Combinatorial Optimization in Computer Vision (IN2245)

Frank R. Schmidt
Csaba Domokos

Winter Semester 2015/2016

---

5. The Expectation Maximization Algorithm

Introduction

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

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)} \propto p(x \mid \theta) \ \ p(\theta) \ .$$

Posterior probability Likelihood Prior probability

We are given finite amount of **measurement** (i.e. observed data) $x_1, x_2, \dots$, 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

Multivariate Gaussian distribution

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

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

$\mathbf{X}$ is said to have **multivariate Gaussian (or Normal) distribution** with parameters $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\mathbf{\Sigma} \in \mathbb{R}^{D \times D}$ assuming that $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ is called the **covariance matrix**. We often use the notation $\mathbf{X} \sim \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma})$ 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 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 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 to estimate the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ with the maximum likelihood framework. The **log-likelihood function** is given by

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

Maximum likelihood for the Gaussian (cont.)

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

Maximum likelihood for μ *

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

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

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

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 Σ *

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

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

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

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

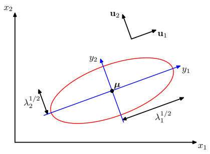
Let us consider the following form

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

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



2D Gaussian

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

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

Two dimensional Gaussian distribution *

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

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

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

Note that this density function can be written equivalently as

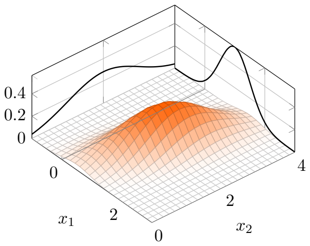
$$\boxed{p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right)}}.$$

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

$$p(x_1, x_2) = \frac{1}{\pi\sqrt{0.75}}\exp\left(-\frac{8(x_1 - 1)^2}{3} + \frac{4(x_1 - 1)(x_2 - 2)}{3} - \frac{2(x_2 - 2)^2}{3}\right) ,$$



and the marginal distributions are defined by

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

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

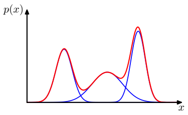
Mixtures of Gaussians

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.

Let us consider a superposition of K Gaussian densities

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



Mixture of three Gaussians

is called a **mixture of Gaussians**.

The parameters π_k are called **mixing coefficients**.

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

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 joint distribution

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

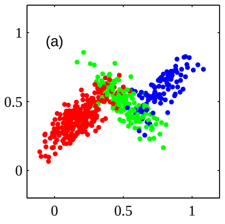
One can view

- $\pi_k = p(k)$ 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.

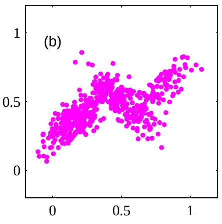
The posterior probabilities $p(k \mid \mathbf{x})$, a.k.a. **responsibilities**, are denoted by $\gamma_k(\mathbf{x})$ and show the probability that a given sample $\mathbf{x}$ belongs to the k^{th} component.

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

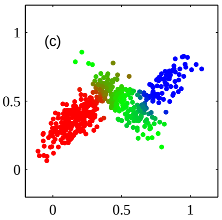
Example: Mixture of three 2D Gaussians



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

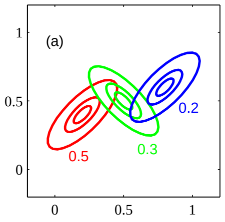


Samples from $p(\mathbf{x})$

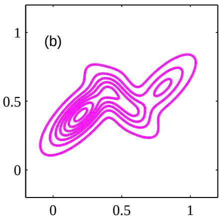


Values of responsibilities coded by `RGB colors` $[\gamma_1(\mathbf{x}) \quad \gamma_2(\mathbf{x}) \quad \gamma_3(\mathbf{x})]$.

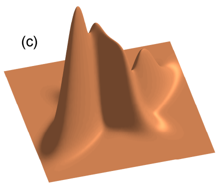
Example: Mixture of three 2D Gaussians



Iso-countours of 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 $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn from a mixture of Gaussians. The data set is 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** $\mathcal{L}(\boldsymbol{\theta})$

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

Note that there is no closed-form solution for this model ⇒ iterative solution.

Maximum likelihood for μ *

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

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

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

By substituting back we get

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

Maximum likelihood for μ (cont.) *

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

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

$$\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \quad \Rightarrow \quad \boxed{\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \, \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}} .$$

Maximum likelihood for Σ *

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

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

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

Maximum likelihood for Σ (cont.) *

We calculate the following derivatives:

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

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

Maximum likelihood for Σ (cont.) *

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

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

Maximum likelihood for Σ (cont.) *

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

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

Maximum likelihood for π *

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

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

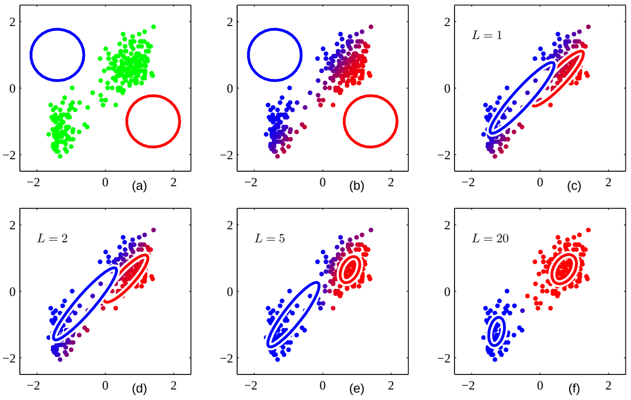
Setting the derivative w.r.t. π_k to 0, we obtain

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

The EM Algorithm for mixtures of Gaussians

```
1: Initialize the means $\boldsymbol{\mu}_k$, covariances $\mathbf{\Sigma}_k$ and mixing coefficients $\pi_k$
2: repeat
3:     E step. Evaluate the responsibilities using the current parameter values
           $\gamma_{nk} = \frac{\pi_k\,\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{l=1}^{K}\pi_l\,\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \mathbf{\Sigma}_l)}$
4:     M step. Re-estimate the parameters using the current responsibilities
           $\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^{N}\gamma_{nk}\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_{nk}}$ ,   $\mathbf{\Sigma}_k^{\text{new}} = \frac{\sum_{n=1}^{N}\gamma_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})^T}{\sum_{n=1}^{N}\gamma_{nk}}$
           $\pi_k^{\text{new}} = \frac{\sum_{n=1}^{N}\gamma_{nk}}{N}$
5: until convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta})$
```

Example



Expectation

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 $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots$, 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 $\int_{-\infty}^{\infty}|x \cdot f(x)|\mathrm{d}x = \int_{-\infty}^{\infty}|x| \cdot f(x)\mathrm{d}x$ is finite).

Suppose a random variable X with density function $f(x)$. The **expected value of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ 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 absolutely 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 absolutely convergent.

Conditional expectation (cont.)

Suppose a (continuous) random vector (X, Y) with joint density function $f_{XY}(x, y)$. The **conditional expectation of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ 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 absolutely convergent.

The Expectation Maximization algorithm

Latent variables

Suppose we are given a set of *i.i.d.* data samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ represented by the matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. The samples are drawn from a model (e.g., mixture of Gaussians) given by its parameters $\boldsymbol{\theta}$.
There are two main applications of the EM algorithm:

1. The data has missing values, due to limitations of the observation process.
2. The likelihood function can be simplified by assuming missing values.

Latent variables gathering the missing values are represented by a matrix $\mathbf{Z}$.

We generally want to maximize the posterior probability

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbf{X}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{X}) .$$

Equivalently, one can maximize the log-likelihood

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

The EM algorithm

```
Choose an initial setting for the parameters $\boldsymbol{\theta}^{(0)}$
$t \to 0$
repeat
    $t \to t + 1$
    E step. Evaluate $q^{(t-1)}(\mathbf{Z}) \triangleq p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)})$
    M step. Evaluate $\boldsymbol{\theta}^{(t)}$ given by
        $\boldsymbol{\theta}^{(t)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$
    where
        $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$
until convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathbf{X})$
```

Remarks

- The EM algorithm is not limited to Mixtures of Gaussians but can also be applied to other probability density functions.
- 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.
- One can think the EM algorithm as an **alternating minimization** procedure. Considering $G(\boldsymbol{\theta}, q)$ as the objective function, one iteration of the EM algorithm can be reformulated as

E-step: $q^{(t+1)} \in \operatorname*{argmax}_{q} G(\boldsymbol{\theta}^{(t)}, q)$

M-step: $\boldsymbol{\theta}^{(t+1)} \in \operatorname*{argmax}_{\boldsymbol{\theta}} G(\boldsymbol{\theta}, q^{(t)})$

Literature *

1. A. P. Dempster, N. M. Laird and D. B. Rubin. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1. pp. 1-38, 1977.
2. Christopher Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006. Note: Chapter 9.
3. Frank Dellaert. **The Expectation Maximization Algorithm**. TechRep: GIT-GVU-02-20, Georgia Institute of Technology, Atlanta, GA, USA, 2002.
4. Shane M. Haas. **The Expectation-Maximization and Alternating Minimization Algorithms**. Unpublished, 2002.
5. Yihua Chen and Maya R. Gupta. **EM Demystified: An Expectation-Maximization Tutorial**. TechRep: UWEETR-2010-0002, University of Washington, Seattle, WA, USA, 2009.