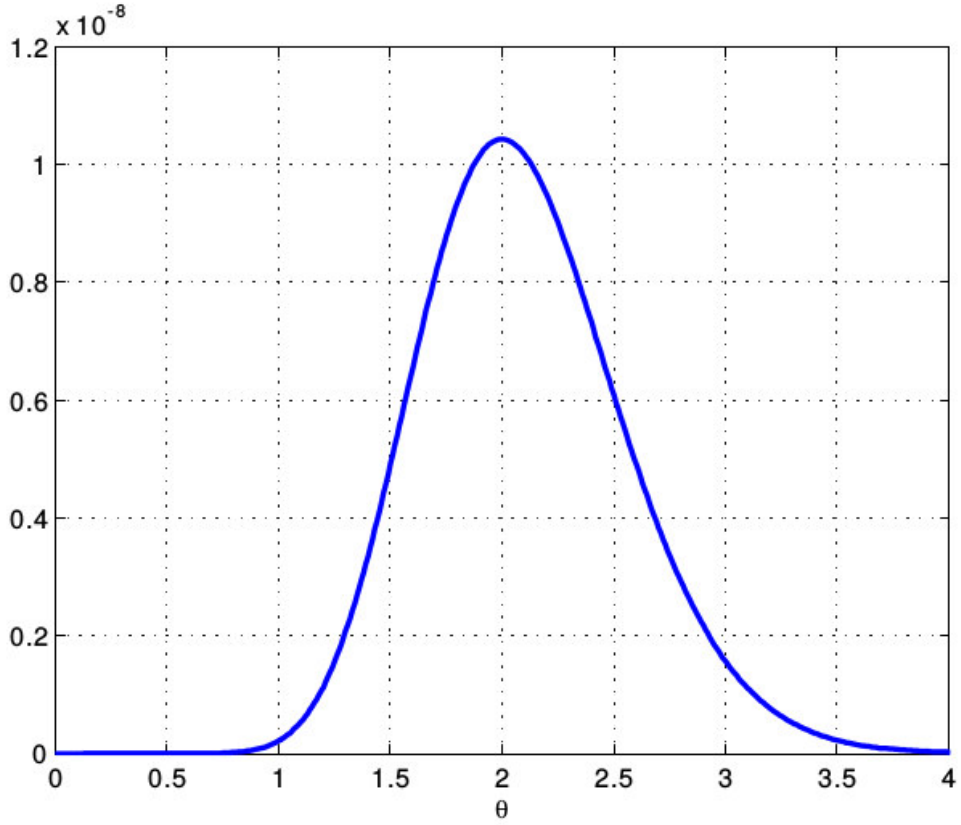
Let's consider some realization of the sample with $N = 10$: $\{5, 0, 1, 1, 0, 3, 2, 3, 4, 1\}$. Then

$$L(\theta; x_1, \ldots, x_{10}) = \frac{e^{-10\Theta}\theta^{20}}{207,360}$$

Example: Poisson distribution

Let's consider some realization of the sample with $N = 10$: $\{5, 0, 1, 1, 0, 3, 2, 3, 4, 1\}$. Then

$$L(\theta; x_1, \ldots, x_{10}) = \frac{e^{-10\Theta}\theta^{20}}{207,360}$$



Example: Poisson distribution

Let's find the value of parameter θ that produces the largest likelihood for given data. Instead of maximizing directly we will maximize logarithm of the likelihood:

$$\ln L(\theta; x_1, \ldots, x_N) = -\theta N + \ln(\theta) \sum_{i=1}^{N} x_i - \ln \left(\prod_{i=1}^{N} x_i! \right)$$

$$\frac{\partial \ln L(\theta; x_1, \ldots, x_N)}{\partial \theta} = -N + \frac{1}{\theta} \sum_{i=1}^{N} x_i$$

$$\frac{\partial^2 \ln L(\theta; x_1, \ldots, x_N)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^{N} x_i < 0$$

Maximum Likelihood Estimate

Maximum Likelihood estimate (estimator) is defined as:

$$\hat{\theta} = \arg\max_{\theta} \ln L(\theta; x_1, \ldots, x_N)$$

Maximum Likelihood Estimate

Maximum Likelihood estimate (estimator) is defined as:

$$\hat{\theta} = \arg\max_{\theta} \ln L(\theta; x_1, \ldots, x_N)$$

$$\left.\frac{\partial \ln L(\theta; x_1, \ldots, x_N)}{\partial \theta}\right|_{\hat{\theta}} = -N + \frac{1}{\hat{\theta}} \sum_{i=1}^{N} x_i = 0$$

$$\iff \hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\left.\frac{\partial^2 \ln L(\theta; x_1, \ldots, x_N)}{\partial \theta^2}\right|_{\hat{\theta}} = -\frac{1}{\hat{\theta}^2} \sum_{i=1}^{N} x_i < 0$$

Maximum Likelihood Estimate

- The Maximum Likelihood **estimate** is a value

$$\hat{\theta} \equiv \hat{\theta}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

  Given some fixed numbers for x_i we'll get a value of estimate.

- The Maximum Likelihood **estimator** is a random variable depending on some other random variables

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

The Likelihood Function

The Likelihood Function is defined to be:

$$L_N : \Theta \times \mathbb{R}^N \to R^+$$

$$(\theta; x_1, \ldots, x_N) \mapsto L_N(\theta; x_1, \ldots, x_N) = \prod_{i=1}^{N} f_X(x_i; \theta)$$

where $f_X(x; \theta)$ is a probability density function of a random variable X.

The log-Likelihood Function is defined to be:

$$\ell_N : \Theta \times \mathbb{R}^N \to R$$

$$(\theta; x_1, \ldots, x_N) \mapsto \ell_N(\theta; x_1, \ldots, x_N) = \sum_{i=1}^{N} \ln f_X(x_i; \theta)$$

Example: Normal distribution

If $Y \sim N(m, \sigma^2)$ then:

$$f_Y(y;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-m)^2}{2\sigma^2}}$$

with

$$\theta = \begin{pmatrix} m \\ \sigma^2 \end{pmatrix}$$

Likelihood and log-Likelihood functions would be of the form:

$$L(\theta; y) = (\sigma^2 2\pi)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-m)^2}$$

$$\ell(\theta; y) = -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - m)^2$$

Example: Normal distribution

What is the MLE of m and σ^2?

$$\frac{\partial \ell(\theta; y)}{\partial m} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - m) \Longrightarrow \hat{m} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

$$\frac{\partial \ell(\theta; y)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (y_i - m)^2 \Longrightarrow \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

Hessian matrix is negative definite (can verify on your own).

Example: Linear regression model

Let's consider linear regression model:

$$y_i = x_i^T \beta + \epsilon_i$$

where $x_i, \beta \in \mathbb{R}^K$ and noise is normally distributed $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then the conditional log-likelihood of the observations (x_i, y_i) is given by

$$\ell(\theta; y|x) = -\frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

where $\theta = (\beta^T \sigma^2)^T \in \mathbb{R}^{K+1}$.

What are the MLE of β and σ^2?

Example: Linear regression model

Finding the derivative of log-likelihood and equating to 0:

$$\frac{\partial \ell(\theta; y)}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i (y_i - x_i^T \beta) \Longrightarrow \hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^T \right)^{-1} \left(\sum_{i=1}^{N} X_i Y_i \right)$$

Example: Linear regression model

Finding the derivative of log-likelihood and equating to 0:

$$\frac{\partial \ell(\theta; y)}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i (y_i - x_i^T \beta) \Longrightarrow \hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^T \right)^{-1} \left(\sum_{i=1}^{N} X_i Y_i \right)$$

$$\frac{\partial \ell(\theta; y)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \Longrightarrow \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - X_i^T \hat{\beta})^2$$

Again, fairly easy to verify that Hessian is negative definite, thus estimators are maximizing log-likelihood.

Invariance principle

Under suitable regularity conditions, the maximum likelihood estimator of a function $g(.)$ of the parameter θ is $g(\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ.

This gives some agility in terms of reparametrization. For example, if parameter θ occurs in the model in denominator, we can substitute it with $\gamma = 1/\theta$

Important assumption

The log-likelihood and the Maximum Likelihood estimator are always based on an assumption about the distribution of the random variable.

$$Y_i \sim \text{ distribution with pdf } f_Y(y;\theta) \Longrightarrow L(y;\theta) \text{ and } \ell(y;\theta)$$

In practice we generally don't know true distribution of Y.

MLE properties

The maximum likelihood estimator has the following properties:

- *Consistency* - as the smple size tends to infinity the MLE tends to the "true" value of the parameter
- *Asymptotic normality* - as the sample size increases, the distribution of the MLE tends to the normal distribution
- *Efficiency* - as the sample size tends to infinity, there are nor any other unbiased estimators with a lower mean squared error

What questions do you have?

Some terminology

- $p(\mathcal{D}|\theta)$ is called the *likelihood*
- $p(\theta)$ is called *prior distribution*
- $p(\theta|\mathcal{D})$ is called *posterior distribution* and in general can be computed using Bayes' Rule

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\theta')p(\mathcal{D}|\theta')d\theta'}$$

- The *posterior predictive distribution* $p(\mathcal{D}'|\mathcal{D})$ is the distribution over unseen observations given known observations

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\theta|\mathcal{D})p(\mathcal{D}'|\theta)d\theta$$

Example

Suppose we are interested in modeling the distribution of temperatures in Tashkent in March. We assume that temperatures are distributed according to Gaussian distribution with unknown mean μ and known standard deviation σ. We want to find μ that is most probable given observations.

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

Example

Suppose we are interested in modeling the distribution of temperatures in Tashkent in March. We assume that temperatures are distributed according to Gaussian distribution with unknown mean μ and known standard deviation σ. We want to find μ that is most probable given observations.

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

We know how to find the likelihood:

$$p(\mathcal{D}|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

What prior distribution should we pick?

Conjugate prior

One source of prior distributions is domain knowledge. More common way is to select **conjugate prior**. Conjugate prior comes from the same family of distributions and has similar functional representation.

Conjugate prior

One source of prior distributions is domain knowledge. More common way is to select **conjugate prior**. Conjugate prior comes from the same family of distributions and has similar functional representation.

Let's look at the data point distribution:

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

Easy to see that this expression also represents a Gaussian distribution over μ. Therefore conjugate prior for this problem will be the Gaussian distribution with μ_p and σ_p for parameters.

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

Example

Now we can express the posterior:

$$p(\mu|\mathcal{D}) \propto \left[\frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2}\right] \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}\right]$$

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

$$\propto e^{-\frac{(\mu-\mu_{\text{post}})^2}{\sigma_{\text{post}}^2}}$$

where

$$\sigma_{\text{post}} = \frac{1}{\sqrt{\frac{1}{\sigma_p^2} + \frac{N}{\sigma^2}}} \qquad \mu_{\text{post}} = \frac{\mu_p/\sigma_p^2 + N/\sigma^2(1/N\sum x_i)}{1/\sigma_p^2 + N/\sigma^2}$$

Maximum a-posteriori

In the example we deduced the distribution posterior will have according to selected prior. To infer some prediction regarding unseen data we need to integrate over all possible values of parameter. This approach is called full Bayesian approach.

Instead we can approximate the parameter with a single value that maximizes the posterior distribution. In this example we will obtain the same value for mean $\mu_{MAP} = \mu_{post}$ but predictions might be different.

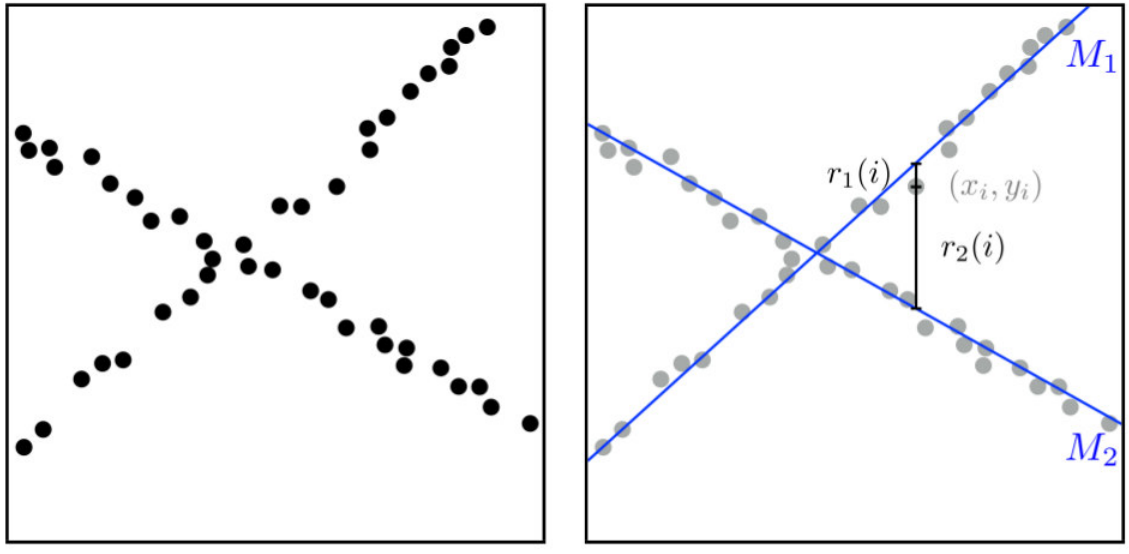
MAP is similar to MLE with prior distribution serving as a regularizer.

What questions do you have?

Yet another example

Let's consider some data that was generated from two different models:

$$y(i) = a_1 x(i) + b_1 + \epsilon_1(i)$$
$$y(i) = a_2 x(i) + b_2 + \epsilon_2(i)$$



Yet another example

In this example we have two problems:

1. We don't know the parameters. If we knew the parameters we would be able to make assignments of points to one or the other model (by selecting which is closer).

Expectation Maximization (EM) algorithm iteratively estimates both assignments and model parameters. Every iteration has 2 steps, at each we assume one is fixed and estimate the other.

Yet another example

In this example we have two problems:

1. We don't know the parameters. If we knew the parameters we would be able to make assignments of points to one or the other model (by selecting which is closer).
2. We don't know what model every point belongs to. If we knew the assignments, we could easily estimate the parameters using ordinary LSP.

Expectation Maximization (EM) algorithm iteratively estimates both assignments and model parameters. Every iteration has 2 steps, at each we assume one is fixed and estimate the other.

E-step

Assuming that model parameters are known. We assign each data point a probability of being generated by one model or the other based on calculated residuals
$r_k(i) = a_k x(i) + b(k) - y(i), \quad k = 1, 2$. Thus we get

$$P(a_k, b_k | r_k(i)) = \frac{P(r_k(i)|a_k, b_k)}{P(r_1(i)|a_1, b_1) + P(r_1(i)|a_2, b_2)}$$

If we assume noise to be Gaussian distributed with 0 mean, probability takes the form

$$w_k(i) = P(a_k, b_k | r_k(i)) = \frac{e^{-r_k^2(i)/2\sigma^2}}{e^{-r_1^2(i)/2\sigma^2} + e^{-r_2^2(i)/2\sigma^2}}$$

M-step

We estimated probabilities of points being generated by both models. We can estimate the parameters assuming these assignments are true. This brings us to weighted least squares problem

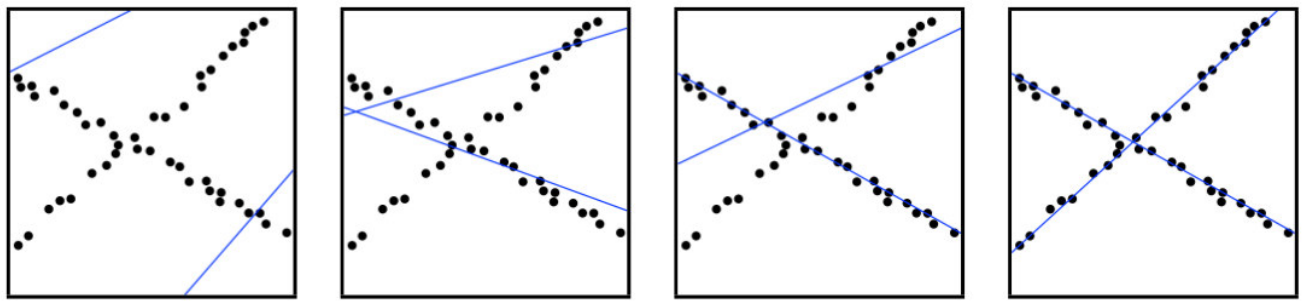
$$E(a_k, b_k) = \sum_{i=1}^{n} (w_k(i)(a_k x(i) + b_k - y(i)))^2$$

Or in matrix form:

$$E(m_k) = ||W_k(Xm_k - y)||^2$$

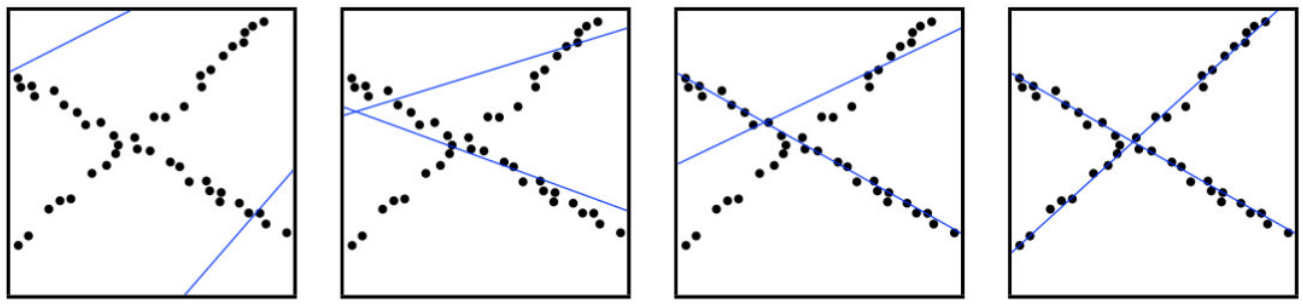
EM algorithm

E and M steps are iterated until convergence.



EM algorithm

E and M steps are iterated until convergence.



EM algorithm is guaranteed to converge. But quality of obtained result largely depends on initial parameters and parameters of noise distribution, like σ. One recommendation is to update σ on each EM iteration

$$\sigma_k = \frac{\sum_{i=1}^{n} w_k(i) r_k^2(i)}{\sum_{i=1}^{n} w_k(i)}$$

The Gaussian Distribution

- Multivariate Gaussian

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

mean

covariance

- Define precision to be the inverse of the covariance

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

- In 1-dimension

$$\tau = \frac{1}{\sigma^2}$$

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

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

Sampling from the Gaussian

- 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

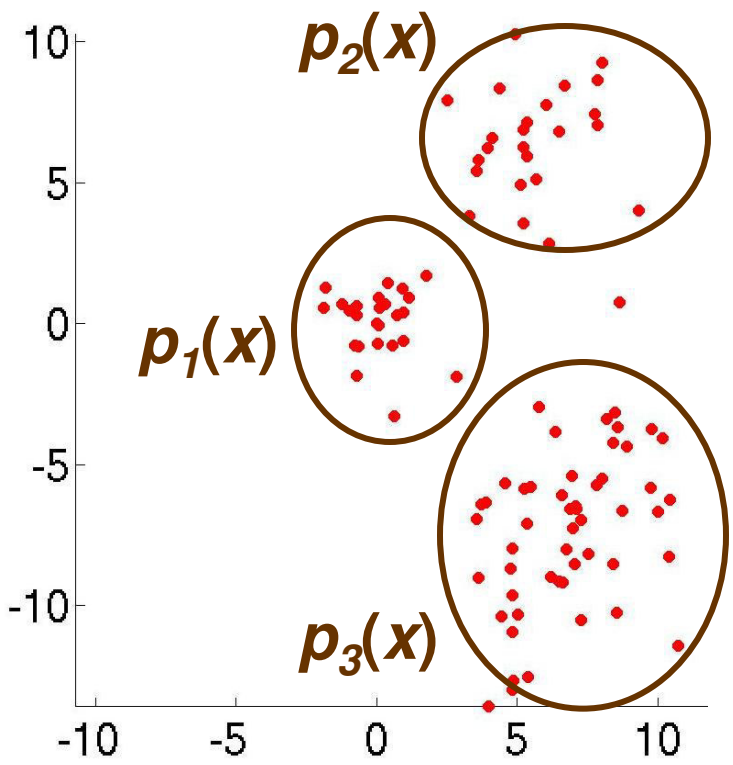
Example: Gaussian Mixture Density

- Mixture of 3 Gaussians

$$p_1(x) \cong N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

$$p_2(x) \cong N\left([6,6], \begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}\right)$$

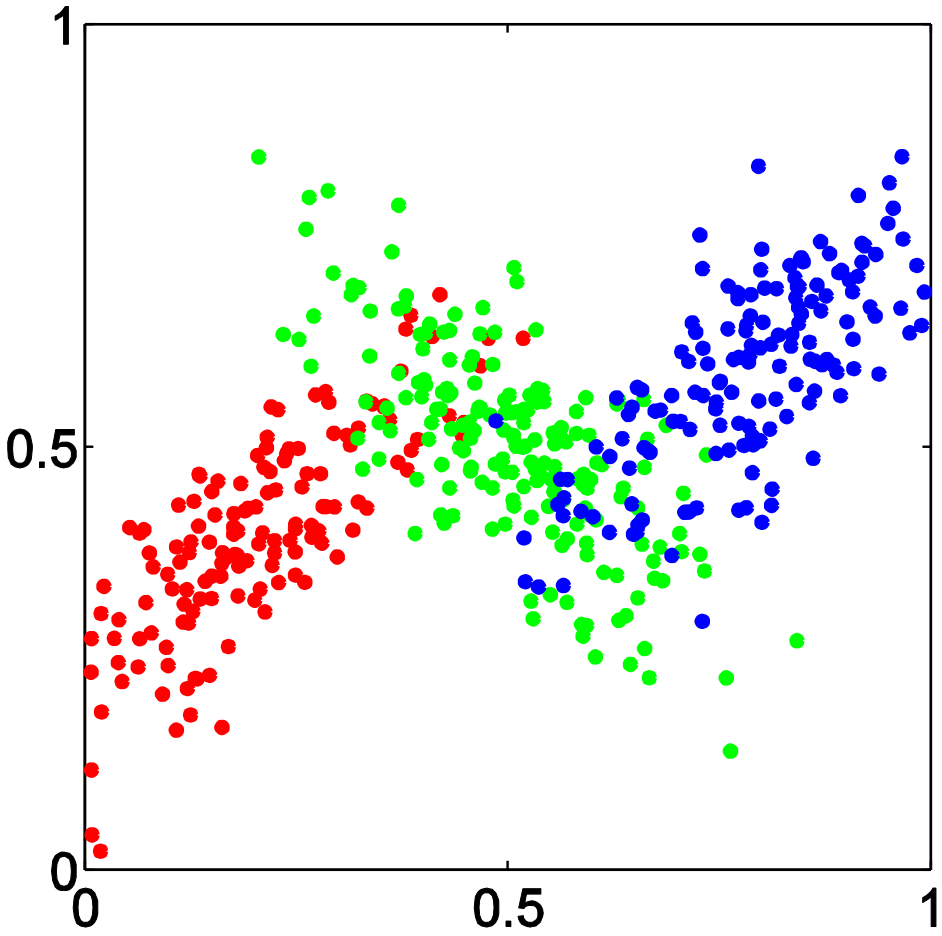
$$p_3(x) \cong N\left([7,-7], \begin{bmatrix}6 & 0\\0 & 6\end{bmatrix}\right)$$



$$p(x) = 0.2p_1(x) + 0.3p_2(x) + 0.5p_3(x)$$

Synthetic Data Set

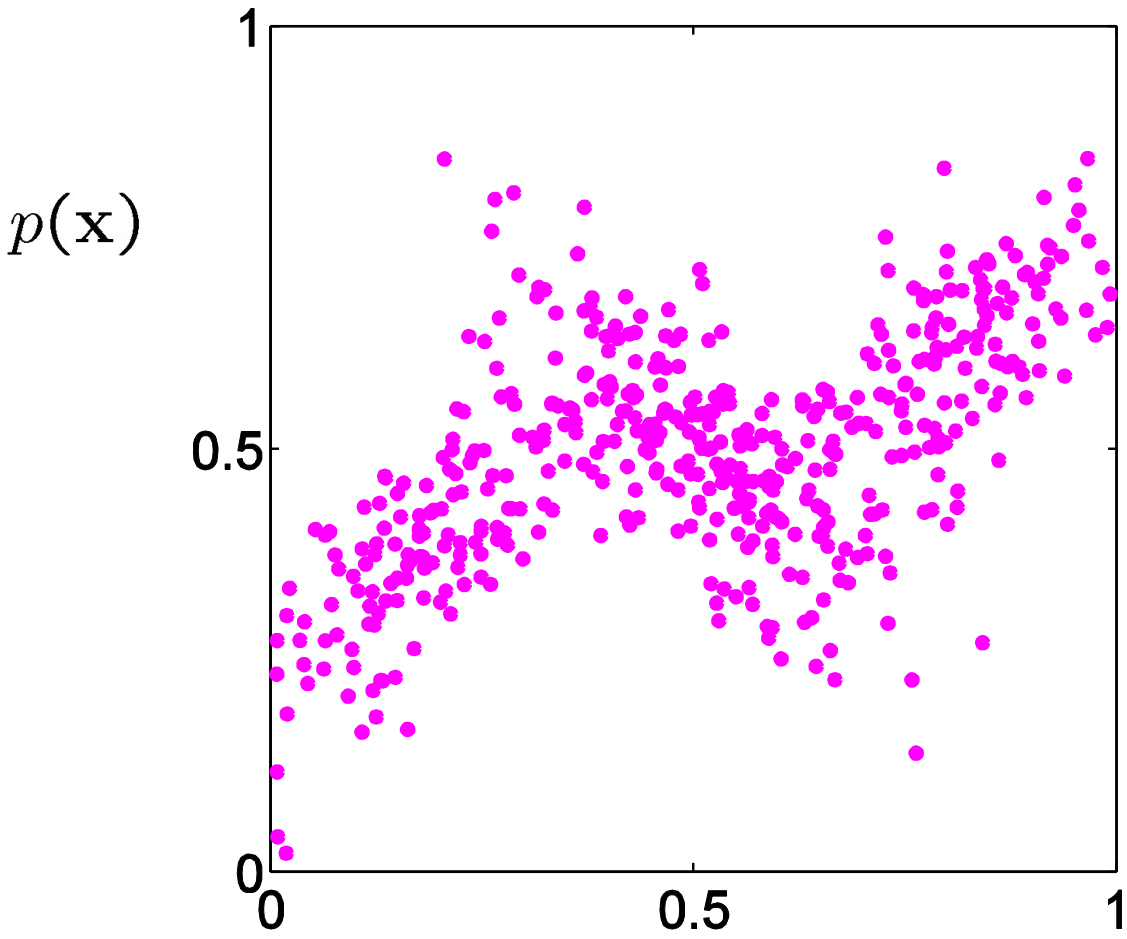
$p(\mathbf{x}|k)$



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

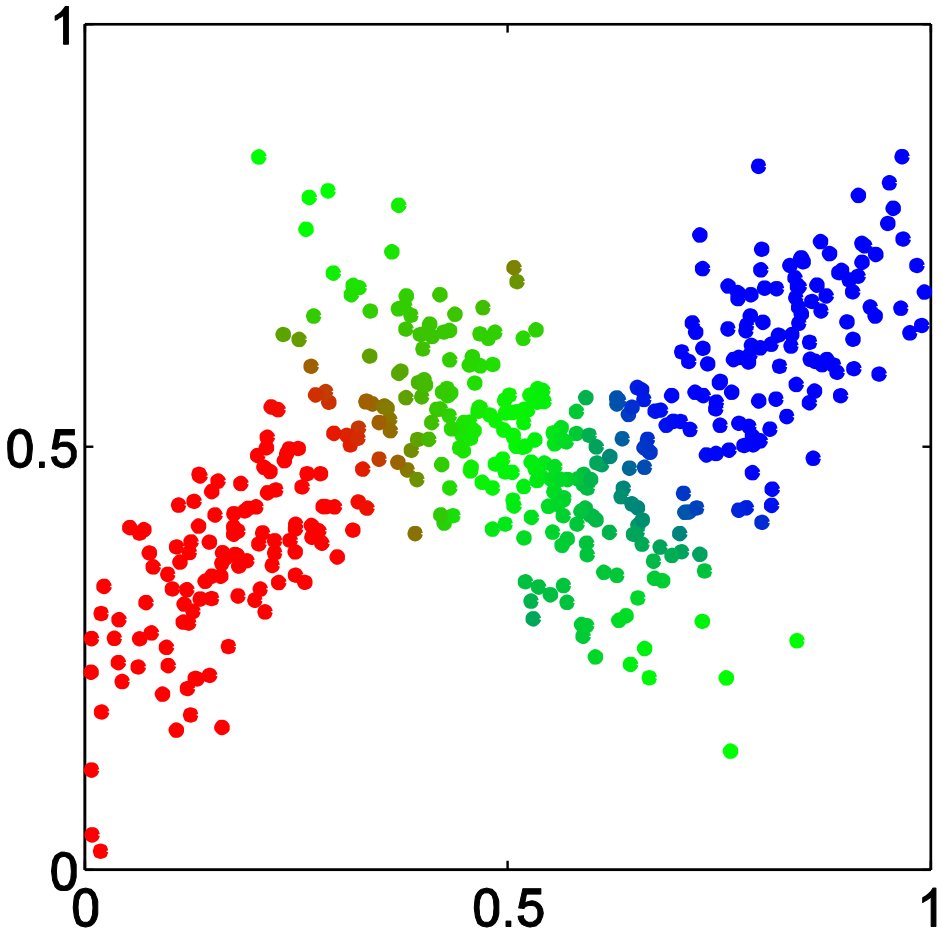
- We can think of the mixing coefficients as prior probabilities for the components
- For a given value of $\mathbf{x}$ we can evaluate the corresponding posterior probabilities, called *responsibilities*
- These are given from Bayes' theorem by

$$\gamma_k(\mathbf{x}) \equiv p(k|\mathbf{x}) = \frac{p(k)p(\mathbf{x}|k)}{p(\mathbf{x})}$$

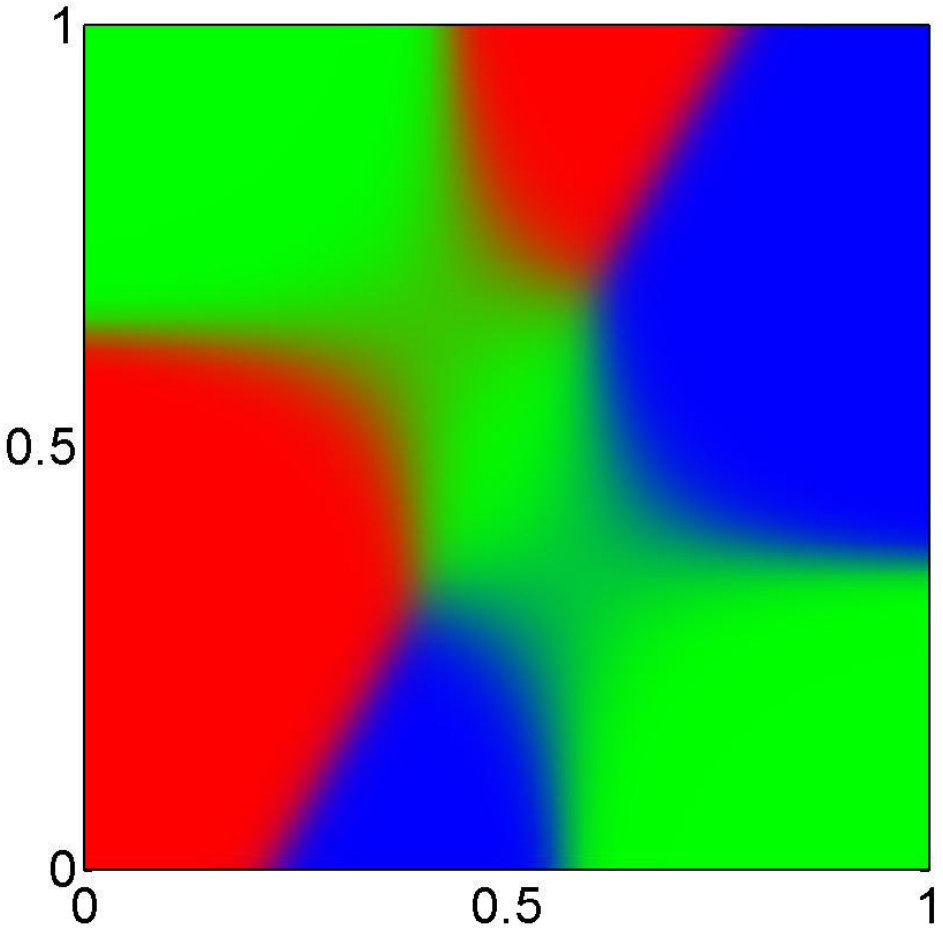
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Posterior Probabilities (colour coded)

$p(k \mid x)$



Posterior Probability Map



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
- How to maximize the log likelihood
  - solved by expectation-maximization (EM) algorithm

EM Algorithm – Informal Derivation

- Let us proceed by simply differentiating the log likelihood
- Setting derivative with respect to $\boldsymbol{\mu}_j$ equal to zero gives

$$-\sum_{n=1}^{N} \underbrace{\frac{\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}}_{\gamma_j(\mathbf{x}_n)} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_j) = 0$$

giving

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

which is simply the weighted mean of the data

EM Algorithm – Informal Derivation

- Similarly for the covariances

$$\boldsymbol{\Sigma}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)^{\mathsf{T}}}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

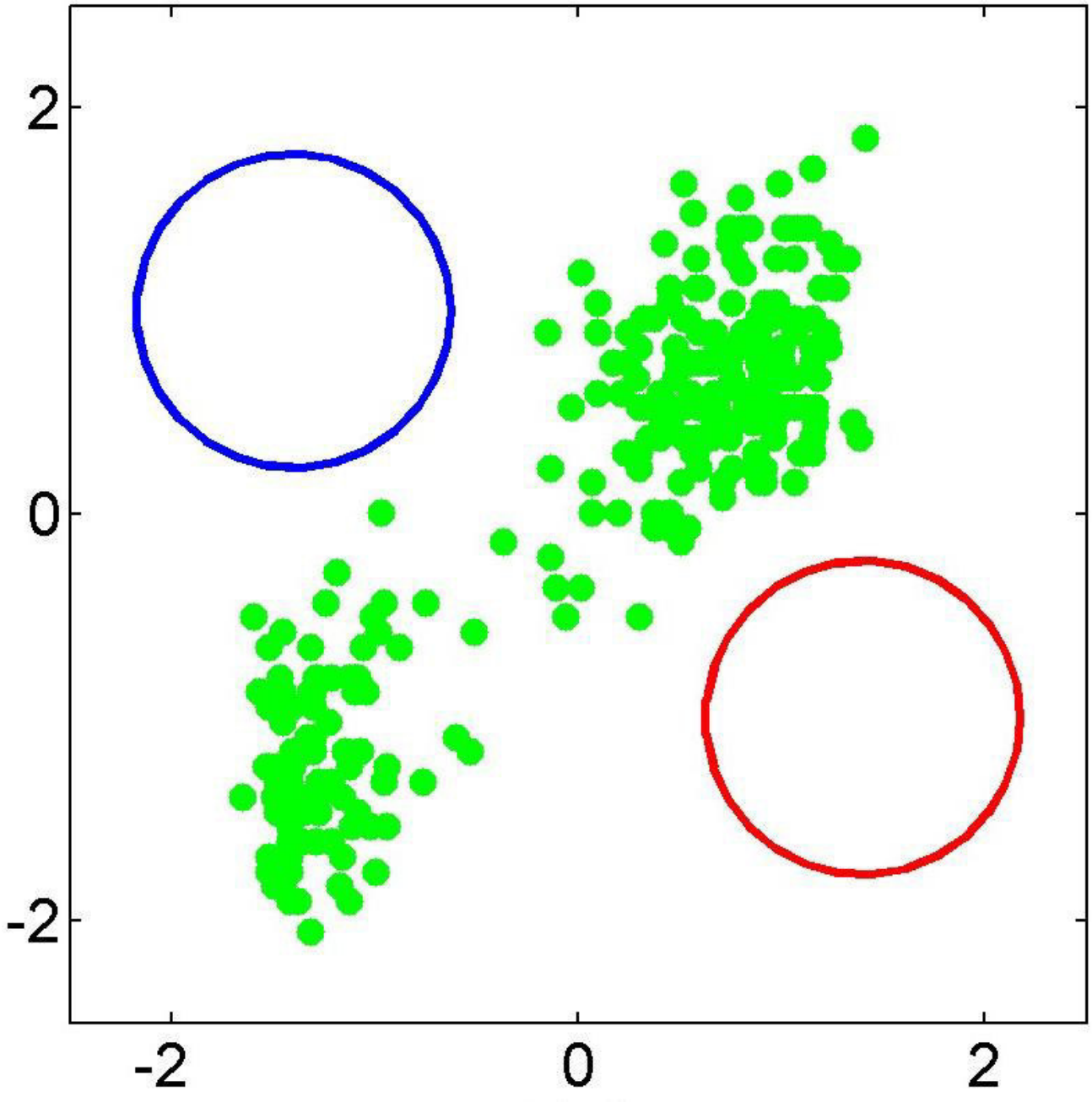
- For mixing coefficients use a Lagrange multiplier to give

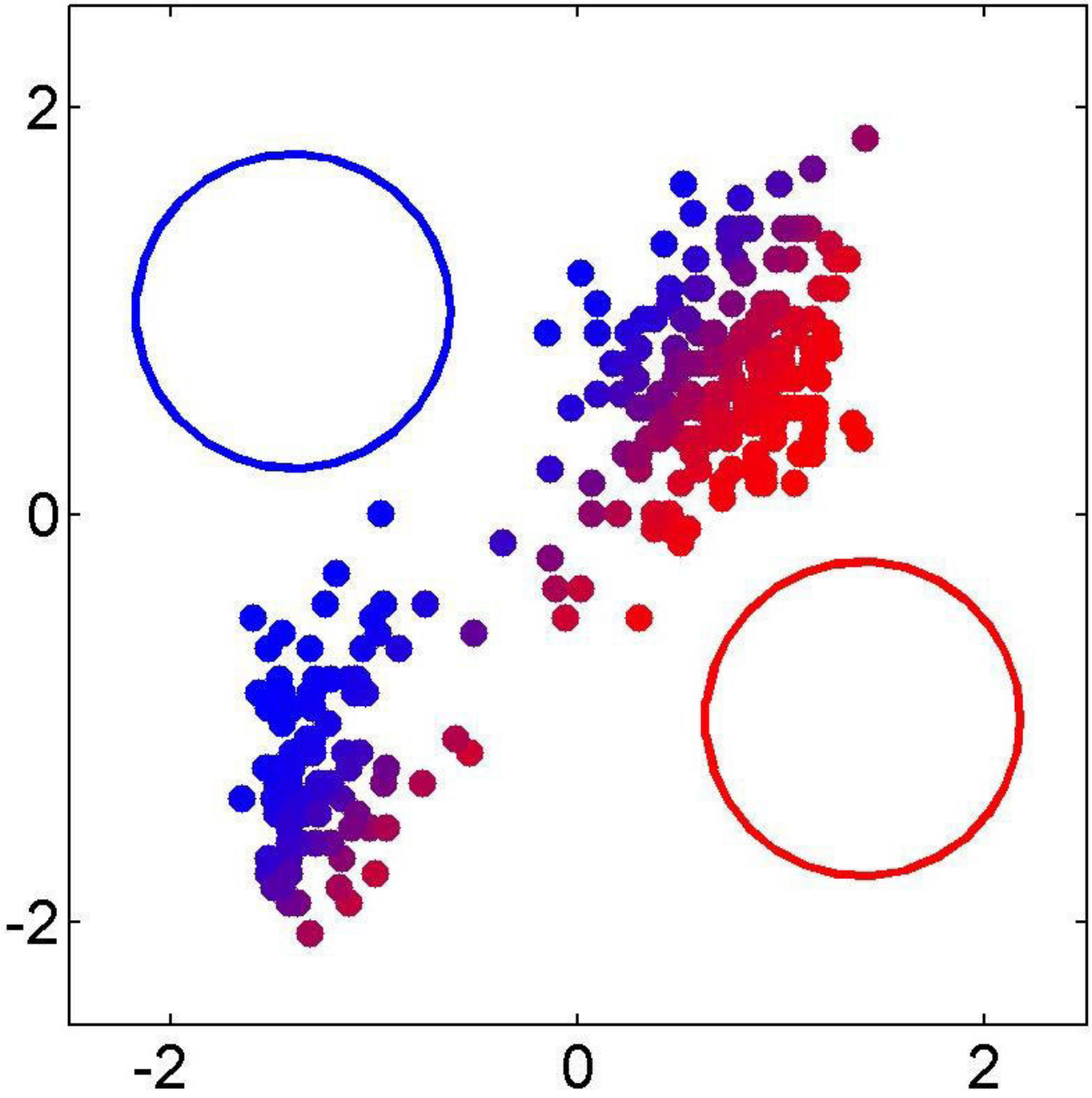
$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)$$

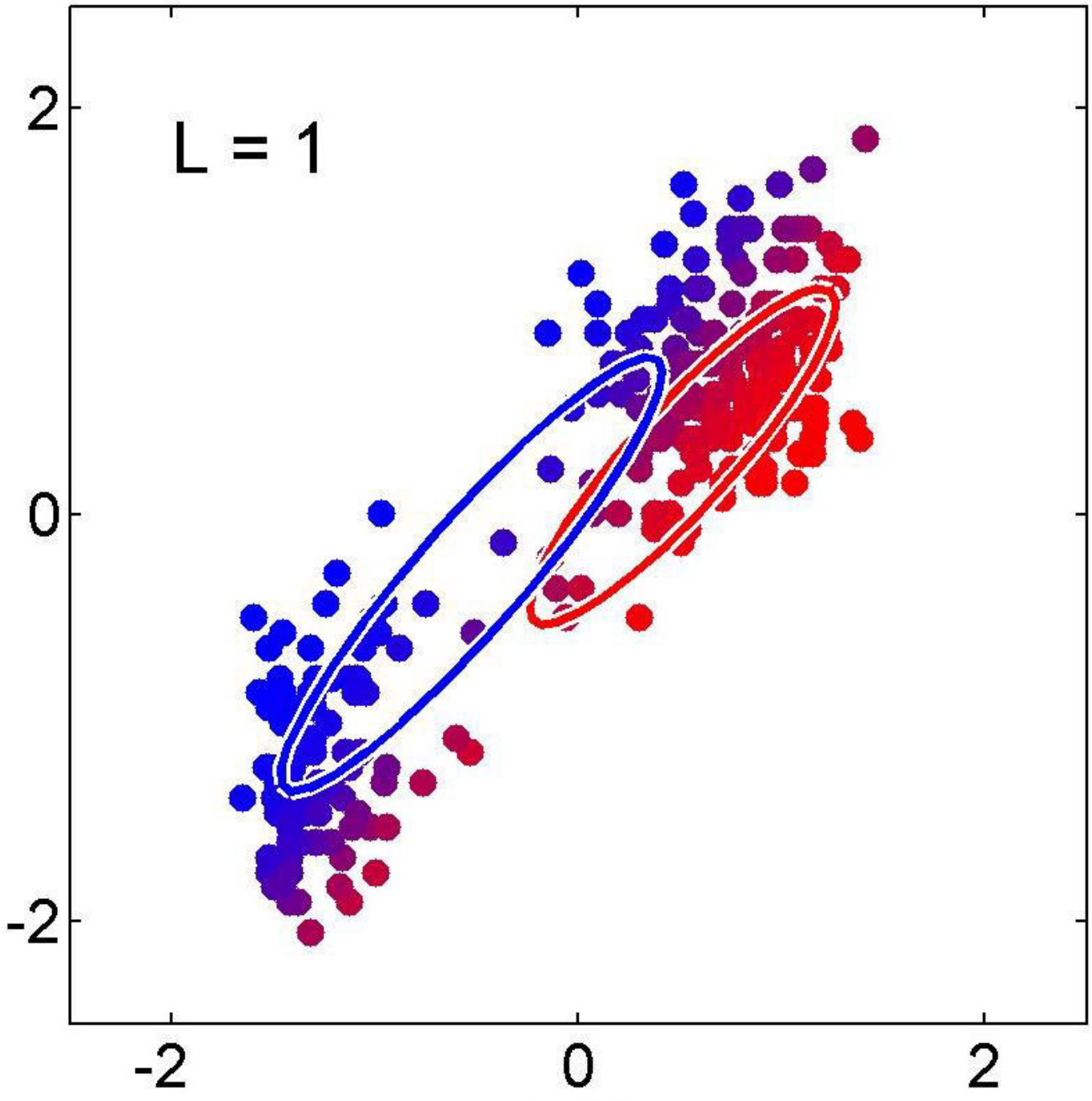
Average responsibility which component j takes for explaining the data points.

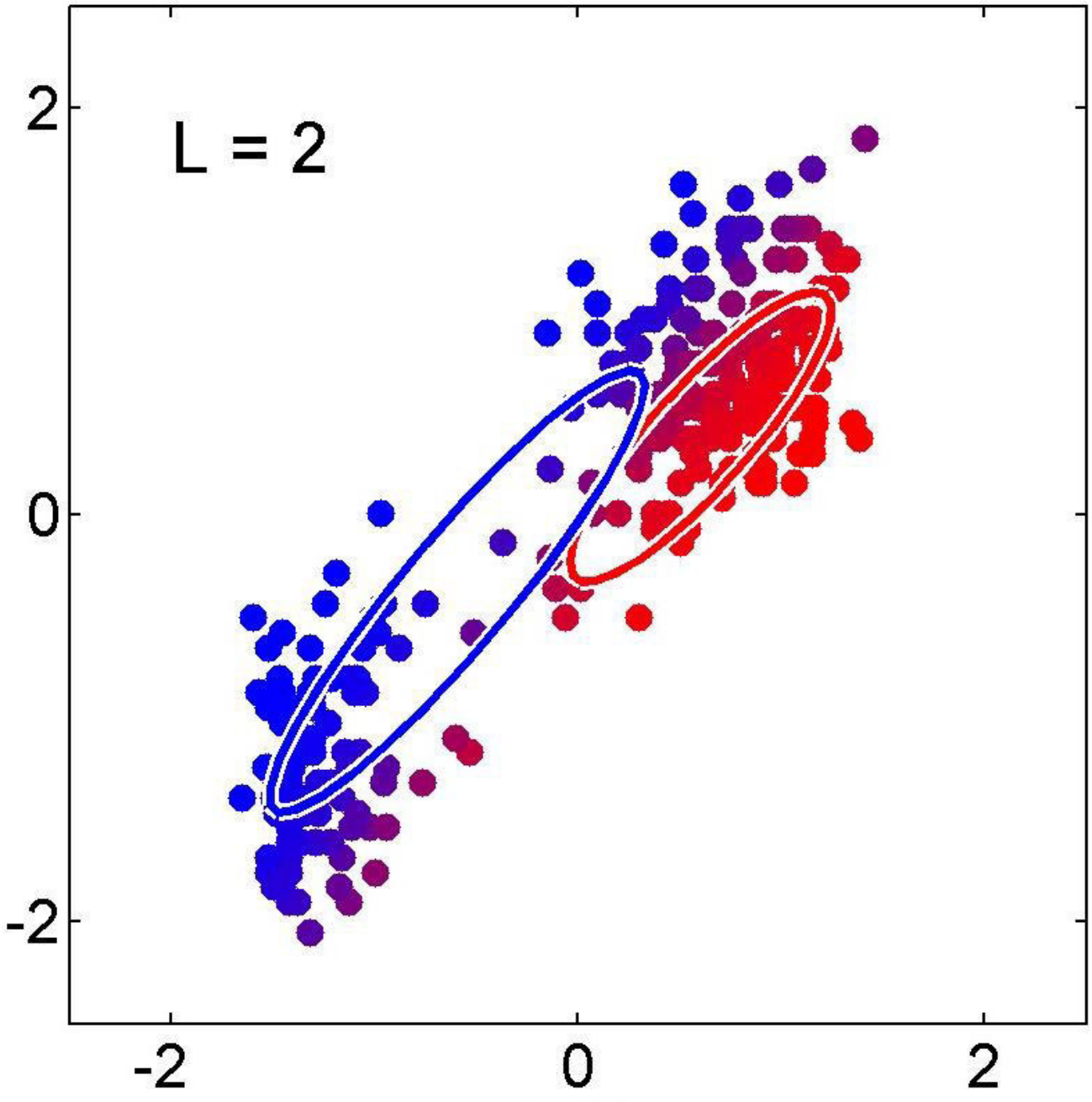
EM Algorithm – Informal Derivation

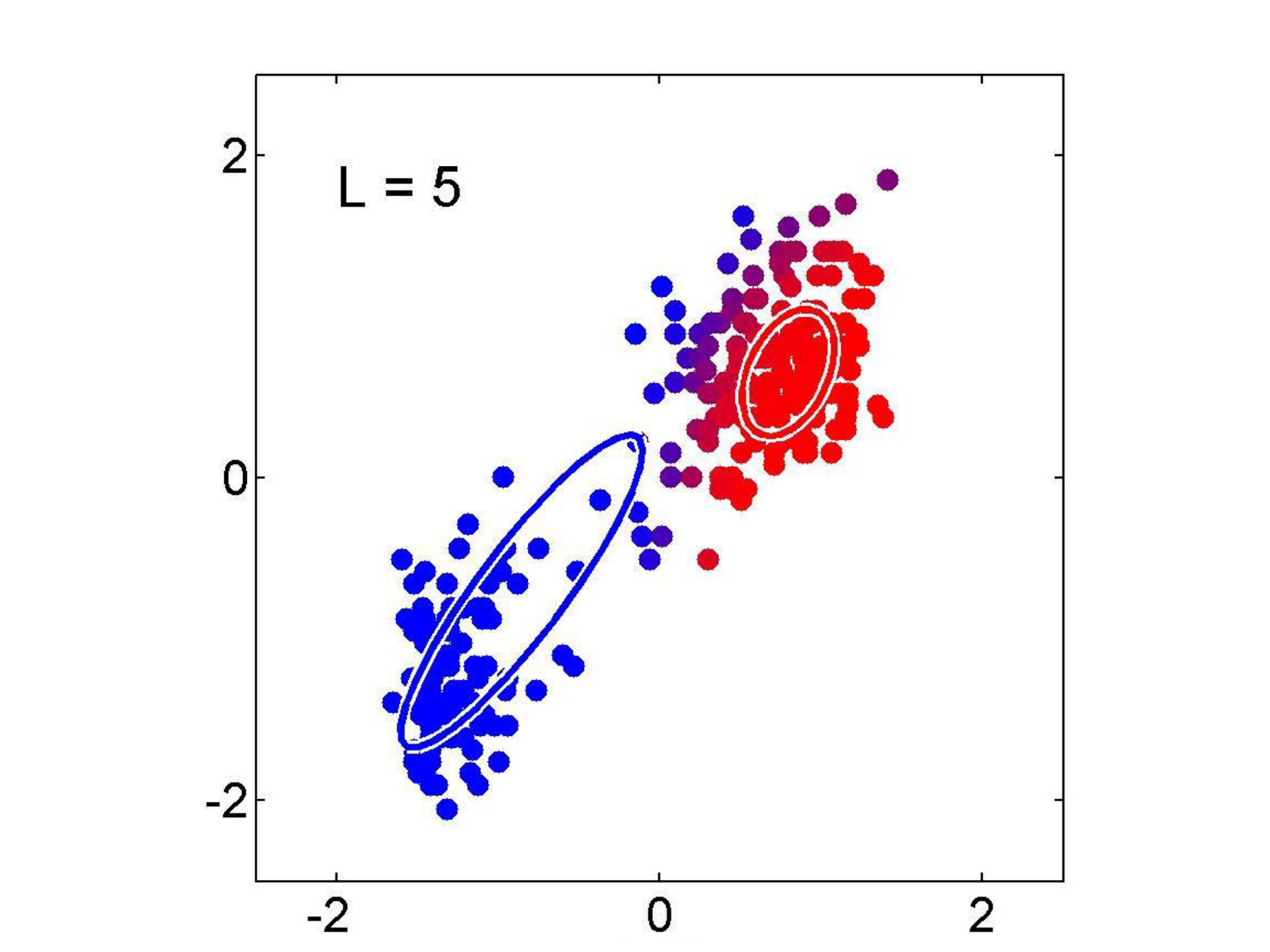
- The solutions are not closed form since they are coupled
- Suggests an iterative scheme for solving them:
  - Make initial guesses for the parameters
  - Alternate between the following two stages:
    1. E-step: evaluate responsibilities
    2. M-step: update parameters using ML results

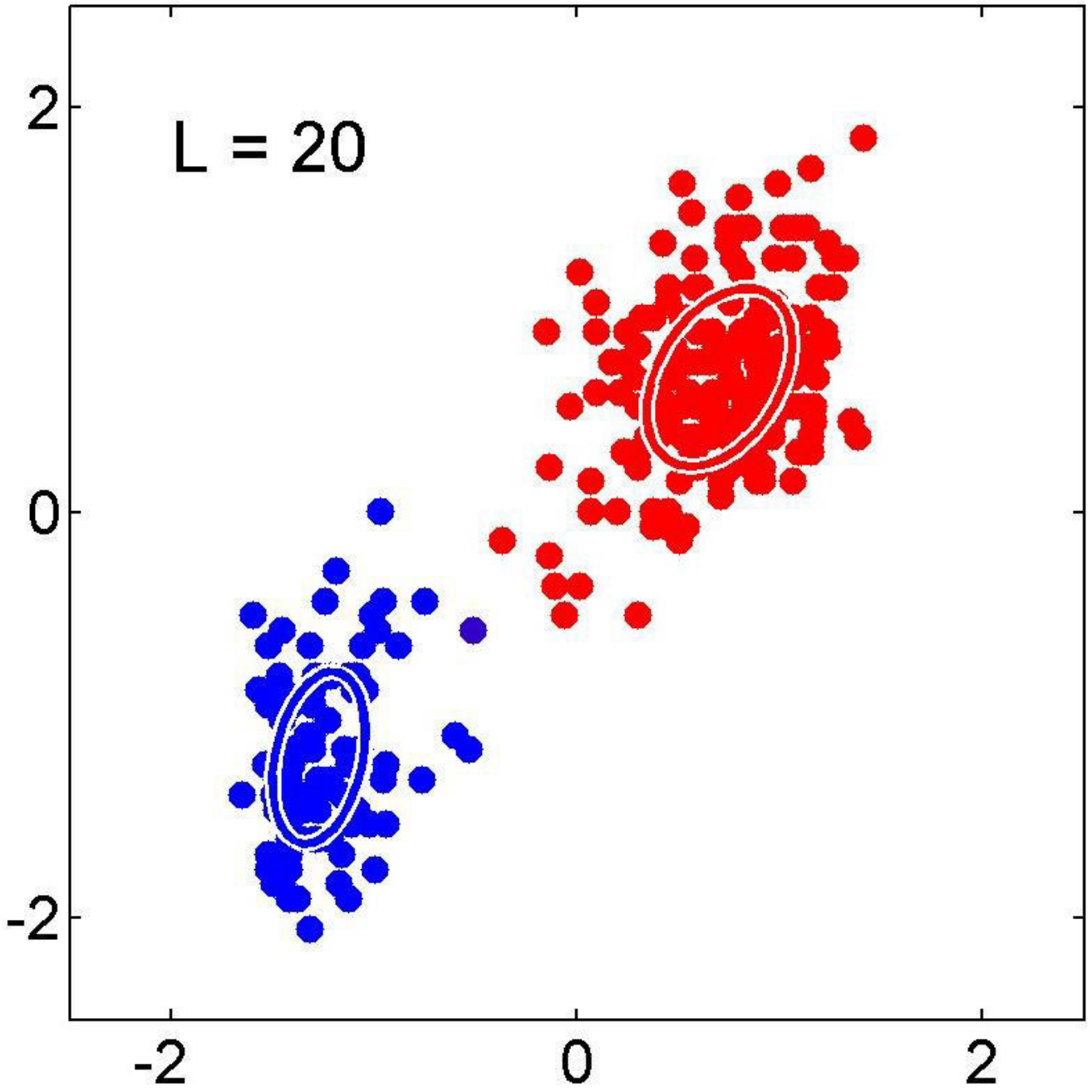

2
0
-2
-2
0
2


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

General EM Algorithm

Repeat until convergence {

- E-step:

  For each i set $Q_i(z_i) := p(z_i | x_i; \theta)$

- M-step:

$$\theta := \arg\max_{\theta} \underbrace{\sum_{i=1}^{m} \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}}_{\text{Lower bound on } l(\theta)}$$

}

EM for MoG revisited

- For $1 \le i \le N,\ 1 \le j \le K$, define hidden variables z_{ij}

$$z_{ij} = \begin{cases} 1 & \text{if sample i was generated by component } \mathbf{k} \\ 0 & \text{otherwise} \end{cases}$$

- z_{ij} are indicator random variables, they indicate which Gaussian component generated sample x_i

- Let $z_i = \{z_{i1}, ..., z_{iK}\}$ indicator r.v. correspond to sample x_i .

  We say that $z_i = k$, when its k'st coordinate is 1 and the rest are 0.

- Conditioned on z_i, distribution of x_i is Gaussian

$$p(\mathrm{x}_i \mid z_i = k) \sim N(\mu_k, \Sigma_k)$$

EM for MoG revisited

E-step:

$$Q_i(z_i = k) = p(z_i = k | x_i; \mu, \Sigma, \pi)$$

$$= \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_i \mid \mu_j, \Sigma_j)} \quad = \gamma_k(x_i)$$

EM for MoG revisited

M-step: $\max_{\mu,\Sigma,\pi} \sum_{i=1}^{m} \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k(x_i) \log \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\gamma_k(x_i)}$$

$$\nabla_\mu(\ldots) \overset{\text{set}}{=} 0 \Longrightarrow \mu_k = \frac{\sum_{i=1}^{N} \gamma_k(x_i) x_i}{\sum_{i=1}^{N} \gamma_k(x_i)}$$

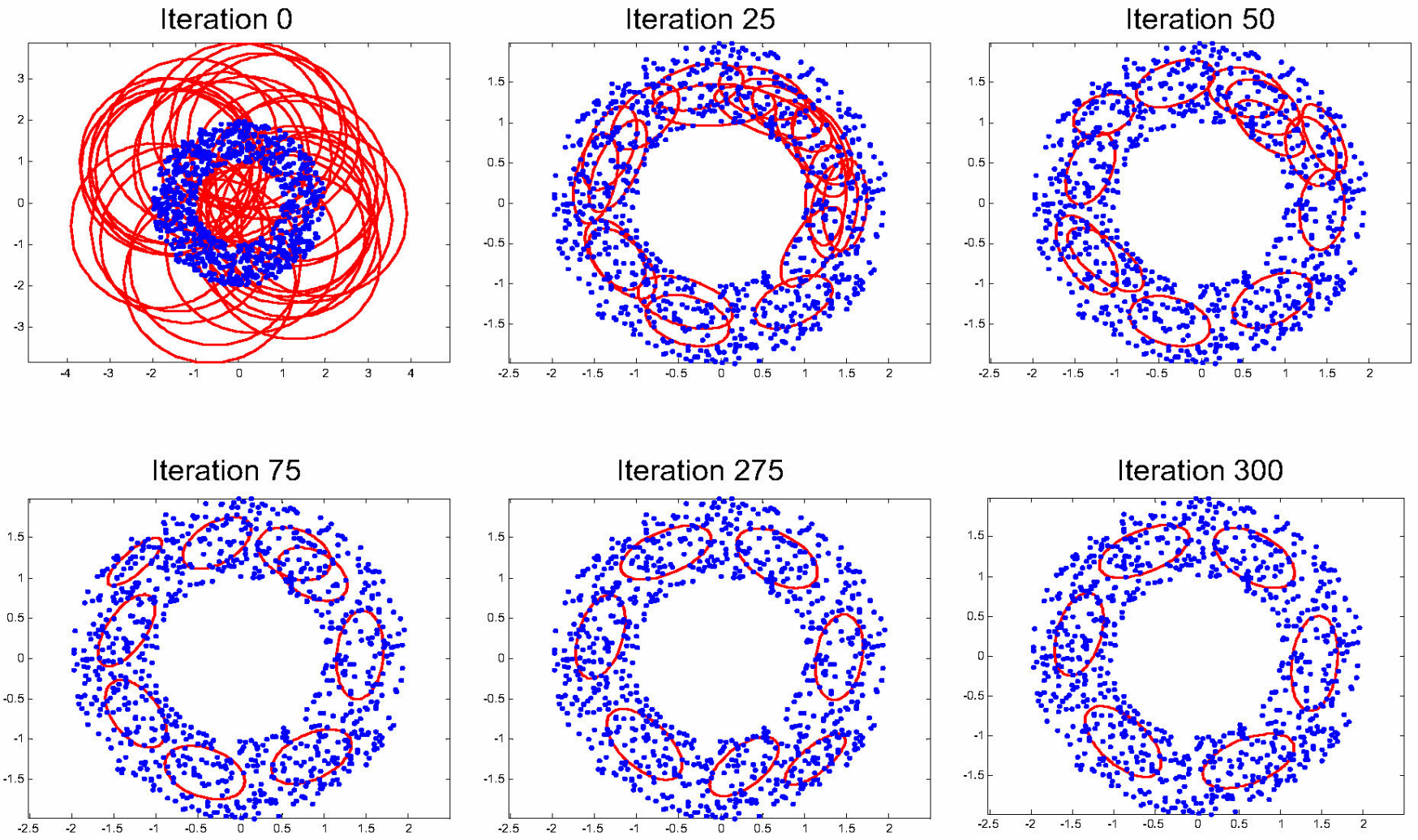
Similarly,

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_k(x_i), \qquad \Sigma_k = \frac{\sum_{i=1}^{N} \gamma_k(x_i)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_k(x_i)}$$

EM Example

- Example from R. Gutierrez-Osuna
- Training set of 900 examples forming an annulus
- Mixture model with **m** = 30 Gaussian components of unknown mean and variance is used
- Training:
  - Initialization:
    - means to 30 random examples
    - covaraince matrices initialized to be diagonal, with large variances on the diagonal (compared to the training data variance)
  - During EM training, components with small mixing coefficients were trimmed
    - This is a trick to get in a more compact model, with fewer than 30 Gaussian components

EM Example



from R. Gutierrez-Osuna