01.001 - Introduction to Probability and Statistics

Alexander Binder

Week 11: First lecture on 03-Apr-2017, Second Lecture on 05-Apr-2017

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity. every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

Summary of key ideas from week's second lecture

The multivariate normal distribution

The Multivariate Normal Density

The one-dimensional normal distribution with parameters μ, σ^2 generates real numbers $x \in \mathbb{R}^1$ according to the density

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

The multi-variate normal generates **vectors** of real numbers $x \in \mathbb{R}^n$ according to the density

$$f(x) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

It has as parameters

1. mean μ - a vector $\mu \in \mathbb{R}^{n\times 1}$
2. Covariance matrix Σ - a matrix $\Sigma \in \mathbb{R}^{n\times n}$. This matrix must satisfy two properties
   (a) Σ is symmetric
   (b) Σ is positive definite.

$det(\Sigma)$ is the Determinant of a matrix which is equal to the product of all its eigenvalues.

Normal Distribution: Independence and the Covariance matrix

For a joint normal distribution, the covariance matrix allows to see independence relationships. However this holds only for this distribution.

What for is that useful? We can design Σ to encode independences that we want to have in our model – by enforcing (block-wise) zeros in the right positions!

Theorem:

if all X_k are independent from each other, then the covariance matrix is a diagonal matrix, that is

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

Remember that this is one case that we plotted above.

Proof:

Theorem:

If

1. the set of random variables are **jointly normally distributed**

2. and two blocks of variables X_{set1},X_{set2} are independent, then we have for all pairs of variables (X_{i1}, X_{k2}) such that X_{i1} is from the first set, and X_{k2} from the second set the following observation in the covariance matrix: $\Sigma_{i1,k2} = \Sigma_{k2,i1} = 0$ - that is the entry in $i1$-th row and $k2$-th column is zero, and also its mirrored entry in $i1$-th row and $k2$-th column.

If

1. the set of random variables are **jointly normally distributed**

2. the covariance matrix has zeros, then the corresponding sets are normally distributed

The theorem is best demonstrated by an **Example:**

Suppose $X_{set1} = (X_1, X_2, X_3)$, $X_{set2} = (X_4, X_5)$.

Then Σ must look like this ($*$ denotes a non-zero entry)

$$\Sigma = \begin{pmatrix} * & * & * & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}$$

That means: $P(X_1, X_2, X_3, X_4, X_5) = P(X_1, X_2, X_3)P(X_4, X_5)$.

Similarly:

$$\Sigma = \begin{pmatrix} * & * & 0 & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

denotes that X_1 and X_3 are independent. This is not of use for the joint distribution $P(X_1, X_2, X_3, X_4, X_5)$, but we know that for the marginal distribution of X_1 and X_3 it holds that: $P(X_1, X_3) = P(X_1)P(X_3)$.

How to see this? Delete from the matrix all those rows and columns that have no 0s, that is: $2, 4, 5$. then we get a 2×2 matrix over variables X_1, X_3 only that looks like:

$$\Sigma = \begin{pmatrix} [X_1] & * & 0 \\ [X_3] & 0 & * \end{pmatrix}$$

That means that in the marginal for (X_1, X_3) we have $P(X_1, X_3) = P(X_1)P(X_3)$.

Same thinking:

$$\Sigma = \begin{pmatrix} * & * & 0 & 0 & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

How to see this ... lets delete rows/columns $2, 5$ (because they have no 0 entries). then we obtain this matrix over X_1, X_3, X_4

$$\Sigma = \begin{pmatrix} [X_1] & * & 0 & 0 \\ [X_3] & 0 & * & * \\ [X_4] & 0 & * & * \end{pmatrix}$$

From here we can see that the block $X_{set1} = (X_1)$ is independent from the block $X_{set2} = (X_3, X_4)$. That means we can say nothing for the joint $P(X_1, X_2, X_3, X_4, X_5)$ but we have for the marginal over variables X_1, X_3, X_4: $P(X_1, X_3, X_4) = P(X_1)P(X_3, X_4)$. In fact this implies more: you can integrate out any variable in this equation to obtain independencies for lower-dimensional marginal distributions.

$$P(X_1, X_3, X_4) = P(X_1)P(X_3, X_4)$$

implies also:

$$P(X_1, X_3) = P(X_1)P(X_3)$$

$$P(X_1, X_4) = P(X_1)P(X_4)$$

This shows that the covariance matrix Σ is – for the normal distribution – linked to independence properties between variables of its marginal distributions.

Warning!!!:

If the set of random variables are **jointly normally distributed**, and the covariance has zeros, then this implies **independence**. But if only marginal distributions are normally distributed, and the variables have zero covariance, then this does NOT imply independence. The deeper reason is: if variables are marginally normal distributed, it does NOT mean that their joint is a normal distribution.

A counterexample can be given by:

$$X \sim N(0,1)$$
$$Y = uX,\ u \in \{-1,+1\},\ P(u=-1) = 0.5$$

This stuff is fascinating: $P(X+Y=0) = P(X+uX=0) = 0.5$. Does this have a density?

Normal Distribution and Linear transformations

How does the normal distribution changes when $x \sim N(\mu,\Sigma)$ and we transform the random vector $x \in \mathbb{R}^n$ by a linear transformation: $\tilde{x} = Ax+b$ where $A \in \mathbb{R}^{n\times n}$ is an $n \times n$-Matrix, and $b \in \mathbb{R}^n$ is a vector.

Property:

Suppose $x \sim N(\mu,\Sigma)$, $A \in \mathbb{R}^{n\times n}$ is an $n \times n$-Matrix, and $b \in \mathbb{R}^n$ is a vector. We have then

$$\begin{aligned} x &\sim N(\mu,\Sigma) \\ \Rightarrow y = Ax+b &\sim N(A\mu+b, A\Sigma A^T) \end{aligned}$$

See worked out example below for how to use this for sampling vectors from a normal distribution with parameters μ,Σ when one can draw only real numbers from the one-dimensional normal distribution $N(0,1)$

Maximum Likelihood Estimator for the multivariate normal distribution

Its important. See the worked out examples.

We obtain

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

Note here: $x_i - \mu$ is a $n \times 1$ vector, $(x_i - \mu)^T$ is a $1 \times n$ vector. The matrix product of $n \times 1$ with $1 \times n$ is a $n \times n$-matrix. So $\widehat{\Sigma} = \ldots$ is an equation between two matrices. For a single component of Σ this decomposes into:

$$\widehat{\Sigma}_{k,l} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,k} - \mu_k)(x_{i,l} - \mu_l)$$

where μ_k is the k-th dimension of vector μ, and $x_{i,k}$ is the k-th dimension of vector x_i.

Compare this to the one-dimensional case, where we had $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - just that in that case we have real numbers and not vectors, and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$.

Worked-out examples related to week's second lecture

1. ...

   Suppose you have a random number generator, that can generate one-dimensional random numbers $x \sim N(0,1)$. You want to draw from a multivariate random number distribution $N(\mu, \Sigma)$ in say $d = 369$ dimensions. How to do that?

Solution:

We know: If we draw $x_i \sim N(0,1) = f(x_i \mid 0,1)$, then the vector

$$\begin{aligned} x = (x_1, \ldots, x_n) &\sim \prod_{i=1}^{n} f(x_i \mid 0,1) \\ &= N(0, I) \end{aligned}$$

where I is the $d \times d$ identity matrix.
We know from the theorem about the normal distribution and linear transformations that:

$$x \sim N(0,1) \Rightarrow Ax + b \sim N(b, AIA^T) = N(b, AA^T)$$

We need

$$b = \mu,\ \Sigma = AA^T$$

So we must choose

(a) $b = \mu$

(b) A such that $\Sigma = AA^T$.
Usually, you cannot do that by pen and paper except when Σ is diagonal. To our rescue `scipy.linalg` (and weka for Java, GNU GSL, eigen3 for C/C++ and many other libraries) has matrix decomposition functions that do right this kind of decomposition.
The solution A is not unique. you can multiply it with any rotation matrix R ($RR^T = I$), then AR is also a solution because of

$$AR(AR)^T = ARR^TA^T = AIA^T = AA^T = \Sigma$$

2. Suppose a vectors $x \in \mathbb{R}^2$ are drawn from $N(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 6 & 2 \\ 2 & 5 \end{pmatrix}$$

What is the distribution of the transformed vectors $Ax + b$ such that

$$b = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$
$$A = \begin{pmatrix} 1 & 5 \\ 3 & 2 \end{pmatrix}$$

Solution:

Its a matrix-vector/matrix-matrix multiplication task.

Remember:

$$
\begin{aligned}
(Av)_k &= \sum_{i=1}^{n} A_{ki} v_i \\
(AB)_{kl} &= \sum_{i=1}^{n} A_{ki} B_{il}
\end{aligned}
$$

We have the result parameters being:

$$
\begin{aligned}
\widetilde{\mu} &= A\mu + b \\
\widetilde{\Sigma} &= A\Sigma A^T
\end{aligned}
$$

$$
\begin{aligned}
\widetilde{\mu} &= A\mu + b \\
&= \begin{pmatrix} 1 & 5 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ -2 \end{pmatrix} + \begin{pmatrix} -1 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} 1 \cdot 3 + 5 \cdot -2 \\ 3 \cdot 3 + 2 \cdot -2 \end{pmatrix} + \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} -8 \\ 7 \end{pmatrix}
\end{aligned}
$$

Equally

$$
\begin{aligned}
A\Sigma A^T &= \begin{pmatrix} 1 & 5 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 6 & 2 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix} \\
&= \begin{pmatrix} 16 & 27 \\ 22 & 16 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix} \\
&= \begin{pmatrix} 151 & 102 \\ 102 & 98 \end{pmatrix}
\end{aligned}
$$

3. Suppose we have a set of vector-valued random variables $(X_1, \ldots, X_n)$, $X_i \sim N(\mu, \Sigma)$, where we restrict Σ to be a diagonal matrix, that is

$$\Sigma = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_n^2 \end{pmatrix}$$

Find the maximum likelihood estimator for μ and Σ.

Solution:

Lets rewrite the density for a diagonal covariance matrix. We have

$$
\begin{aligned}
det(\Sigma) &= \prod_{i=1}^{n} s_i^2 \\
(\Sigma^{-1})_{lr} &= \begin{cases} \frac{1}{s_l^2} & \text{if } l = r \\ 0 & \text{otherwise} \end{cases} \\
f(x) &= \frac{1}{(2\pi)^{n/2} det(\Sigma))^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{n/2} \prod_{l=1}^{n} |s_l|} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{(2\pi)^{n/2} \prod_{l=1}^{n} |s_l|} \exp\left(-\frac{1}{2}\sum_{l=1}^{n}\sum_{r=1}^{n}(x_l-\mu_l)^T \Sigma_{lr}^{-1}(x_r-\mu_r)\right) \\
&= \frac{1}{(2\pi)^{n/2} \prod_{l=1}^{n} |s_l|} \exp\left(-\frac{1}{2}\sum_{l=1}^{n}(x_l-\mu_l)^T \Sigma_{ll}^{-1}(x_l-\mu_l)\right) \\
&= \frac{1}{(2\pi)^{n/2} \prod_{l=1}^{n} |s_l|} \exp\left(-\frac{1}{2}\sum_{l=1}^{n}\frac{1}{s_l^2}(x_l-\mu_l)^2\right)
\end{aligned}
$$

The log-likelihood of it for one sample x becomes (under assumption $s_l > 0$, so that $|s_l| = s_l$):

$$
\begin{aligned}
&-\log((2\pi)^{n/2}) - \log(\prod_{l=1}^{n}|s_l|) - \frac{1}{2}\sum_{l=1}^{n}\frac{1}{s_l^2}(x_l-\mu_l)^2 \\
=&-\log((2\pi)^{n/2}) - \sum_{l=1}^{n}\log(s_l) - \frac{1}{2}\sum_{l=1}^{n}\frac{1}{s_l^2}(x_l-\mu_l)^2
\end{aligned}
$$

Now lets assume that we have samples $x^{(1)}, \ldots, x^{(S)}$ (with the k-th dimension of sample $x^{(i)}$ being: $x_k^{(i)}$). Then the log likelihood for the dataset becomes

$$
\begin{aligned}
&-\sum_{i=1}^{S}\left(\sum_{l=1}^{n}\log(s_l) - \frac{1}{2}\sum_{l=1}^{n}\frac{1}{s_l^2}(x_l^{(i)} - \mu_l)^2\right) \\
&= -S\sum_{l=1}^{n}\log(s_l) - \frac{1}{2}\sum_{i=1}^{S}\sum_{l=1}^{n}\frac{1}{s_l^2}(x_l^{(i)} - \mu_l)^2
\end{aligned}
$$

Its derivative for the l-th dimension of the mean μ_l is:

$$
\begin{aligned}
-\frac{1}{2}\sum_{i=1}^{S}\frac{1}{s_l^2}(-2)\cdot(x_l^{(i)} - \mu_l)^1 &= 0 \\
\frac{1}{s_l^2}\sum_{i=1}^{S}(x_l^{(i)} - \mu_l) &= 0 \\
\sum_{i=1}^{S}(x_l^{(i)} - \mu_l) &= 0 \\
\sum_{i=1}^{S}x_l^{(i)} &= S\mu_l \\
\mu_l &= \frac{1}{S}\sum_{i=1}^{S}x_l^{(i)}
\end{aligned}
$$

Writing this in vector notation this yields:

$$
\mu = \frac{1}{S}\sum_{i=1}^{S}x^{(i)}
$$

Its derivative for the l-th dimension of the variance diagonal s_l is:

$$
\begin{aligned}
-S\frac{1}{s_l} - \frac{1}{2}\sum_{i=1}^{S}(-2)\cdot\frac{1}{s_l^3}(x_l^{(i)} - \mu_l)^2 = 0 \mid \cdot s_l^3 \\
\Leftrightarrow -Ss_l^2 + \sum_{i=1}^{S}(x_l^{(i)} - \mu_l)^2 = 0 \\
\Leftrightarrow \sum_{i=1}^{S}(x_l^{(i)} - \mu_l)^2 = Ss_l^2 \\
\Leftrightarrow s_l^2 = \frac{1}{S}\sum_{i=1}^{S}(x_l^{(i)} - \mu_l)^2
\end{aligned}
$$

This is the variance estimate for the l-th dimension of the data if it would have been treated as a separate dimension.