# Manipulating the Multivariate Gaussian Density

Thomas B. Schön and Fredrik Lindsten
Division of Automatic Control
Linköping University
SE–58183 Linköping, Sweden.
E-mail: {schon, lindsten}@isy.liu.se

January 11, 2011

**Abstract**

In these note we provide some important properties of the multivariate Gaussian, which are important building blocks for more sophisticated probabilistic models. We also illustrate how these properties can be used to derive the Kalman filter.

## 1 Introduction

The most common way of parameterizing the multivariate Gaussian (a.k.a. Normal) density function is according to

$$\mathcal{N}(x\,;\,\mu,\Sigma) \triangleq \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{1}$$

where $x \in \mathbb{R}^n$ denotes a random vector that is Gaussian distributed, with mean value $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in S^n_{++}$ (i.e., the $n$-dimensional positive definite cone). Furthermore, $x \sim \mathcal{N}(\mu,\Sigma)$ states that the random vector $x$ is Gaussian with mean $\mu$ and covariance $\Sigma$. The parameterizing (1) is sometimes referred to as the *moment* form. An alternative parameterization of the Gaussian density is provided by the *information* form, which is also referred to as the *canonical* form or the *natural* form. This parameterization

$$\mathcal{N}^{-1}(x;\nu,\Lambda) \triangleq \frac{\exp\left(-\frac{1}{2}\nu^T\Lambda^{-1}\nu\right)}{(2\pi)^{n/2}\sqrt{\det\Lambda^{-1}}}\exp\left(-\frac{1}{2}x^T\Lambda x + x^T\nu\right), \tag{2}$$

is parameterized by the information vector $\nu$ and (Fisher) information matrix $\Lambda$. It is straightforward to see the relationship between the two parameterizations (1) and (2),

$$\begin{aligned}\mathcal{N}(x\,;\,\mu,\Sigma) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right)\\ &= \frac{\exp\left(-\frac{1}{2}\mu^T\Sigma^{-1}\mu\right)}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu\right),\end{aligned} \tag{3}$$

which also reveals the that the parameters used in the information form are given by a scaled version of the mean,

$$\nu = \Sigma^{-1}\mu \tag{4a}$$

and the inverse of the covariance matrix (sometimes also called the precision matrix)

$$\Lambda = \Sigma^{-1}. \tag{4b}$$

The moment form (1) and the information form (2) results in different computations. It is useful to understand these differences in order to derive efficient algorithms.

## 2 Partitioned Gaussian Densities

Let us, without loss of generality, assume that random vector $x$, its mean $\mu$ and its covariance $\Sigma$ can be partitioned according to

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{5}$$

where for reasons of symmetry $\Sigma_{ba} = \Sigma_{ab}^T$. It is also useful to write down the partitioned information matrix

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}, \tag{6}$$

since this form will provide simpler calculations below. Note that, since the inverse of a symmetric matrix is also symmetric, we have $\Lambda_{ab} = \Lambda_{ba}^T$.

We will now derive two important and very useful theorems for partitioned Gaussian densities. These theorems will explain the two operations marginalization and conditioning.

**Theorem 1** ***(Marginalization)*** *Let the random vector $x$ be Gaussian distributed according to* (1) *and let it be partitioned according to* (5)*, then the marginal density $p(x_a)$ is given by*

$$p(x_a) = \mathcal{N}(x_a\,;\,\mu_a, \Sigma_{aa}). \tag{7}$$

**Proof:** See Appendix A.1. ■

**Theorem 2** ***(Conditioning)*** *Let the random vector $x$ be Gaussian distributed according to* (1) *and let it be partitioned according to* (5)*, then the conditional density $p(x_a \mid x_b)$ is given by*

$$p(x_a \mid x_b) = \mathcal{N}\left(x_a\,;\,\mu_{a|b}, \Sigma_{a|b}\right), \tag{8a}$$

*where*

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \tag{8b}$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}, \tag{8c}$$

*which using the information matrix can be written,*

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b), \tag{8d}$$

$$\Sigma_{a|b} = \Lambda_{aa}^{-1}. \tag{8e}$$

**Proof:** See Appendix A.1. ■

## 3 Affine Transformations

In the previous section we dealt with partitioned Gaussian densities, and derived the expressions for the marginal and conditional densities expressed in terms of the parameters of the joint density. We shall now take a different starting point, namely that we are given the marginal density $p(x_a)$ and the conditional density $p(x_b \mid x_a)$ (affine in $x_a$) and derive expressions for the joint density $p(x_a, x_b)$, the marginal density $p(x_b)$ and the conditional density $p(x_a \mid x_b)$.

**Theorem 3** ***(Affine transformation)*** *Assume that $x_a$, as well as $x_b$ conditioned on $x_a$, are Gaussian distributed according to*

$$p(x_a) = \mathcal{N}\left(x_a\, ;\, \mu_a, \Sigma_a\right), \tag{9a}$$

$$p(x_b \mid x_a) = \mathcal{N}\left(x_b\, ;\, Mx_a + b, \Sigma_{b|a}\right), \tag{9b}$$

*where $M$ is a matrix (of appropriate dimension) and $b$ is a constant vector. The joint distribution of $x_a$ and $x_b$ is then given by*

$$p(x_a, x_b) = \mathcal{N}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix}; \begin{pmatrix} \mu_a \\ M\mu_a + b \end{pmatrix}, R\right), \tag{9c}$$

*with*

$$R = \begin{pmatrix} M^T\Sigma_{b|a}^{-1}M + \Sigma_a^{-1} & -M^T\Sigma_{b|a}^{-1} \\ -\Sigma_{b|a}^{-1}M & \Sigma_{b|a}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_a & \Sigma_a M^T \\ M\Sigma_a & \Sigma_{b|a} + M\Sigma_a M^T \end{pmatrix}. \tag{9d}$$

**Proof:** See Appendix A.2. ■

Combining the results in Theorems 1, 2 and 3 we also get the following corollary.

**Corollary 1** ***(Affine transformation – marginal and conditional)*** *Assume that $x_a$, as well as $x_b$ conditioned on $x_a$, are Gaussian distributed according to*

$$p(x_a) = \mathcal{N}\left(x_a\, ;\, \mu_a, \Sigma_a\right), \tag{10a}$$

$$p(x_b \mid x_a) = \mathcal{N}\left(x_b\, ;\, Mx_a + b, \Sigma_{b|a}\right), \tag{10b}$$

*where $M$ is a matrix (of appropriate dimension) and $b$ is a constant vector. The marginal density of $x_b$ is then given by*

$$p(x_b) = \mathcal{N}\left(x_b\, ;\, \mu_b, \Sigma_b\right), \tag{10c}$$

*with*

$$\mu_b = M\mu_a + b, \tag{10d}$$

$$\Sigma_b = \Sigma_{b|a} + M\Sigma_a M^T. \tag{10e}$$

*The conditional density of $x_a$ given $x_b$ is*

$$p(x_a \mid x_b) = \mathcal{N}\left(x_a\, ;\, \mu_{a|b}, \Sigma_{a|b}\right), \tag{10f}$$

*with*

$$\mu_{a|b} = \Sigma_{a|b}\left(M^T\Sigma_{b|a}^{-1}(x_b - b) + \Sigma_a^{-1}\mu_a\right) = \mu_a + \Sigma_a M^T \Sigma_b^{-1}(x_b - b - M\mu_a), \tag{10g}$$

$$\Sigma_{a|b} = \left(\Sigma_a^{-1} + M^T\Sigma_{b|a}^{-1}M\right)^{-1} = \Sigma_a - \Sigma_a M^T \Sigma_b^{-1} M \Sigma_a. \tag{10h}$$

## 4 Example – Kalman filter

Consider the following linear Gaussian state-space model

$$x_{t+1} = Ax_t + b_t + w_t, \tag{11a}$$
$$y_t = Cx_t + d_t + e_t, \tag{11b}$$

where

$$w_t \sim \mathcal{N}(0, Q), \qquad e_t \sim \mathcal{N}(0, R), \tag{12}$$

and $b_t$ and $d_t$ are known vectors (e.g. from a known input). Furthermore, assume that the initial condition is Gaussian according to

$$x_1 \sim \mathcal{N}(\hat{x}_{1|0}, P_{1|0}). \tag{13}$$

As pointed out in the previous section, affine transformations conserve Gaussianity. Hence, the filtering density $p(x_t \mid y_{1:t})$ and the 1-step prediction density $p(x_t \mid y_{1:t-1})$ will also be Gaussian for any $t \geq 1$ (with $p(x_1 \mid y_0) \triangleq p(x_1)$).

In this probabilistic setting, we can view the Kalman filter as a filter keeping track of the mean vector and covariance matrix of the filtering density (and the 1-step prediction density). Since these statistics are sufficient for the Gaussian distribution, the Kalman filter is clearly optimal, since it holds all information about the filtering density and 1-step prediction density.

To derive that Kalman filter, all that we need is Corollary 1. The derivation is done inductively. Hence, assume that

$$p(x_t \mid y_{1:t-1}) = \mathcal{N}\left(x_{t-1}\,;\, \hat{x}_{t|t-1}, P_{t|t-1}\right) \tag{14a}$$

(which, according to (13) is true at $t = 1$). From (11b) we have

$$p(y_t \mid x_t, y_{1:t-1}) = p(y_t|x_t) = \mathcal{N}\left(y_t\,;\, Cx_t + d_t, R\right). \tag{14b}$$

By identifying $x_t \leftrightarrow x_a$ and $y_t \leftrightarrow x_b$ and using Corollary 1 (the part concerning the conditional density) we get

$$p(x_t \mid y_{1:t}) = \mathcal{N}\left(x_t\,;\, \hat{x}_{t|t}, P_{t|t}\right), \tag{15}$$

with

$$P_{t|t} = P_{t|t-1} - K_t C P_{t|t-1}, \tag{16a}$$
$$K_t = P_{t|t-1} C^T S_t^{-1}, \tag{16b}$$
$$S_t = C P_{t|t-1} C^T + R. \tag{16c}$$

Furthermore,

$$\begin{aligned}
\hat{x}_{t|t} &= P_{t|t}\left(C^T R^{-1}(y_t - d_t) + (P_{t|t-1})^{-1}\hat{x}_{t|t-1}\right) \\
&= (P_{t|t-1} - K_t C P_{t|t-1}) C^T R^{-1}(y_t - d_t) + P_{t|t}(P_{t|t-1})^{-1}\hat{x}_{t|t-1} \\
&= (\underbrace{P_{t|t-1}C^T}_{=K_t S_t} - K_t \underbrace{C P_{t|t-1} C^T}_{=S_t - R}) R^{-1}(y_t - d_t) + P_{t|t}(P_{t|t-1})^{-1}\hat{x}_{t|t-1} \\
&= K_t(S_t - S_t + R)R^{-1}(y_t - d_t) + (P_{t|t-1} - K_t C P_{t|t-1})(P_{t|t-1})^{-1}\hat{x}_{t|t-1} \\
&= K_t(y_t - d_t) + (I - K_t C)\hat{x}_{t|t-1} \\
&= \hat{x}_{t|t-1} + K_t(y_t - d_t - C\hat{x}_{t|t-1}),
\end{aligned} \tag{17}$$

which completes the measurement update of the filter.

For the time update, note that by (11a)

$$p(x_{t+1} \mid x_t, y_{1:t}) = p(x_{t+1} \mid x_t) = \mathcal{N}\left(x_{t+1}\,;\, Ax_t + b_t, Q\right). \tag{18}$$

From (15) and (18) we can again make use of Corollary 1 (the part concerning the marginal density). By identifying $x_t \leftrightarrow x_a$, $x_{t+1} \leftrightarrow x_b$ we get

$$p(x_{t+1} \mid y_{1:t}) = \mathcal{N}\left(x_{t+1}\,;\, \hat{x}_{t+1|t}, P_{t+1|t}\right), \tag{19}$$

with

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t} + b_t, \tag{20}$$

$$P_{t+1|t} = AP_{t|t}A^T + Q, \tag{21}$$

which completes the time update and the derivation of the Kalman filter.

**Remark 1** *The extra conditioning on* $y_{1:t}$ *in both densities* (15) *and* (18) *does not change the properties of the Gaussians according to Corollary 1. Since we are conditioning on it in both distributions,* $y_{1:t}$ *can be viewed as a constant.*

# A Proofs

## A.1 Partitioned Gaussian Densities

**Proof 1** ***(Proof of Theorem 1)*** *The marginal density $p(x_a)$ is by definition obtained by integrating out the $x_b$ variables from the joint density $p(x_a, x_b)$ according to*

$$p(x_a) = \int p(x_a, x_b) dx_b, \tag{22}$$

*where*

$$p(x_a, x_b) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp(E) \tag{23}$$

*and $\Sigma$ was defined in (5) and the exponent $E$ is given by*

$$\begin{aligned}
E = & -\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b) \\
& - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b) \\
= & -\frac{1}{2}\big(x_b^T \Lambda_{bb} x_b - 2x_b^T \Lambda_{bb}(\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)) - 2x_a^T \Lambda_{ab}\mu_b \\
& + 2\mu_a^T \Lambda_{ab}\mu_b + \mu_b^T \Lambda_{bb}\mu_b + x_a^T \Lambda_{aa} x_a - 2x_a^T \Lambda_{aa}\mu_a + \mu_a^T \Lambda_{aa}\mu_a\big)
\end{aligned} \tag{24}$$

*Completing the squares in the above expression results in*

$$\begin{aligned}
E = & -\frac{1}{2}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)))^T \Lambda_{bb}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a))) \\
& + \frac{1}{2}\big(x_a^T \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba} x_a - 2x_a^T \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a + \mu_a \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a\big) \\
& - \frac{1}{2}\big(x_a^T \Lambda_{aa} x_a - 2x_a^T \Lambda_{aa}\mu_a + \mu_a^T \Lambda_{aa}\mu_a\big) \\
= & -\frac{1}{2}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)))^T \Lambda_{bb}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a))) \\
& - \frac{1}{2}(x_a - \mu_a)^T (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(x_a - \mu_a)
\end{aligned} \tag{25}$$

*Using the block matrix inversion provided in (56) we have*

$$\Sigma_{aa}^{-1} = \Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}. \tag{26}$$

*which together with (25) results in*

$$p(x_a, x_b) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp(E_1)\exp(E_2) \tag{27}$$

*where*

$$E_1 = -\frac{1}{2}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)))^T \Lambda_{bb}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a))), \tag{28a}$$

$$E_2 = -\frac{1}{2}(x_a - \mu_a)^T (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(x_a - \mu_a). \tag{28b}$$

*Now, since $E_2$ is independent of $x_b$ we have*

$$p(x_a) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \int \exp(E_1) dx_b \exp(E_2). \tag{29}$$

*Since the integral of a density function is equal to one, we have*

$$\int \exp(E_1) dx_b = (2\pi)^{n_b/2}\sqrt{\det \Lambda_{bb}^{-1}} \tag{30}$$

*which inserted in* (29) *results in*

$$p(x_a) = \frac{\sqrt{\det \Lambda_{bb}^{-1}}}{(2\pi)^{n_a/2}\sqrt{\det\Sigma}} \int \exp((x_a - \mu_a)^T \Sigma_{aa}^{-1}(x_a - \mu_a)). \tag{31}$$

*Finally, the determinant property* (54b) *gives*

$$\det\Sigma = \det\Sigma_{aa}\det(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}) \tag{32}$$

*and the block matrix inversion property* (56) *provides*

$$\Lambda_{bb}^{-1} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}. \tag{33}$$

*Now, inserting* (32) *and* (33) *into* (31) *concludes the proof.*

**Proof 2** ***(Proof of Theorem 2)*** *We will make use of the fact that*

$$p(x_a \mid x_b) = \frac{p(x_a, x_b)}{p(x_b)}, \tag{34}$$

*which according to the definition* (1) *is*

$$p(x_a \mid x_b) = \sqrt{\frac{\det\Sigma_{bb}}{(2\pi)^{n_a/2}\det\Sigma}} \exp(E), \tag{35}$$

$$E = -\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu) + \frac{1}{2}(x_b - \mu_b)\Sigma_{bb}^{-1}(x_b - \mu_b). \tag{36}$$

*Let us first consider the constant in front of the exponential in* (35),

$$\sqrt{\frac{\det\Sigma_{bb}}{(2\pi)^{n_a/2}\det\Sigma}}. \tag{37}$$

*Using the result on determinants given in* (54b) *we have*

$$\det\Sigma = \det\begin{pmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix} = \det\Sigma_{bb}\det(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}) \tag{38}$$

*which inserted into* (37) *results in the following expression for the constant in front of the exponent*

$$\frac{1}{(2\pi)^{n_a/2}\det(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})} \tag{39}$$

*Let us now continue by studying the exponent of* (35)*, which using the precision matrix is given by*

$$
\begin{aligned}
E &= -\frac{1}{2}(x-\mu)^T \Lambda (x-\mu) - \frac{1}{2}(x_b-\mu_b)\Sigma_{bb}^{-1}(x_b-\mu_b) \\
&= -\frac{1}{2}(x_a-\mu_a)^T \Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T \Lambda_{ab}(x_b-\mu_b) \\
&\quad - \frac{1}{2}(x_b-\mu_b)^T \Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T (\Lambda_{bb}-\Sigma_{bb}^{-1})(x_b-\mu_b).
\end{aligned} \tag{40}
$$

*Reordering the terms in* (40) *results in*

$$
\begin{aligned}
E &= -\frac{1}{2}x_a^T \Lambda_{aa} x_a + x_a^T \left(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\right) \\
&\quad - \frac{1}{2}\mu_a^T \Lambda_{aa}\mu_a + \mu_a^T \Lambda_{ab}(x_b-\mu_b) - \frac{1}{2}(x_b-\mu_b)^T(\Lambda_{bb}-\Sigma_{bb}^{-1})(x_b-\mu_b).
\end{aligned} \tag{41}
$$

*Using the block matrix inversion result* (56)

$$
\Sigma_{bb}^{-1} = \Lambda_{bb} - \Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab} \tag{42}
$$

*in* (41) *results in*

$$
\begin{aligned}
E &= -\frac{1}{2}x_a^T \Lambda_{aa} x_a + x_a^T \left(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\right) \\
&\quad - \frac{1}{2}\mu_a^T \Lambda_{aa}\mu_a + \mu_a^T \Lambda_{ab}(x_b-\mu_b) - \frac{1}{2}(x_b-\mu_b)^T \Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b).
\end{aligned} \tag{43}
$$

*We can now complete the squares, which gives us*

$$
E = -\frac{1}{2}\left(x_a - (\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b))\right)^T \Lambda_{aa}\left(x_a - (\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b))\right). \tag{44}
$$

*Finally, combining* (39) *and* (44) *results in*

$$
p(x_a \mid x_b) = \frac{1}{(2\pi)^{nx/2}\det \Lambda_{aa}^{-1}} \exp(E) \tag{45}
$$

*which concludes the proof.*

## A.2 Affine Transformation

**Proof 3** ***(Proof of Theorem 3)*** *We start by introducing the vector*

$$
x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \tag{46}
$$

*for which the joint distribution is given by*

$$
p(x) = p(x_b \mid x_a)p(x_a) = \frac{(2\pi)^{-(n_a+n_b)/2}}{\sqrt{\det \Sigma_{b|a} \det \Sigma_a}} e^{-\frac{1}{2}E}, \tag{47}
$$

*where*

$$
E = (x_b - Mx_a - b)^T \Sigma_{b|a}^{-1}(x_b - Mx_a - b) + (x_a-\mu_a)^T \Sigma_a^{-1}(x_a-\mu_a). \tag{48}
$$

*Introduce the following variables*

$$e = x_a - \mu_a, \tag{49a}$$

$$f = x_b - M\mu_a - b, \tag{49b}$$

*which allows us to write the exponent* (48) *as*

$$\begin{aligned}
E &= (f - Me)^T \Sigma_{b|a}^{-1}(f - Me) + e^T \Sigma_a^{-1} e \\
&= e^T (M^T \Sigma_{b|a}^{-1} M + \Sigma_a^{-1}) e - e^T M^T \Sigma_{b|a}^{-1} f - f \Sigma_{b|a}^{-1} M e + f^T \Sigma_{b|a}^{-1} f \\
&= \begin{pmatrix} e \\ f \end{pmatrix}^T \underbrace{\begin{pmatrix} M^T \Sigma_{b|a}^{-1} M + \Sigma_a^{-1} & -M^T \Sigma_{b|a}^{-1} \\ -\Sigma_{b|a}^{-1} M & \Sigma_{b|a}^{-1} \end{pmatrix}}_{\triangleq R^{-1}} \begin{pmatrix} e \\ f \end{pmatrix} \\
&= \begin{pmatrix} x_a - \mu_a \\ x_b - M\mu_a - b \end{pmatrix}^T R^{-1} \begin{pmatrix} x_a - \mu_a \\ x_b - M\mu_a - b \end{pmatrix}
\end{aligned} \tag{50}$$

*Furthermore, from* (54b) *we have that*

$$\begin{aligned}
\frac{1}{\det R} &= \det R^{-1} = \det\left(\Sigma_{b|a}^{-1}\right) \det\left(M^T \Sigma_{b|a}^{-1} M + \Sigma_a^{-1} - M^T \Sigma_{b|a}^{-1} \Sigma_{b|a} \Sigma_{b|a}^{-1} M\right) \\
&= \det\left(\Sigma_{b|a}^{-1}\right) \det\left(\Sigma_a^{-1}\right) = \frac{1}{\det\left(\Sigma_{b|a}\right) \det\left(\Sigma_a\right)}.
\end{aligned} \tag{51}$$

*Hence, from* (47), (50) *and* (51) *we can write the joint PDF for* $x$ *as*

$$\begin{aligned}
p(x) &= \frac{(2\pi)^{-(n_a+n_b)/2}}{\sqrt{\det R}} \exp\left(-\frac{1}{2}\left(\begin{pmatrix} x_a - \mu_a \\ x_b - M\mu_a - b \end{pmatrix}^T R^{-1} \begin{pmatrix} x_a - \mu_a \\ x_b - M\mu_a - b \end{pmatrix}\right)\right) \\
&= \mathcal{N}\left(x\,;\begin{pmatrix} \mu_a \\ M\mu_a + b \end{pmatrix}, R\right)
\end{aligned} \tag{52}$$

*which concludes the proof.*

# B Matrix Identities

This appendix provides some useful results from matrix theory.

Consider the following block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{53}$$

The following results hold for the determinant

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det A \det(\underbrace{D - CA^{-1}B}_{\Delta_A}), \tag{54a}$$

$$= \det D \det(\underbrace{A - BD^{-1}C}_{\Delta_D}), \tag{54b}$$

where $\Delta_D = A - BD^{-1}C$ is referred to as the *Schur complement* of $D$ in $M$ and $\Delta_A = D - CA^{-1}B$ is referred to as the Schur complement of $A$ in $M$.

When the block matrix (53) is invertible its inverse can be written according to

$$\begin{aligned} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & \Delta_A^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} A^{-1} + A^{-1}B\Delta_A^{-1}CA^{-1} & -A^{-1}B\Delta_A^{-1} \\ -\Delta_A^{-1}CA^{-1} & \Delta_A^{-1} \end{pmatrix}, \end{aligned} \tag{55}$$

where we have made use of the Schur complement of $A$ in $M$. We can also use the Schur complement of $D$ in $M$, resulting in

$$\begin{aligned} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} &= \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} \Delta_D^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \Delta_D^{-1} & \Delta_D^{-1}BD^{-1} \\ -D^{-1}C\Delta_D^{-1} & D^{-1} + D^{-1}C\Delta_D^{-1}BD^{-1} \end{pmatrix}. \end{aligned} \tag{56}$$

The *matrix inversion lemma*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{57}$$

under the assumption that the involved inverses exist. It is worth noting that the matrix inversion lemma is sometimes also referred to as the Woodbury matrix identity, the Sherman-Morrison-Woodbury formula or the Woodbury formula.