Announcements:

The final project should be done individually and should be highly related to graphical models. It doesn’t have to be pattern recognition. The abstract is due next Friday in class. The progress report will be based on this and so about the final paper. One should work as if the final report will be a publishable paper.

8.1 Review

In the last lecture, we formulated the factorization of multi-variate Gaussian $N(\mu, \Sigma)$ in the following way

$$ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T $$

$$ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} $$

Therefore the probability density function (pdf) can be written as

$$ P(x) = P(x_1, x_2) = P(x_1 | x_2) P(x_2), $$

where

$$ X_2 \sim N(\mu_2, \Sigma_{22}) $$

$$ X_1 | X_2 \sim N(\mu_{1|2}, \Sigma_{1|2}) $$

$$ = N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) $$

are all Gaussian distributions. We also formulated the concentration matrix $K$ as the inverse of the covariance matrix. In the multi-variant Gaussian case, it is useful to partition covariance matrix and concentration matrix into blocks. It is also useful to formulate the relationships between the blocks in covariance matrix and concentration matrix. To do this, we need the Schur’s formula which creates holes in a blocked matrix.

Lemma 8.1. (Schur’s formula) Suppose matrix $A$ invertible, $A \in GR$, then

1. $$ \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & B \\ 0 & D - CA^{-1}B \end{pmatrix}. $$

2. $$ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & D - CA^{-1}B \end{pmatrix}. $$

3. $$ \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix}. $$
By Schur’s formula, we have
\[ K = \Sigma^{-1} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \]
\[ = \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & - (\Sigma/\Sigma_{22})^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1} \Sigma_{21} (\Sigma/\Sigma_{22})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1} \Sigma_{21} (\Sigma/\Sigma_{22})^{-1} \Sigma_{12} \Sigma_{22}^{-1} \end{pmatrix}. \]

From the result, we can see that
\[ K_{11} = (\Sigma/\Sigma_{22})^{-1} = \Sigma_{1|2}^{-1}, \]
which means that inversion of the concentration matrix gives the conditional covariance matrix. On the other hand, since
\[ K_{12} = -(\Sigma/\Sigma_{22})^{-1} \Sigma_{12} \Sigma_{22}^{-1} = -K_{11} \Sigma_{12} \Sigma_{22}^{-1} \Rightarrow K_{11} K_{12} = -\Sigma_{12} \Sigma_{22}^{-1}, \]
we can rewrite the conditional mean as
\[ \mu_{1|2} = \mu_1 - K_{11} K_{12} (x_2 - \mu_2). \]

In conclusion, the concentration matrix gives the full information about the factorization of a multi-variate Gaussian.

\[ \left\{ \begin{array}{lcl} \mu_{1|2} &=& \mu_1 - K_{11} K_{12} (x_2 - \mu_2) \\
\Sigma_{1|2} &=& K_{11}^{-1} = \Sigma/\Sigma_{22} \\
&=& \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}. \end{array} \right. \] (8.1)

### 8.2 Subpartition of Multi-variate Gaussians

We have examined the conditional distribution in the multi-variate Gaussian case. What about the conditional independent (CI) statements in mixture models? What’s the relationship between conditional independence and the properties of the covariance/concentration matrices and how are they related to graphical models. In order to answer these questions, let’s partition \( x_1 \) furthermore as \( x_1 = (x_{11} \ x_{12})^T \). Therefore we have
\[ x = (x_{11} \ x_{12} \ x_2)^T, \]
\[ K_{11} = \begin{pmatrix} K_{11,11} & K_{11,12} \\ K_{11,21} & K_{11,22} \end{pmatrix}, \]
\[ \mu_{1|2} = \begin{pmatrix} \mu_{1|2,1} \\ \mu_{1|2,2} \end{pmatrix}, \]
\[ \Sigma_{1|2} = \begin{pmatrix} \Sigma_{1|2,11} & \Sigma_{1|2,12} \\ \Sigma_{1|2,21} & \Sigma_{1|2,22} \end{pmatrix}. \]

With this subpartition, we have the following statement:

**Lemma 8.2.** The following three conditions are identical:
\[ X_{1,1} \perp \perp X_{1,2} | X_2 \Leftrightarrow \Sigma_{1|2,12} = 0 \Leftrightarrow K_{11,12} = 0, \]
\[ \Rightarrow \Sigma_{1|2,21} = 0 \Rightarrow K_{11,21} = 0. \]

**Proof.** The first equivalence statement is easy to see above. The second is because for a block diagonized matrix, we have
\[ \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix}. \]

\[ \square \]
The reason we split random variables is we want to analyze the conditional or marginal independency. We already know that zeros in the covariance matrix gives marginal independency. The lemma shows that analyzing conditional independency is equivalent to finding zeros in the concentration matrix. More formally, we can prove the following theorem for conditional independent (CI) statements in multivariate gaussian partitioning.

**Theorem 8.3.** Suppose multidimension random variable $X$ obeys Gaussian distribution

$$X \sim N(\mu, \Sigma),$$

where $\Sigma > 0$ is a definite positive matrix. We partition $X$ in the following manner:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix},$$

$$K = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix},$$

then we have

$$X_1 \perp\!\!\!\!\!\!\perp X_2 | X_3 \text{ iff } K_{12} = K_{21}^T = 0.$$

The theorem says zeros in the concentration matrix gives conditional indepence. Another claim is zeros in the covariance matrix gives marginal independence. One should keep in mind the difference of these two statements.

### 8.3 Multi-variate Gaussian as Undirected Graphical Model

Now let’s consider how multi-variate Gaussians can be related to graphical models. In the undirected graphical model (UDM) case, each dimension coresponds to a node in the graph. In the general case, the graph is fully connected. Now let’s consider the relationship between the missing edge in UGM (which implies some CI statements) and the properties of the concentration matrix.

**Example 8.1** As an example, we look at a UGM which coresponds to a five dimensional Gaussian distribution.

![Figure 8.1: Fully dependent 5-d Gaussian model](image)

Right now, since the full connectivity of the graph, we can’t induce any CI statements. If we remove one of the edge, say between $X_3$ and $X_4$, as shown in Figure 8.2, we get the following CI statement

$$X_3 \perp\!\!\!\!\!\!\perp X_4 | \{X_1, X_2, X_5\}.$$

![Figure 8.2: 5-d Gaussian model with one missing edge](image)
One the other hand, in the concentration matrix, we have $K_{34} = K_{43}^T = 0$. Furthermore, if have another missing edge between $X_1$ and $X_3$ as in shown Figure 8.3, we can have the following CI statements:

$$X_1 \perp\!\!\!\!\!\!\perp X_3 | \{X_2, X_5\}$$
$$X_3 \perp\!\!\!\!\!\!\perp X_4 | \{X_2, X_5\}.$$

At the same time, we can find the concentration satisfies $K_{13} = K_{31}^T = 0$ and $K_{34} = K_{43}^T = 0$.

By examing the above example, we got the following statement:

**Claim 8.4.** *In multi-variate Gaussians, zeros in the concentration matrix corespond to the missing edge in the UGM.*

So far, we have several equivalent statements: conditional indepence, zeros in the concentration matrix, and missing edge in the UGM. When analyzing a Gaussian probablistic system, it is free choose whichever is convenient. In conclusion, for a multi-variance Gaussian,

- marginal independency $\equiv$ zeros in covariance matrix.
- conditional independency $\equiv$ zeros in concentration matrix *equiv* missing edge in the UGM.

**Example 8.2** Now let’s look at another UGM which coresponds to a four dimension Gaussian distribution as in Figure 8.4.

As we already know, this is an example which has no coresponding directed graphical to it. The coresponding concentration matrix has the form of

$$K = \begin{pmatrix}
* & * & * & 0 \\
* & * & 0 & * \\
* & 0 & * & * \\
0 & * & * & *
\end{pmatrix}.$$

### 8.4 Multi-variate Gaussian as Directed Graphical Model

We have already annalized how Gaussians can be viewed as undirected graphical models. As we mentioned in previous lectures, Gaussian models can be both under directed and undirected graphical models in the class relationship diagrams. In this section, we will try to see how multi-variate Gaussians can be related to directed graphical models (DGMs). A DGM corresponding to Gaussian with no CI assumptions has the following kind of graphical models:
It is not a fully connected graph since a GM must be a DAG. But if we neglect the arrows, it becomes a fully connected UGM.

Now let’s factorize the Gaussian and see how Gaussians can be regarded as DGMs. In an $n$-d Gaussian distribution, we can factorize the pdf as

$$p(x_1:n) = \prod_i p(x_i|x_{\pi_i}).$$

In each term of the factor, the conditional probability can be written as

$$p(x_j|x_{\pi_j}) = N(\sum_{i\in \pi_j} w_{ij} x_j + b_j, v_j),$$

which shows that $x_i$ is a scalar, and individually Gaussian. The relationship can also be seen from Figure 8.6.

The weight matrix $W = (w_{ij})$ is constructed such that $i$ is the parent and $j$ is the child. We will see later the $W$ matrix is exactly the $B$ matrix in the modified Cholesky decomposition. Now we reverse the order of dependency and rewrite the factorization as

$$p(x_{1:n}) = \prod_i p(x_i|x_{i+1:n}),$$

where each factor is a Gaussian and

$$p(x_i|x_{i+1:n}) = N(\mu_{i|i+1:n}, \Sigma_{i|i+1:n}),$$

$$\mu_{i|i+1:n} = \mu_i + \Sigma_{i,i+1:n} \Sigma_{i+1:n,i+1:n}^{-1} (x_{i+1:n} - \mu_{i+1:n}),$$

$$\Sigma_{i|i+1:n} = \Sigma_{ii} + \Sigma_{i,i+1:n} \Sigma_{i+1:n,i+1:n}^{-1} \Sigma_{i+1:n,i}.$$ 

Keep in mind that the conditional mean and variance are both scalars, the overall pdf can be written as

$$p(x_{1:n}) = \prod_i (2\pi \Sigma_{i|i+1:n})^{-1/2} \exp \left[ -\frac{1}{2} (x_i - \mu_{i|i+1:n})^2 \Sigma_{i|i+1:n}^{-1} \right]. \quad (8.2)$$

Because the concentration matrix is positive definite ($K > 0$), we can use the modified Cholesky decomposition and factor the matrix as

$$K = U^T DU,$$
where $U$ is an upper triangle matrix with diagonal ones, and $D$ is a diagonal matrix:

$$U = \begin{pmatrix} 1 & * & \cdots & * \\ 0 & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

$$D = \begin{pmatrix} D_{00} & 0 & \cdots & 0 \\ 0 & D_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{dd} \end{pmatrix} \Rightarrow |K| = |D|.$$

With this factorization, we can write the pdf as

$$p(x_1:n) = (2\pi)^{-d/2}|D|^{1/2} \exp \left[ -\frac{1}{2}(x - \mu)^T U^T D U (x - \mu) \right].$$

Furthermore, let’s focus on the exponential index term.

$$U = I - B \Rightarrow (U(x - \mu))^T D (U(x - \mu)) = ((I - B)(x - \mu))^T D ((I - B)(x - \mu))$$

$$\tilde{\mu} = (I - B) \mu = (x - Bx - \tilde{\mu})^T D(x - Bx - \tilde{\mu}).$$

Here, the $B$ matrix is an upper triangle matrix with zeros in the diagonal:

$$B = \begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$ 

This matrix fully describes the relationship between components in a DGM and is the $W$ matrix we mentioned before.

The overall pdf is

$$p(x_1:n) = (2\pi)^{-d/2}|D|^{1/2} \exp \left[ -\frac{1}{2}(x - Bx - \tilde{\mu})^T D(x - Bx - \tilde{\mu}) \right]$$

$$= \prod_i (2\pi)^{-1/2} D_{ii}^{1/2} \exp \left[ -\frac{1}{2} (x_i - B_{i,i+1:n} x_{i+1:n} - \tilde{\mu}_i)^2 D_{ii} \right], \quad (8.3)$$

where $D_{ii}$ is the $i$th diagonal term of $D$, and $x_i$ is the $i$th component of vector $x$.

Compare the result with Equation 8.2, we got

$$B_{i,i+1:n} = \Sigma_{i,i+1:n} \Sigma_{i+1:n,i+1:n}^{-1}$$

$$\tilde{\mu}_i = \mu_i - B_{i,i+1:n} \mu_{i+1:n}$$

$$D_{ii} = \Sigma_{i,i+1:n}^{-1} \mu_i$$

**Claim 8.5.** In multi-variate Gaussians, zeros in the upper triangle part of B matrix corresponds to the missing edge in the DGM.

$$X_i \perp \perp X_{\{i+1:n\} \setminus \pi_i} | X_{\pi_i}, \text{iff } B_{i,\{i+1:n\} \setminus \pi_i} = 0.$$
**Example 8.3** Consider a Gaussian model associated with the following DGM:

\[X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5\]

Figure 8.7: 5-d Gaussian model as DGM

Then the corresponding \(B\) matrix has the format

\[
B = \begin{pmatrix}
0 & * & * & 0 & 0 \\
0 & 0 & * & 0 & 0 \\
0 & 0 & 0 & * & * \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

So far, we look at how multi-variate Gaussian can be viewed as undirected and directed graphical models. We should keep in mind that zeros in \(B\) matrix does not imply zeros in \(K\) matrix. The two canonical graphs below can be good examples.

\[X_1 \rightarrow X_2\]
\[X_3 \rightarrow X_4\]
\[X_1 \rightarrow X_2 \rightarrow X_3\]

Figure 8.8: Two canonical examples

In the first case, the \(K\) matrix has zeros in the entries \((1, 4), (2, 3), (3, 2), (4, 1)\). However, since there is no corresponding DGM to this model, the \(B\) matrix doesn’t have zeros in the upper triangle. Similarly, in the second case, the \(B\) matrix has zeros in the entries \((1, 2)\). But since there is no corresponding UGM to it, the \(K\) matrix is full.

### 8.5 Hidden Markov Models

Mixture model can be viewed as the following graphical model.

\[Y_n \rightarrow X_n \rightarrow Y_{n-1} \cdots \]

Figure 8.9: Graphical model for mixture models
where \( y_i \in \{1, 2, \ldots, M\} \) is discrete and \( x_i \) can be either discrete or continuous. By analyzing the graphical model above we get the following CI statements.

\[
X_i \perp \perp X_j, Y_i \perp \perp Y_j, \forall i \neq j.
\]

Therefore if \( Y_i \) is iid, \( X_i \) is also iid.

In this model, the rv \( Y_i \)'s are independent of each other. But suppose the condition \( P(y_i|y_{i-1}) = P(y_i) \) no longer holds. Instead, we relax the condition and have the fact that \( Y_i \) only depends on the previous value \( y_{i-1} \):

\[
P(y_i|y_{1:i-1}) = P(y_i|y_{i-1}).
\]

For convenience, let’s also change the notation so that \( Q_t = Y_i \) and assume the index variable \( t \) represents time. The transition matrix is

\[
P(q_t|q_{t-1}) = a_{q_{t-1}, q_t},
\]

or

\[
P(Q_t = i|Q_{t-1} = j) = a_{ji}.
\]

Now we get the hidden Markov model as shown below.

The name hidden Markov model (HMM) comes from the fact that the rv \( Q_t \) is hidden (not observable) and \( X_t \) is observed (evidence). But this doesn’t have to be true.

Properties

Given a hidden markov model described as above, we have the following properties.

1. The rv \( Q_t \) forms a Markov chain \( Q_1 \to Q_2 \to \cdots \to Q_T \).
2. We don’t have the independence statement \( X_i \perp \perp X_j, \forall i \neq j \).
3. We have the following CI statements:
   
   (a) \( Q_t \perp \perp \{Q_{1:t-2}, X_{1:t-1}\}|Q_{t-1} \),
   
   (b) \( X_t \perp \perp \{Q_{1:t-1}, X_{1:t-1}, Q_{t+1:T}, X_{t+1:T}\}|Q_t \).

   We should note that the first CI statement we are not saying that \( Q_t \) is independent of the future given the past.
4. \( P(X_T|X_1) \neq P(X_T) \) and \( P(X_t|q_1) \neq P(X_T) \) even if \( T \gg 1 \).
5. \( P(Q_{1:T}) \neq \prod_{t} P(Q_t), P(Q_{1:T}|X_{1:T}) \neq \prod_{t} P(Q_t|X_{1:T}) \).
One should keep in mind that the meaning of probabilities is actually

\[ P(q_t|q_{t-1}) = P(Q_t = q_t|Q_{t-1} = q_{t-1}) = P_{Q_t|Q_{t-1}}(Q_t = q_t|Q_{t-1} = q_{t-1}), \]

which says the distribution may depend on time. But if we have time homogeneity, we denote

\[ P(Q_t = i|Q_{t-1} = j) = a_{ij}, \]

and the initial condition

\[ P(q_0) = \pi, \pi_i = P(Q_0 = i). \]

Therefore, the overall probability can be factored as

\[ P(q_0:T, x_0:T) = \prod_i P(x_t|q_t)P(q_t|q_{t-1}), \]

where \( P(q_0|q_{-1}) = P(q_0) \), and the marginal distribution is

\[ P(x_0:T) = \sum_{q_0:T} P(x_0:T, q_0:T) \]

\[ = \sum_{q_0=1}^{|Q|} \sum_{q_1=1}^{|Q|} \cdots \sum_{q_T=1}^{|Q|} P(x_0:T, q_0, q_1, \ldots, q_T). \]

In the naïve algorithm, it requires \(|Q|^T\) computation to get the marginals. Usually the length \( T \) is very big so the total number of operations is huge. We will see how CI statements can help us reduce the computation in the next lecture.

References
