5.1 Review

This section reviews the introduction to the expectation-maximization (EM)-algorithm as a method to get a maximum likelihood estimate for the parameters of a statistical model with hidden variables. The EM algorithm can be seen as coordinated ascent on the function $\mathcal{L}(Q, \theta)$ and is guaranteed to converge to a local maximum (Note: see also [NH98]).

In the E-step we assign new expected values to the hidden variables

$$Q^{(t+1)} = \arg\max_Q \mathcal{L}(Q, \theta^t)$$

and in the M-step new parameters $\theta$ are chosen to maximize the complete-log likelihood

$$\theta^{(t+1)} = \arg\max_\theta \mathcal{L}(Q^{(t+1)}, \theta)$$

We defined the function $\mathcal{L}(Q, \theta)$ as

$$\mathcal{L}(Q, \theta) = \sum_z Q(z|x) \log \frac{P(x, z|\theta)}{Q(z|x)}$$

where $x$ are observed variables and $z$ are hidden variables. We also gave an upper bound for $\mathcal{L}$

$$\mathcal{L}(Q, \theta) \leq l(\theta) = \log P(x|\theta) = \log P(x_1, \ldots, x_n|\theta)$$

with $x$ standing for the entire observed data set. To maximize the likelihood of the complete data given the $x$ and $\theta$ we assigned the most likely $z$ according to the current parameters

$$\arg\max_Q \mathcal{L}(Q, \theta^t) = P(z|x, \theta^{(t)}) = Q^{(t+1)}$$

and the new expected complete likelihood is

$$\arg\max_\theta \mathcal{L}(Q^{(t+1)}, \theta) = \arg\max_\theta \sum_z P(z|x, \theta^{(t)}) \cdot \frac{\log P(x, z|\theta)}{\text{complete log likelihood}}$$

“marginalizes” away $z$

5.2 EM and directed graphical models (DGM)

In analogy to the observed and hidden variables we partition the nodes of our graphical model into evidence (observed) and hidden nodes:
If each node $x_i$ has a parameter vector $\theta_i$ that is independent of the other parameter vectors (so there is no parameter sharing) and each node $x_i$ only depends on its parents $\pi_{x_i}$, then we can write the probability to observe the vector $x$ as

$$P(x) = \prod_i P(x_i | \pi_{x_i}, \theta_{x_i})$$

Let’s first assume that everything is observed, i.e., if $H=\emptyset$. Then we can use the formula for $P(x)$ to rewrite $l(\theta)$ as a sum:

$$l(\theta) = \log P(x) = \log \prod_i P(x_i | \pi_{x_i}, \theta_{x_i}) = \sum_i \log P(x_i | \pi_{x_i}, \theta_{x_i})$$

This corresponds to a decoupling of the optimization problems for different nodes, i.e. we can optimize the following formulas separately:

$$\arg\max_{\theta_{x_1}} \log P(x_1 | \pi_{x_1})$$

$$\arg\max_{\theta_{x_2}} \log P(x_2 | \pi_{x_2})$$

$$\ldots$$

However, if $H \neq \emptyset$, i.e. we assume that we have observed and hidden variables, then our formula becomes more complicated. First we introduce a simpler notation for our nodes:

$$\mathcal{X} = \mathcal{X}_{\text{observed}} \cup \mathcal{X}_{\text{hidden}}$$

Now we can write the probability to observe the combined vector $(x, z)$ as

$$P(x, z|\theta) = \prod_i P(x_i | x_{\pi_i}, \theta_{x_i}) \prod_i P(z_i | z_{\pi_i}, \theta_{z_i})$$

but since we do not know $z$ we can only compute

$$\arg\max_{\theta} \log P(x|\theta) = \arg\max_{\theta} \log \sum_z P(x, z|\theta)$$

This last formula contains the logarithm of a sum and therefore this problem does not decouple and is hard to optimize. But we know that EM can be used to optimize the following equations (see review section), we substitute $\prod_i P(x_i | x_{\pi_i}, \theta_{x_i}) \prod_i P(z_i | z_{\pi_i}, \theta_{z_i})$ for $P(x, z|\theta)$ and get

$$\arg\max_{\theta} \sum_z P(z | x, \theta^{(t)}) \log P(x, z|\theta)$$
In the last line the terms for the observed variables $x$ and the hidden variables $z$ occur in separate sum formulas. This means observed and the hidden variables are decoupled and we can optimize the sum formulas separately.

Note: probabilistic inference arises wherever we have $P(x_A|x_B, \theta)$ i.e. inference is needed for learning with EM. Therefore, we will be spending considerable time in this course on the mechanisms to provide for efficient inference.

### 5.3 Finite Mixture Models

This section introduces density functions that are linear combinations of a finite set of known and parameterized density functions, i.e. of the general form

$$P(x) = \sum_{i=1}^{M} \alpha_i f_i(x)$$

or, in the parameterized form

$$P(x|\theta) = \sum_{i=1}^{M} \alpha_i f_i(x|\theta_i)$$

This defines a mixture of densities that is a linear combination of $M$ densities. The coefficients are $\alpha_i$ and the parameters for the densities are $\theta_i$. We require the following properties to hold for the coefficients

$$\alpha_i \geq 0; \sum_{i=1}^{M} \alpha_i = 1$$

The properties of the density functions $f_i$ are

$$f_i(x) \geq 0; \int f_i(x) dx = 1$$

If we mix $M$ density functions our parameter vector looks like $\{\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M\}$. Suppose we have observed values (i.e. a sample) from a density function that is linear combination of functions $f_i$. All we know is how many functions are mixed ($M$), but the coefficients $\alpha_i$ and the parameter vectors $\theta_i$ are unknown. But we will see that we can use the EM-algorithm to find the coefficients parameters of the underlying density functions.
The following formula shows how $f_i$ and $\alpha_i$ are related to the graphical models we know. Here $i$ is the component (density function) a value is drawn from, $\alpha_i$ is the probability to draw a value from a particular density $i$ and $f_i(x|\theta_i)$ is the probability to draw $x$ from $i$. Summing over $i$ we get the likelihood for our data ($x$) given the $\alpha_i$ and $\theta_i$.

\[
P(x|\theta) = \sum_i \alpha_i f_i(x|\theta_i) P(i|\theta)
\]

We will also use

\[
P(x, i) = P(x|i) \cdot P(i)
\]

and Bayes rule

\[
p(i|x) = \frac{P(x|i)P(i)}{\sum_j P(x|j)P(j)}
\]

### 5.4 Mixtures of unimodal distributions

This section describes how a mixture of unimodal distributions is used to approximate a multimodal distribution. Figure 2 shows a unimodal distribution. Figure 3 gives an example, how two unimodal distributions (the dashed and the dotted plot) can be used to approximate a bimodal distribution.
Figure 5.3: A mixture of two unimodals (e.g. $p(x)$). The unimodal distributions are dashed and dotted, here the same $\alpha$ and $\theta$ is used for both distributions.

One example for an unimodal distribution is the Gaussian distribution

$$f_i(x|\theta) = |2\pi \Sigma|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

but any other unimodal distribution can be used.

Figure 4a and 4b show the data model we assume here. The $x_n$ is the observed data from the distribution, it can be discrete or continuous, the $y_n$ corresponds to the chosen unimodal distribution and is hidden and discrete.

Figure 5.4a: The graphical model for sampling.
To sample from the mixture model we choose one of the (hidden) density functions according to the distribution over those functions. Then we sample the value for \( x \) from the density function we have chosen. For an example recall figure 2: first we choose one of the two Gaussians (here with probability 0.5) and then we generate a value using the density function for this Gaussian.

Learning in the case of complete data is easy, given \( x \) and \( y \) are known we have:

\[
\log P(x, y | \theta) = \log \prod_i P(x_i | y_i, \theta_i) \cdot P(y_i | \theta_i) \\
= \sum_i \log P(x_i | y_i, \theta_i) + \sum_i \log P(y_i | \theta_i) \\
= \sum_{l=1}^{M} \sum_{i: y_i = l} \log P(x_i | l, \theta_l) + \sum_i \log \alpha_{y_i}
\]

The first term of the last equation corresponds to learning the parameters for a single Gaussian (\( l \)) from a sample set \( x_i \), i.e. from all values that “were generated” by this Gaussian. We can easily find these parameters using the sample means and covariances. The second part coressponds to learning the discrete mixture parameters and just amounts to bin counting (or constructing a histogram).

For Gaussian components, we just chose the best fitting parameters for each of the \( M \) Gaussians as follows

\[
\mu_l = \frac{1}{|i : y_i = l|} \sum_{i : y_i = l} x_i \\
\Sigma_l = \frac{1}{|i : y_i = l|} \sum_{i : y_i = l} (x_i - \mu_l)(x_i - \mu_l)
\]

and

\[
\alpha_l = |i : y_i = l|/N
\]

where \( N \) is the number of samples. Note: we just calculated mean and covariance from our dataset and we assume that this data is normal.

If \( y \) happens to not observed, things are much more difficult.
\[ l(\theta) = \log \prod_{j=1}^{N} P(x_i | \theta) = \sum_{i=1}^{N} \log \left( \sum_{l=1}^{M} a_l P_l(x_i | \theta_l) \right) \]

This is hard to optimize, because we have the logarithm of a sum. But if we have initial parameter estimates \( \Theta^0 = (\alpha_{1..M}^0, \theta_{1..M}^0) \) and since the samples are i.i.d. we have

\[ P(y|x, \Theta^0) = \prod_{i=1}^{N} P(y_i | x_i, \Theta^0) \]

and

\[ P(y_i | x_i, \Theta^0) = \frac{P(x_i | y_i, \Theta^0) P(y_i | \Theta^0)}{\sum_k P(x_i | k, \Theta^0) P(k, \Theta^0)} \]

The expected complete log likelihood is:

\[ E \left[ \log P(x, y | \theta^{(l)}) \right] = \sum_{y} \log P(x, y | \theta) P(y | x, \Theta^0) \]

\[ = \sum_{y} \sum_{i=1}^{N} \log(\alpha_{y_i} P(x_i | y_i, \Theta_{y_i})) P(y | x, \Theta^0) \]

\[ = \sum_{i=1}^{N} \sum_{y_i} \log(\alpha_{y_i} P(x_i | y_i \Theta_{y_i})) P(y_i | x_i, \Theta^0) \cdot \sum_{y \setminus y_i} P(y \setminus y_i | x, \Theta^0) \]

\[ = \sum_{i=1}^{N} \sum_{l=1}^{M} \log(\alpha_l P(x_i | l, \theta_l)) P(l | x_i, \Theta^0) \]

\[ = \sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) P(l | x_i, \Theta^0) + \sum_{i=1}^{M} \sum_{l=1}^{N} \log(P_l(x_i | \theta_l)) P(l | x_i, \Theta^0) \]

Here \( \alpha_l \) and \( \theta_l \) are decoupled again, i.e. we can optimize the terms that contain \( \alpha_l \) and \( \theta_l \) separately.

We introduce a Lagrange multiplier \( \lambda \) on the constraint \( \sum_l \alpha_l = 1 \) to solve the following optimization problem

\[ \frac{\partial}{\partial \alpha_l} \left[ \sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) P(l | x_i, \Theta^0) + \lambda \left( \sum_{l=1}^{M} \alpha_l - 1 \right) \right] = 0 \]

or

\[ \sum_{i=1}^{N} \frac{1}{\alpha_l} P(l | x_i, \Theta^0) + \lambda = 0 \]

If we sum over \( l \) we have \( \lambda = -N \) and we get:

\[ \alpha_l = \frac{1}{N} \sum_{i=1}^{N} P(l | x_i, \Theta^0) \]

If we are using Gaussian components, we need some results from matrix algebra first.
5.4.1 Matrix Algebra Summary

Here we recall some results from matrix algebra:

\[ tr(A) = \sum_{i,j} A_{ij} \] (The trace of a matrix A is the sum of the A's diagonal elements.)

\[ tr(b) = b \] (The trace of a scalar is the scalar itself.)

\[ tr(A + B) = tr(A) + tr(B) \]

\[ tr(AB) = tr(BA) \] When well defined

\[ tr(ABC) = tr(CAB) = tr(BCA) \]

\[ \sum_{i} x_i^T A x_i = tr(\sum_{i} x_i^T A x_i) = tr(\sum_{i} A x_i x_i^T) = tr(A \sum_{i} x_i x_i^T) = tr(AB) \] where \( B = \sum_{i} x_i x_i^T \)

\[ |A| = det(A) \text{ and } |A^{-1}| = \frac{1}{|A|} \]

We use the following properties (for A, B symmetric matrices)

\[ \frac{\partial x^T A x}{\partial x} = (A + A^T)x \]

\[ \frac{\partial \log A}{\partial A} = (2A^{-1} - \text{diag}(A^{-1})) \]

\[ \frac{\partial tr(AB)}{\partial A} = (2B - \text{diag}(B)) \]

and formulas [B98]:

\[ \mu_i = \frac{\sum_{j=1}^{N} x_j p(l|x_i, \Theta^\theta)}{\sum_{j=1}^{N} p(l|x_i, \Theta^\theta)} \]

We first write out the likelihood

\[ \sum_{i} \left[ \frac{1}{2} \log(|\Sigma_i^{-1}|) \sum_{i=1}^{N} p(l|x_i, \Theta^\theta) - \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)tr(\Sigma_i^{-1}(x_i - \mu_i)(x_i - \mu_i^T)) \right] \]

and then introduce \( N_{i,i} \)

\[ \sum_{i} \left[ \frac{1}{2} \log(|\Sigma_i^{-1}|) \sum_{i=1}^{N} p(l|x_i, \Theta^\theta) - \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)tr(\Sigma_i^{-1}N_{i,i}) \right] \]

as

\[ N_{i,i} = (x_i - \mu_i)(x_i - \mu_i^T) \]
Taking the partial derivative of the last formula with respect to $\Sigma_i^{-1}$ we get:

$$\frac{\partial}{\partial \Sigma_i^{-1}} = \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)(2\Sigma_i - \text{diag}(\Sigma_i)) - \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)(2N_i - \text{diag}(N_i, i))$$

$$= \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)(2M_i, i - \text{diag}(M_i, i))$$

$$= 2S - \text{diag}(S)$$

where $M_i, i = \Sigma_i - N_i, i = \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)M_i, i$ and $S = \frac{1}{2} \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)M_i, i$. Furthermore

$$0 = 2S - \text{diag}(S) \Rightarrow S = 0 \Rightarrow \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)(\Sigma_i - N_i, i) = 0$$

For the last step we just substituted for $S$ and $M_i, i$, if we solve this for $\Sigma_i$ and substitute for $N_i, i$ we get

$$\Sigma_i, i = \sum_{i=1}^{N} p(l|x_i, \Theta^\theta)(x_i - \mu_i)(x_i - \mu_i)^T$$

$$\sum_{i=1}^{N} p(l|x_i, \Theta^\theta)$$

### 5.5 Multivariate Gaussian

Multivariate Gaussians can themselves be seen as graphical models. In the general case, all nodes are connected to all other nodes in the figure, but as we will see, depending on the values of the covariance and inverse covariance matrix, we will have independence and conditional independence properties within the Gaussian.

![Graphical Model](image)

Figure 5.5: The most general graphical model for a multivariate Gaussian

where

$$P(x|\mu, \Sigma) = |2\pi \Sigma|^{-1/2} \exp{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

which we indicate by saying that $x \sim N(\mu, \Sigma)$. The mean is $\mu = EX$ and covariance $\Sigma = E[(x - \mu)(x - \mu)^T]$.

Here is a formal definition of a multivariate normal distribution.

**Definition 5.1.** $x$ has a $p$-variate normal distribution, iff $a^T x$ is univariate normal for all fixed $a$ vectors.

$$a^T x \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{(a - \mu)^2}{\sigma^2}\right)}$$
Note: in the function

\[ f(x) = \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \]

the term

\[ (x - \mu)^T \Sigma^{-1} (x - \mu) \]

is a generalized form of the Euclidean distance, sometimes called the 'Mahalanobis' distance between \( x \) and \( \mu \). Setting the function to a constant, e.g. \( f(x) = c \) leads to ellipsoids with center \( \mu \), and with the principle axes pointing in the direction of the eigenvectors (Figure 6) of the covariance matrix.

![Figure 5.6: Ellipsoids with center \( \mu \) and axes in the direction of the eigenvectors.](image)

Here are some other properties (given without proof):

1. \( x \sim N(\mu, \Sigma) \) and \( y = Ax + c \Rightarrow y \sim N(A\mu + c, A\Sigma A^T) \)
2. \( x \sim N(\mu, \Sigma) \), \( \Sigma \) is positive definite, and \( y = \Sigma^{-1/2}(x - \mu) \) with \( \Sigma^{-1/2} \) defined so that \( \Sigma^{1/2}\Sigma^{1/2} = \Sigma \), \( \Rightarrow y \sim N(0, I) \) with \( I \) the identity matrix.
3. \( x_1 \sim N(\mu_1, \Sigma_1), x_2 = N(\mu_2, \Sigma_2), x_1 \perp x_2 \)
   \( \Rightarrow x_1 + x_2 \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2) \) (note that this need not hold in general. In general, if \( X \) and \( Y \) are two r.v.'s \( \sim f(X,Y) \), and we form \( Z = X + Y \), then \( f(z) = \int f(x, z-x)dx \).
4. These properties and many others can be found in most multivariate statistics books.

We are interested in the relationship between Gaussians and graphical models with respect to independence and conditional independence properties. The next portion of the lecture leads to an analysis so that we can find when variables are independent and conditionally independent.

We partition the \( n \times 1 \) vector \( x \) into a \( p \times 1 \) vector \( x_1 \) and a \( q \times 1 \) vector \( x_2 \), such that \( n = p + q \). Further we partition \( \mu \) and \( \Sigma \) in a similar way:

\[
\begin{align*}
x &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, & \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, & \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\end{align*}
\]

Where \( \Sigma_{11} \) is \( p \times p \), \( \Sigma_{12} \) is \( p \times q \), ...
We can write:

\[
p(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\}
\]

**Definition 5.2.** \( K \triangleq \Sigma^{-1} \) is the concentration matrix, \( K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \)

Further, to make sure there is no ambiguity about the order that we take an inverse or take a submatrix, we define: \( \Sigma^{-1}_{11} \triangleq (\Sigma_{11})^{-1} \) and \( K_{11} \triangleq (\Sigma^{-1})_{11} \).

**Theorem 5.3.** Let \( x \) be partitioned as above. Then \( x_1 \perp x_2 \) iff \( \Sigma_{12}^T = 0 \).

Note: \( \Sigma \) is a symmetric matrix.

**Proof.**

\[
|\Sigma| = \begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}| |\Sigma_{22}|
\]

\[
P(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma_{11}|^{1/2} |\Sigma_{22}|^{1/2}} \exp\left\{ -\frac{1}{2} \left[ (x_1 - \mu_1)^T \Sigma_{11} (x_1 - \mu_1) + (x_2 - \mu_2)^T \Sigma_{22} (x_2 - \mu_2) \right] \right\}
\]

\[
P(x | \mu, \Sigma) = \exp\left( -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_{11} (x_1 - \mu_1) \right) \cdot \frac{1}{(2\pi)^{d/2} |\Sigma_{22}|^{1/2}} \exp\left( -\frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22} (x_2 - \mu_2) \right)
\]

\[
P(x | \mu, \Sigma) = P(x_1 | \mu_1, \Sigma_1) \cdot P(x_2 | \mu_2, \Sigma_2)
\]

Note: recall how we wrote \( P(x | \mu, \Sigma) \) before.

Note: the last line corresponds to \( x_1 \perp x_2 \), zeros in the covariance matrix correspond to marginal independence.

\[\square\]

**References**


