Tutorial Session
Multivariate Gaussian

Suggested Reading

Murphy: chapter 4

- section 4.2 provides you with an idea on how gaussians can be simply used as class conditional densities, either with individual \Sigmas or with shared \Sigmas.
- section 4.3 will become important for Gaussian Processes (compare especially section 4.3.2.2) and contains helpful examples for marginals and conditionals of gaussians.
- section 4.3.4 may be skipped if not enough time available.
- 4.4 is also important and has a more simple example (4.4.2.1) than the more general one on the slides (4.4.2.2)
- 4.5 may be skipped if not enough time available.
- 4.6 as well as the chapters 5 and 6 provide an excellent further insight into bayesian thinking and the connection to classical (frequentist) statistics. this is VERY well invested reading time, but from the course's perspective may be skipped if not enough time available.

Bishop: section 2.3.

- contains most of the relevant stuff in compact form. Actually the whole presentation here is more didactically well thought out, more accessible and more compact than the stuff in Murphy. Even for Murphy addicts this section of Bishop is a true alternative.

Problem 1: Show that the sum of two independent Gaussian random variables ($\boldsymbol{X_1}$ and $\boldsymbol{X_2}$) is Gaussian. Some of the properties of Gaussians mentioned in the lecture can help.

Problem 1: Show that the sum of two independent Gaussian random variables ($\boldsymbol{X}_1$ and $\boldsymbol{X}_2$) is Gaussian. Some of the properties of Gaussians mentioned in the lecture can help.

Let $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, two n dimensional random vectors. *Stack* both vectors to form a random vector $\boldsymbol{Y}$. $\boldsymbol{Y}$ is a Gaussian random variable:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{pmatrix} \right)$$

Consider $\boldsymbol{Z} = \boldsymbol{X}_1 + \boldsymbol{X}_2$. $\boldsymbol{Z}$ can be written as $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{Y}$, with $\boldsymbol{A} = [\boldsymbol{I}_n \ \boldsymbol{I}_n]$. So $\boldsymbol{Z}$ is Gaussian with mean $\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$ and covariance $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ (see the rule for linear transformations of Gaussians from the slides).

Problem 2: Let $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$. Show that the Kullback-Leibler divergence of q from p is $KL(p, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$.

Problem 2: Let $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$. Show that the Kullback-Leibler divergence of q from p is $KL(p,q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$.

We know that

$$KL(p,q) = -\int p(x) \log q(x) dx + \int p(x) \log p(x) dx$$

where integration is done over the real line, and that

$$-\int p(x) \log p(x) dx = \frac{1}{2}(1 + \log 2\pi\sigma_1^2),$$

which is the entropy of a Gaussian. We restrict ourselves to $-H(p,q) = \int p(x) \log q(x) dx$—the negative cross-entropy of p and q—which we can write out as

$$-H(p,q) = \int p(x) \log \frac{1}{(2\pi\sigma_2^2)^{(1/2)}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx,$$

which we can separate into

$$\begin{aligned}
-H(p,q) &= -\frac{1}{2}\log(2\pi\sigma_2^2) - \int p(x) \log e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx \\
&= -\frac{1}{2}\log(2\pi\sigma_2^2) - \int p(x) \frac{(x-\mu_2)^2}{2\sigma_2^2} dx \\
&= -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{\int p(x) x^2 dx - \int p(x) 2x\mu dx + \int p(x) \mu^2 dx}{2\sigma_2^2}
\end{aligned}$$

where we separated the sums and got σ_2^2 out of the integral.

Letting $\langle\rangle$ denote the expectation operator under p, we can rewrite this as

$$-H(p,q) = -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{\langle x^2\rangle - 2\langle x\rangle\mu_2 + \mu_2^2}{2\sigma_2^2}.$$

We know that $var(x) = \langle x^2\rangle - \langle x\rangle^2$. Thus $\langle x^2\rangle = \sigma_1^2 + \mu_1^2$ and therefore

$$-H(p,q) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2}{2\sigma_2^2},$$

which we can put as

$$-H(p,q) = -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}.$$

Putting everything together, we get to

$$KL(p,q) = -\int p(x)\log q(x)dx + \int p(x)\log p(x)dx$$

$$= H(p,q) - H(p)$$

Putting everything together, we get to

$$
\begin{aligned}
KL(p,q) &= -\int p(x)\log q(x)dx + \int p(x)\log p(x)dx \\
&= H(p,q) - H(p) \\
&= \frac{1}{2}\log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}(1+\log 2\pi\sigma_1^2) \\
&= \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.
\end{aligned}
$$

For a discussion see `http://stats.stackexchange.com/questions/7440/kl-divergence-between-two-univariate-gaussians`

Problem 3: We can sample from any multivariate Gaussian by using:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z} \Rightarrow \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Show that this works by using the change of variable theorem.

Problem 3: We can sample from any multivariate Gaussian by using:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z} \Rightarrow \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Show that this works by using the change of variable theorem.

We apply the linear transformation $y = \boldsymbol{L}x + \boldsymbol{\mu}$. $\boldsymbol{x}(\boldsymbol{y})$ is therefore $x = \boldsymbol{L}^{-1}(y - \boldsymbol{\mu})$

Together with the change of variable theorem we get:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}(\boldsymbol{y}))\left|\frac{\boldsymbol{dx}}{\boldsymbol{dy}}\right| = f(\boldsymbol{L}^{-1})\,|\boldsymbol{L}| \propto e^{-\frac{1}{2}(\boldsymbol{L}^{-1}(y-\boldsymbol{\mu}))^T \boldsymbol{I}^{-1}(\boldsymbol{L}^{-1}(y-\boldsymbol{\mu}))}$$

$$= e^{-\frac{1}{2}(y-\boldsymbol{\mu})^T \boldsymbol{L}^{-T}\boldsymbol{I}^{-1}\boldsymbol{L}^{-1}(y-\boldsymbol{\mu})}$$

$$= e^{-\frac{1}{2}(y-\boldsymbol{\mu})^T (\boldsymbol{L}\boldsymbol{L}^T)^{-1}(y-\boldsymbol{\mu})}$$

showing that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$.

Excursus: Cholesky Decomposition

Source: course slides ee103: applied numerical computing,
UCLA, fall 2011, prof Vandenberghe

Triangular matrix

a square matrix A is **lower triangular** if $a_{ij} = 0$ for $j > i$

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 & 0 \\ a_{21} & a_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ a_{n-1,1} & a_{n-1,2} & \cdots & a_{n-1,n-1} & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix}$$

A is **upper triangular** if $a_{ij} = 0$ for $j < i$ (A^T is lower triangular)

a triangular matrix is **unit upper/lower triangular** if $a_{ii} = 1$ for all i

Forward substitution

solve $Ax = b$ when A is lower triangular with nonzero diagonal elements

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

algorithm:

$$\begin{aligned} x_1 &:= b_1/a_{11} \\ x_2 &:= (b_2 - a_{21}x_1)/a_{22} \\ x_3 &:= (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \\ &\vdots \\ x_n &:= (b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1})/a_{nn} \end{aligned}$$

cost: $1 + 3 + 5 + \cdots + (2n-1) = n^2$ flops

Back substitution

solve $Ax = b$ when A is upper triangular with nonzero diagonal elements

$$\begin{bmatrix} a_{11} & \cdots & a_{1,n-1} & a_{1n} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

algorithm:

$$\begin{aligned} x_n &:= b_n/a_{nn} \\ x_{n-1} &:= (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n-1} \\ x_{n-2} &:= (b_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n)/a_{n-2,n-2} \\ &\vdots \\ x_1 &:= (b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n)/a_{11} \end{aligned}$$

cost: n^2 flops

Inverse of a triangular matrix

triangular matrix A with nonzero diagonal elements is nonsingular

- $Ax = b$ is solvable via forward/back substitution; hence A has full range
- therefore A has a zero nullspace, is invertible, etc. (see p.4-8)

inverse

- can be computed by solving $AX = I$ column by column

$$A \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix} = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}$$

- inverse of lower triangular matrix is lower triangular
- inverse of upper triangular matrix is upper triangular

Cholesky factorization

every positive definite matrix A can be factored as

$$A = LL^T$$

where L is lower triangular with positive diagonal elements

cost: $(1/3)n^3$ flops if A is of order n

- L is called the *Cholesky factor* of A
- can be interpreted as 'square root' of a positive define matrix

Cholesky factorization algorithm

partition matrices in $A = LL^T$ as

$$\begin{aligned}\left[\begin{array}{cc} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{array}\right] &= \left[\begin{array}{cc} l_{11} & 0 \\ L_{21} & L_{22} \end{array}\right] \left[\begin{array}{cc} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{array}\right] \\ &= \left[\begin{array}{cc} l_{11}^2 & l_{11}L_{21}^T \\ l_{11}L_{21} & L_{21}L_{21}^T + L_{22}L_{22}^T \end{array}\right]\end{aligned}$$

algorithm

1. determine l_{11} and L_{21}:

$$l_{11} = \sqrt{a_{11}}, \qquad L_{21} = \frac{1}{l_{11}} A_{21}$$

2. compute L_{22} from

$$A_{22} - L_{21}L_{21}^T = L_{22}L_{22}^T$$

this is a Cholesky factorization of order $n-1$

proof that the algorithm works for positive definite A of order n

- step 1: if A is positive definite then $a_{11} > 0$
- step 2: if A is positive definite, then

  $$A_{22} - L_{21}L_{21}^T = A_{22} - \frac{1}{a_{11}}A_{21}A_{21}^T$$

  is positive definite (see page 4-23)
- hence the algorithm works for $n = m$ if it works for $n = m - 1$
- it obviously works for $n = 1$; therefore it works for all n

Example

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

- first column of L

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & l_{22} & 0 \\ -1 & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 5 & 3 & -1 \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

- second column of L

$$\begin{bmatrix} 18 & 0 \\ 0 & 11 \end{bmatrix} - \begin{bmatrix} 3 \\ -1 \end{bmatrix} \begin{bmatrix} 3 & -1 \end{bmatrix} = \begin{bmatrix} l_{22} & 0 \\ l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{22} & l_{32} \\ 0 & l_{33} \end{bmatrix}$$

$$\begin{bmatrix} 9 & 3 \\ 3 & 10 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & l_{33} \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & l_{33} \end{bmatrix}$$

- third column of L: $10 - 1 = l_{33}^2$, *i.e.*, $l_{33} = 3$

conclusion:

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 3 & 0 \\ -1 & 1 & 3 \end{bmatrix} \begin{bmatrix} 5 & 3 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$

End of Excursus: Cholesky Decomposition

Source: course slides ee103: applied numerical computing, UCLA, fall 2011, prof Vandenberghe

Problem 4: The unbiased estimates for the covariance of a d-dimensional Gaussian based on n samples is given by

$$\hat{\Sigma} = C_n = \frac{1}{n-1}\sum_{i=1}^{n}(x - \mu_n)(x - \mu_n)^T$$

It is clear that it takes $O(nd^2)$ time to compute C_n. If the data points arrive one at a time, it is more efficient to incrementally update these estimates than to recompute from scratch.

Show that the covariance can be sequentially udpated as follows

$$C_{n+1} = \frac{n-1}{n}C_n + \frac{1}{n+1}(x_{n+1} - \mu_n)(x_{n+1} - \mu_n)^T$$

Problem 4: The unbiased estimates for the covariance of a d-dimensional Gaussian based on n samples is given by

$$\hat{\Sigma} = C_n = \frac{1}{n-1}\sum_{i=1}^{n}(x-\mu_n)(x-\mu_n)^T$$

It is clear that it takes $O(nd^2)$ time to compute C_n. If the data points arrive one at a time, it is more efficient to incrementally update these estimates than to recompute from scratch.

Show that the covariance can be sequentially udpated as follows

$$C_{n+1} = \frac{n-1}{n}C_n + \frac{1}{n+1}(x_{n+1}-\mu_n)(x_{n+1}-\mu_n)^T$$

$$\begin{aligned}
\mathbf{C}_{n+1} &= \frac{1}{n}\sum_{k=1}^{n+1}(\mathbf{x}_k-\mathbf{m}_{n+1})(\mathbf{x}_k-\mathbf{m}_{n+1})^T \\
&= \frac{1}{n}\left[\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_{n+1})(\mathbf{x}_k-\mathbf{m}_{n+1})^T + (\mathbf{x}_{n+1}-\mathbf{m}_{n+1})(\mathbf{x}_{n+1}-\mathbf{m}_{n+1})^T\right] \\
&= \frac{1}{n}\left[\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)(\mathbf{x}_k-\mathbf{m}_n)^T - \frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)^T\right. \\
&- \frac{1}{(n+1)}\left(\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)\right)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T + \frac{1}{(n+1)^2}\sum_{k=1}^{n}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n) \\
&+ \left.\frac{1}{n}\left((\mathbf{x}_{n+1}-\mathbf{m}_n)-\frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)\right)\left((\mathbf{x}_{n+1}-\mathbf{m}_n)-\frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)\right)\right] \\
&= \frac{1}{n}\left[(n-1)\mathbf{C}_n + \frac{n}{(n+1)^2}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T\right] \\
&+ \frac{1}{n}\left(\left(\frac{n}{n+1}\right)^2(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T\right) \\
&= \frac{n-1}{n}\mathbf{C}_n + \left(\frac{1}{(n+1)^2}+\frac{n}{(n+1)^2}\right)(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T \\
&= \frac{n-1}{n}\mathbf{C}_n + \frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^T
\end{aligned}$$

Problem 5: We consider a partitioning of the components of x into three groups $\boldsymbol{x}_a$, $\boldsymbol{x}_b$, and $\boldsymbol{x}_c$, with a corresponding partitioning of the mean vector $\boldsymbol{\mu}$ and of the covariance matrix $\boldsymbol{\Sigma}$ in the form

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bc} \\ \boldsymbol{\Sigma}_{ca} & \boldsymbol{\Sigma}_{cb} & \boldsymbol{\Sigma}_{cc} \end{pmatrix}$$

Find an expression for the conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ in which $\boldsymbol{x}_c$ has been marginalized out.

Problem 5: We consider a partitioning of the components of x into three groups $\boldsymbol{x}_a$, $\boldsymbol{x}_b$, and $\boldsymbol{x}_c$, with a corresponding partitioning of the mean vector $\boldsymbol{\mu}$ and of the covariance matrix $\boldsymbol{\Sigma}$ in the form

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bc} \\ \boldsymbol{\Sigma}_{ca} & \boldsymbol{\Sigma}_{cb} & \boldsymbol{\Sigma}_{cc} \end{pmatrix}$$

Find an expression for the conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ in which $\boldsymbol{x}_c$ has been marginalized out.

We first of all take the joint distribution $p(\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c)$ and marginalize to obtain the distribution $p(\boldsymbol{x}_a, \boldsymbol{x}_b)$. This is again a Gaussian distribution with mean and covariance given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

The distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ is then Gaussian with mean and covariance given by

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

and

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

respectively.