Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt
Csaba Domokos

Winter Semester 2015/2016

Introduction

We are interested in a method to find *maximum likelihood estimator* of a **parameter** θ of a **probability distribution** $p(x \mid \theta)$.
Reminiscent of naming conventions:

$$\underbrace{p(\theta \mid x)}_{\text{Posterior probability}} = \frac{p(x \mid \theta) p(\theta)}{p(x)} \propto \underbrace{p(x \mid \theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior probability}}.$$

We are given finite amount of **measurement** (or observation data) $x_1, x_2, \ldots$, and also know the probability distribution $p(x \mid \theta)$. The maximum likelihood estimate of θ is given by

$$\hat{\theta} \in \operatorname*{argmax}_{\theta} p(x \mid \theta) \; .$$

A possible solution: **Expectation Maximization Algorithm**, which iteratively makes guesses about the data x, and iteratively maximizes $p(x \mid \theta)$ over θ.

Multivariate Gaussian distribution 4 / 39

Multivariate Gaussian distribution

Assume a D-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_D)$, i.e. a vector whose components are random variables, with the joint density function

$$p(x_1, \ldots, x_D) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) .$$

$\mathbf{X}$ is said to have **multivariate Gaussian (or Normal) distribution**, with parameters $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ assuming that $\boldsymbol{\Sigma}$ is positive definite.

Reminder: A symmetric $\mathbf{A} \in \mathbb{R}^{n \times n}$ matrix is said to be **positive definite**, if $\mathbf{u}^T \mathbf{A} \mathbf{u} > 0$ for all $\mathbf{u} \in \mathbb{R}^n$.

μ is called the **mean vector** and $\boldsymbol{\Sigma}$ is called the **covariance matrix**. We often use the notation $\mathbf{X} \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denoting $\mathbf{X}$ has Normal distribution.

Note that the Gaussian distribution has many important analytical properties. For example, it is “closed” under marginalization.

Maximum likelihood for the Gaussian

Suppose we have a set of **independent and identically distributed** (*i.i.d.*) data samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from a Gaussian distribution. The data set can be represented as an $\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{bmatrix}^T = \mathbf{X} \in \mathbb{R}^{N \times D}$ matrix.

We are interested in to estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by maximum likelihood. The **log-likelihood function** is given by

$$\begin{aligned}\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln \prod_{n=1}^{N} p(\mathbf{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{n=1}^{N} \ln \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right) \\ &= \sum_{n=1}^{N} \left(-\frac{1}{2}\ln\left((2\pi)^D|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right)\end{aligned}$$

Maximum likelihood for the Gaussian (cont.)

$$\begin{aligned}\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^{N} \left(-\frac{1}{2}\ln\left((2\pi)^D|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right) \\ &= \sum_{n=1}^{N} \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right) \\ &= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).\end{aligned}$$

Maximum likelihood for $\boldsymbol{\mu}$

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

Setting the derivative of the log-likelihood function w.r.t. $\boldsymbol{\mu}$ to 0, we obtain

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{-1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial \boldsymbol{\mu}}\left(\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= -\frac{1}{2}\sum_{n=1}^{N}\left(-\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1} - \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1} - 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0 \quad \Rightarrow \quad \boxed{\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n}.
\end{aligned}$$

The maximum likelihood estimator for $\boldsymbol{\mu}$ is simply given by the center of the mass of the data, i.e. the sample mean.

Maximum likelihood for $\boldsymbol{\Sigma}$

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

Setting the derivative of the log-likelihood function w.r.t. $\boldsymbol{\Sigma}$ to 0, we obtain

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial \boldsymbol{\Sigma}}\big((\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\big) \\
&= -\frac{N}{2}\frac{1}{|\boldsymbol{\Sigma}|}|\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\sum_{n=1}^{N} -\boldsymbol{\Sigma}^{-T}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T} \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

Maximum likelihood for Σ (cont.)

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} = 0$$

$$\Rightarrow \boxed{\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T}.$$

This is, by definition, called the sample **covariance matrix** of the data.

The geometry of the Multivariate Gaussian distribution

Let us consider the quadratic form

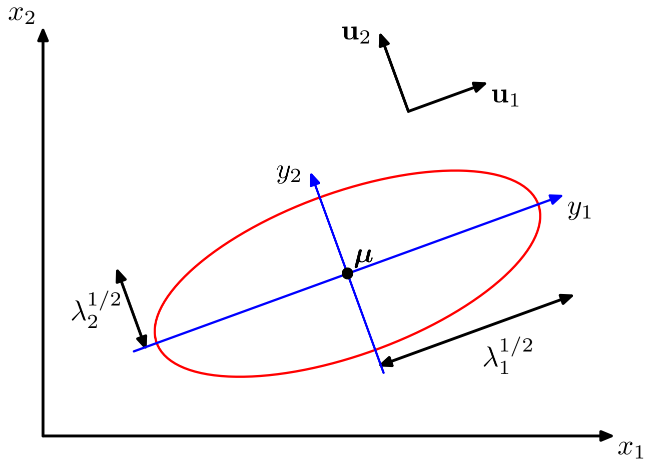
$$\Delta = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \ ,$$

which is called the **Mahalanobis-distance** from $\boldsymbol{\mu}$ to $\mathbf{x}$. In case of $\boldsymbol{\Sigma} = \boldsymbol{I}$ we get the Euclidean-distance. Note that the quantity Δ^2 appears in the exponent in the density function.

The covariance matrix $\boldsymbol{\Sigma}$ is a real, symmetric matrix, hence its

- eigenvalues $\lambda_1, \dots, \lambda_D$ will be real,
- eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_D \in \mathbb{R}^D$ from an orthonormal set.

Therefore $\boldsymbol{\Sigma}^{-1}$ can be written as



2D Gaussian

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \ , \quad \text{which yields} \quad \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \ , \quad \text{where} \quad y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu}) \ .$$

Two dimensional Gaussian distribution

The density function of the two dimensional Gaussian distribution is given by

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}\begin{bmatrix}x_1 - \mu_1 & x_2 - \mu_2\end{bmatrix}\mathbf{\Sigma}^{-1}\begin{bmatrix}x_1 - \mu_1 \\ x_2 - \mu_2\end{bmatrix}\right) ,$$

where $\boldsymbol{\mu} = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}$ and $\mathbf{\Sigma} = \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$ for $\sigma_1, \sigma_2 > 0$ and $-1 < \rho < 1$.

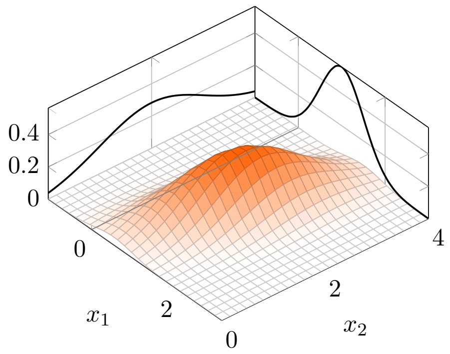
Note that this density function can be written equivalently as

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right)}.$$

Example: 2D Gaussian and its marginals

Assume $\mathbf{X} \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 1 \end{bmatrix}$ that is $\rho = 0.5$. The density function is given by

$$p(x_1, x_2) = \frac{1}{\pi\sqrt{0.75}} \exp\left(-\frac{2(x_1-1)^2}{3} + \frac{4(x_1-1)(x_2-2)}{3} - \frac{(x_2-2)^2}{3}\right) ,$$



and the marginal distributions are defined by

$$p_{X_1}(x_1) = \frac{1}{0.5\sqrt{2\pi}} \exp\left(-\frac{(x_1-1)^2}{0.5}\right) ,$$

$$p_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2-2)^2}{2}\right) .$$

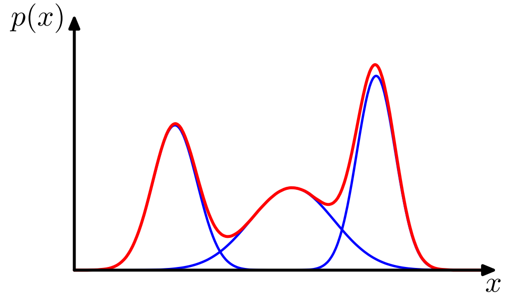
Mixtures of Gaussians

While the Gaussian distribution has some important analytical properties, it suffers from limitations when it comes to modelling real data sets. However the **linear combination of Gaussians** can give rise to very complex densities.

We consider a superposition of K Gaussian densities

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

is called a **mixture of Gaussians**. The parameters π_k are called **mixing coefficients**.



Mixture of three Gaussians

$$1 = \int_{\mathbb{R}^D} p(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^D} \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \mathrm{d}\mathbf{x} = \sum_{k=1}^{K} \pi_k \, .$$

All the density functions are non-negative, hence $\pi_k \geqslant 0$, therefore

$$0 \leqslant \pi_k \leqslant 1 \quad \text{for all} \quad k = 1, \ldots, K \, .$$

Mixtures of Gaussians (cont.)

We are provided with the following joint distribution

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

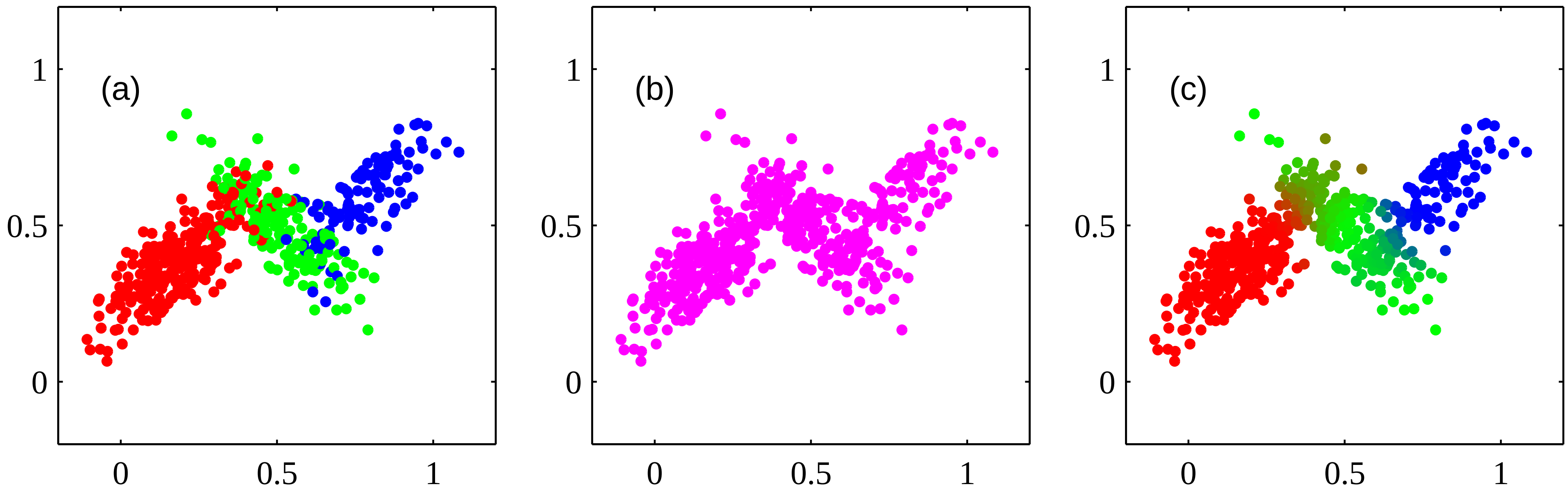
One can view

- $\pi_k = p(k)$ as the prior probability of picking the k^{th} component;
- $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\mathbf{x} \mid k)$ as the probability of $\mathbf{x}$ conditioned on k.

The posterior probabilities $p(k \mid \mathbf{x})$ are also known as **responsibilities**, denoted by $\gamma_k(\mathbf{x})$.

$$\begin{aligned}\gamma_k(\mathbf{x}) \triangleq p(k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid k) p(k)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid k) p(k)}{\sum_{l=1}^{K} p(l, \mathbf{x})} &= \frac{p(k) p(\mathbf{x} \mid k)}{\sum_{l=1}^{K} p(l) p(\mathbf{x} \mid l)} \\ &= \frac{\pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \; .\end{aligned}$$

Example: Mixture of three 2D Gaussians

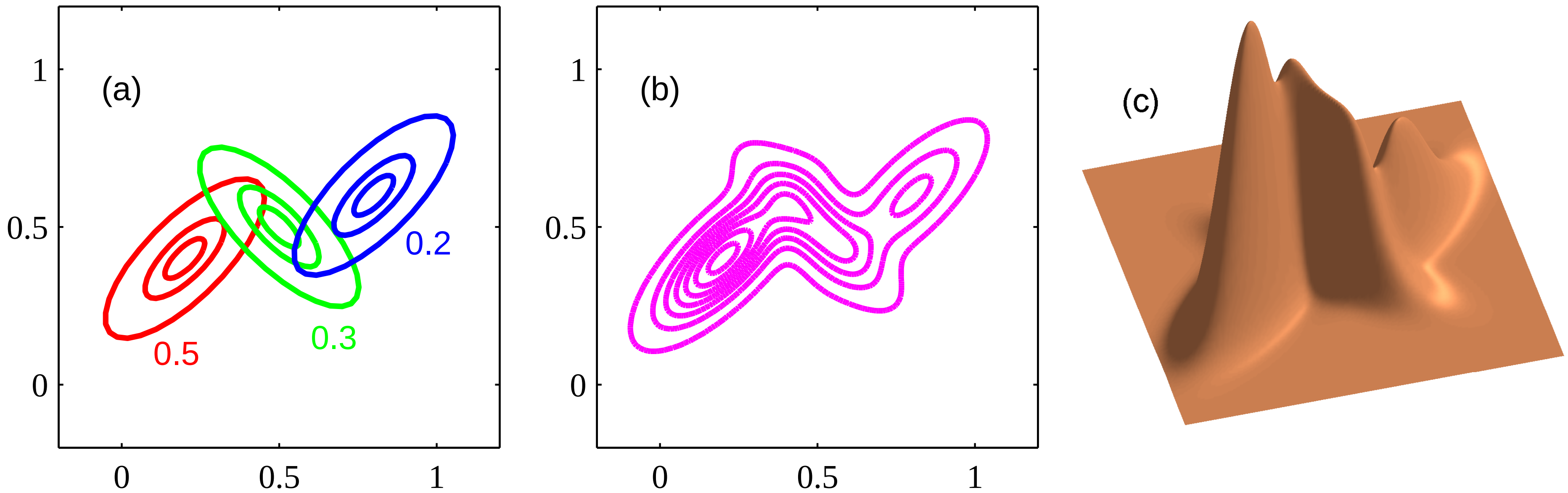


Samples from $p(k)p(\mathbf{x} \mid k)$

Samples from $p(\mathbf{x})$

Values of responsibilities coded by colors $\begin{bmatrix}\gamma_1(\mathbf{x}) & \gamma_2(\mathbf{x}) & \gamma_3(\mathbf{x})\end{bmatrix}$.

Example: Mixture of three 2D Gaussians



Iso-countours for each component

Iso-contours of $p(\mathbf{x})$

Surface plot of $p(\mathbf{x})$

Maximum likelihood for mixture of Gaussians

Suppose we have a set of *i.i.d.* data samples $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ drawn from a mixture of Gaussians. The data set is also represented by $\mathbf{X} \in \mathbb{R}^{N \times D}$.

The goal is to find the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, specifying the model from which the samples $\mathbf{x}_n$ have most likely been drawn. We may find the parameters which maximize the *likelihood function*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} p(\mathbf{X} \mid \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{n=1}^{N} p(\mathbf{x}_n \mid \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,.$$

To simplify the optimization we use the **log-likelihood function** $\mathcal{L}(\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,.$$

Note that there is no closed-form solution for this model $\Rightarrow$ Iterative solution.

Maximum likelihood for μ

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{s.t.} \quad \pi_k \geqslant 0, \sum_{k=1}^{K} \pi_k = 1 \,.$$

We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \frac{1}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

Maximum likelihood for μ (cont.)

Let us now consider the derivative of a Gaussian only

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) =& \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}}\frac{\partial}{\partial \boldsymbol{\mu}_k}\exp\Big(-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\Big)\\ =&\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}}\exp\Big(\frac{-1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\Big)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\\ =&\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) \ .\end{aligned}$$

By substituting back we get

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \underbrace{\boxed{\frac{\pi_k\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K}\pi_l\ \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}}}_{\gamma_{nk}\triangleq\gamma_k(\mathbf{x}_n)}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) \ .$$

Maximum likelihood for μ (cont.)

Setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$ to 0, we obtain

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N}\gamma_{nk}\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) =& 0\\ \sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k) =& 0\\ \boldsymbol{\mu}_k =& \frac{\sum_{n=1}^{N}\gamma_{nk}\ \mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{nk}} \ .\end{aligned}$$

Maximum likelihood for Σ

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{s.t.} \quad \pi_k \geqslant 0, \sum_{k=1}^{K} \pi_k = 1 \, .$$

We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\pi_k}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Let us now consider the derivative of a Gaussian only

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \, .$$

Maximum likelihood for Σ (cont.)

We calculate the following derivatives:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{-1}{2} |\boldsymbol{\Sigma}_k|^{-\frac{3}{2}} |\boldsymbol{\Sigma}_k| \boldsymbol{\Sigma}_k^{-1} = \frac{-\boldsymbol{\Sigma}_k^{-1}}{2\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \, .$$

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \\
&= \exp\Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \\
&= \exp\Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \frac{-1}{2} (-\boldsymbol{\Sigma}^{-T}) (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-T} \\
&= \frac{1}{2} \exp\Big(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \Big) \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \, .
\end{aligned}$$

Maximum likelihood for Σ (cont.)

Now we are at the position to calculate the derivative of a Gaussian w.r.t. $\boldsymbol{\Sigma}$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
=&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\Big) \\
=&\frac{-\boldsymbol{\Sigma}_k^{-1}}{2\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\Big) \\
&+ \frac{1}{2} \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\Big) \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \\
=&-\frac{1}{2}\boldsymbol{\Sigma}_k^{-1} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \frac{1}{2}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \,.
\end{aligned}
$$

Maximum likelihood for Σ (cont.)

Setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$ to 0, we obtain

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= -\frac{1}{2}\sum_{n=1}^{N} \frac{\boldsymbol{\Sigma}_k^{-1}\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \\
&\quad + \frac{1}{2}\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_k^{-1}}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \\
&= \frac{-\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_{nk} + \frac{\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_k^{-1} = 0 \\
\Rightarrow \quad \boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} \gamma_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}} \; .
\end{aligned}
$$

Maximum likelihood for $\boldsymbol{\pi}$

To integrate the conditions on $\boldsymbol{\pi}$ we use the Lagrange multiplier method

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \lambda(1 - \sum_{k=1}^{K} \pi_k) \, .$$

Setting the derivative w.r.t. π_k to 0, we obtain

$$\begin{aligned}
\sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} - \lambda &= 0 \\
\sum_{n=1}^{N} \frac{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} &= \lambda \sum_{l=1}^{K} \pi_l \quad \Rightarrow \quad N = \lambda \\
\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}}_{\gamma_{nk}} - \pi_k N = 0 \quad &\Rightarrow \quad \pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N} \, .
\end{aligned}$$

The EM Algorithm for mixtures of Gaussians

Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k

repeat

E step. Evaluate the responsibilities using the current parameter values

$$\gamma_{nk} = \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

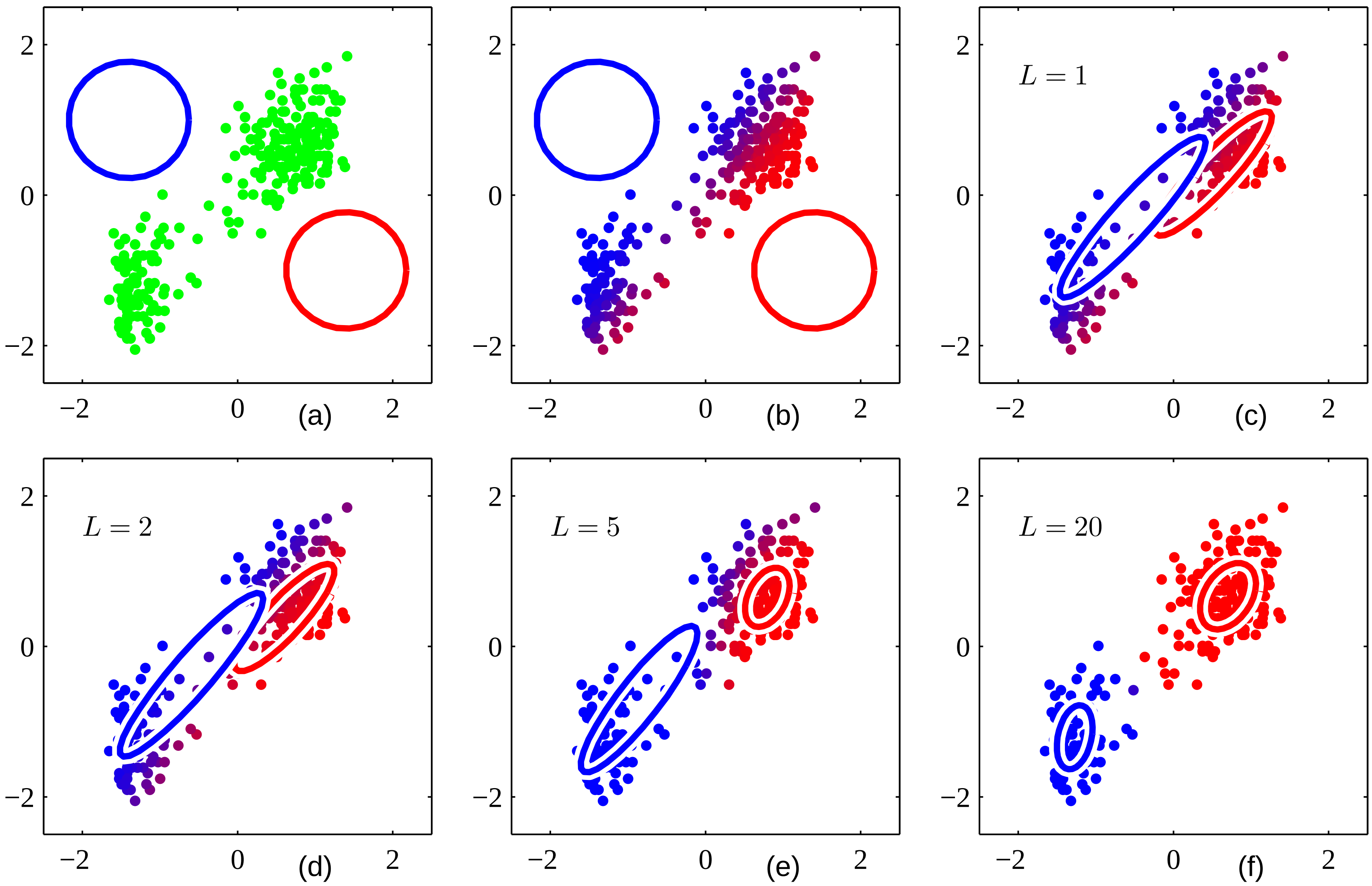
M step. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}, \quad \boldsymbol{\Sigma}_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T}{\sum_{n=1}^{N} \gamma_{nk}}$$

$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

until convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta})$

Example



2
0
−2
−2
0
(a)
2
2
0
−2
−2
0
(b)
2
L = 1
2
0
−2
−2
0
(c)
2
L = 2
2
0
−2
−2
0
(d)
2
L = 5
2
0
−2
−2
0
(e)
2
L = 20
2
0
−2
−2
0
(f)
2

Expectation 30 / 39

Expectation

The expectation of a random variable is intuitively the long-run average value of repetitions of the experiment it represents.

Let X be a discrete random variable taking values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots,$ respectively. The **expectation** (or **expected value**) of X is defined as

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i \,,$$

assuming that this series is absolute convergent (that is $\sum_{i=1}^{\infty} |x_i| p_i$ is convergent).

Example: throwing two "fair" dice and the value of X is is the sum the numbers showing on the dice.

$$\begin{aligned}\mathbb{E}[X] =& 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + 5\frac{4}{36} + 6\frac{5}{36} \\ &+ 7\frac{6}{36} + 8\frac{5}{36} + 9\frac{4}{36} + 10\frac{3}{36} + 11\frac{2}{36} + 12\frac{1}{36} = 7 \,.\end{aligned}$$

Expectation (cont.)

Let X be a (continuous) random variable with density function $f(x)$. The **expectation** of X is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x) \mathrm{d}x \ ,$$

assuming that this integral is absolutely convergent (that is the value of the integral $\int_{-\infty}^{\infty} |x| \cdot f(x) \mathrm{d}x$ is finite).

Suppose a random variable X with density function $f(x)$. Let $g(x)$ be a measurable function. The **expected value of the function $g(x)$** is defined as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \mathrm{d}x \ ,$$

assuming that this integral is absolutely convergent.

Conditional expectation

Let (X, Y) be a *discrete random vector*. The **conditional expectation** of X given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[X \mid Y = y] = \sum_{i=1}^{\infty} x_i P(X = x_i \mid Y = y) \ ,$$

assuming that this series is absolute convergent.

Let (X, Y) be a (continuous) random vector with joint density function $f_{XY}(x, y)$. The **conditional expectation** of X given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x \mid Y = y) \mathrm{d}x \ ,$$

assuming that this integral is absolute convergent.

Conditional expectation (cont.)

Suppose a (continuous) random vector (X, Y) with joint density function $f_{XY}(x, y)$. Let $g(x)$ be a measurable function. The **conditional expectation of the function $g(x)$** given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) \cdot f_{X|Y}(x \mid Y = y) \mathrm{d}x \; ,$$

assuming that this integral is absolute convergent.

EM algorithm 35 / 39

Expectation Maximization algorithm 35 / 39

Latent variables

Suppose we are given a set of *i.i.d.* data samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ observed data $\mathbf{X} \in \mathbb{R}^{N \times D}$ represented by $\mathbf{X} \in \mathbb{R}^{N \times D}$ matrix. The model parameters are given by $\boldsymbol{\theta}$. Moreover, we assume some unknown (or **latent**) variables denoted by $\mathbf{Z}$. The log-likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{X} \mid \boldsymbol{\theta}) \; .$$

We consider the following expectation

$$\mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}] = \sum_{\mathbf{Z}} \ln p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$$
$$\triangleq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \; .$$

The general EM algorithm

Choose an initial setting for the parameters $\boldsymbol{\theta}^{(0)}$
$t \rightarrow 0$
repeat
  $t \rightarrow t + 1$
  E step. Evaluate $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)})$
  M step. Evaluate $\boldsymbol{\theta}^{(t)}$ given by

$$\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$$

  where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)}) \ln p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$$

until convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta})$

The General EM Algorithm

- The EM algorithm is not limited to Mixtures of Gaussians but can also be applied to other probability density functions. How to choose the value for K is an open question?
- The algorithm does not necessary yield global maxima. In practice, it is restarted with different initializations and the result with the highest log likelihood after convergence is chosen.
- The estimated covariance matrices can become singular if the data points lie on a lower dimensional subspace. A possible remedy is to add a constant matrix $\varepsilon\mathbf{I}$ in each step to the covariance matrix.

Literature

1. Christopher Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006. Note: Chapter 9.
2. Yihua Chen and Maya R. Gupta. **EM Demystified: An Expectation-Maximization Tutorial**. TechRep: UWEETR-2010-0002, University of Washington, Seattle, WA, USA, 2009.
3. Frank Dellaert. **The Expectation Maximization Algorithm**. TechRep: GIT-GVU-02-20, Georgia Institute of Technology, Atlanta, GA, USA, 2002.