RJ 10202 (95076) December 1, 2000
Computer Science

Research Report

BAYESIAN MARGINAL LIKELIHOOD FOR MULTIVARIATE GAUSSIAN WITH CONJUGATE PRIORS

Byron Dom and Alex Cozzi

Research Division
Yorktown Heights, New York • San Jose, California • Zurich, Switzerland

BAYESIAN MARGINAL LIKELIHOOD FOR MULTIVARIATE GAUSSIAN WITH CONJUGATE PRIORS

Byron Dom and Alex Cozzi

ABSTRACT: This report is intended as a tutorial derivation of the Bayesian marginal likelihood/probability corresponding to multivariate Gaussian (Normal) models and conjugate priors, which are Gaussian for the mean vector μ and Inverse Wishart for the covariance matrix $\mathbf{\Sigma}$. While we have found this result stated elsewhere, we have not found an easily accessible treatment of its derivation allowing various terms to be easily interpreted unambiguously and so on. We hope that what is presented here will be at least somewhat usefull in filling this apparent void.

1. Background

1.1. General Background

The background and context for this problem lie in the subject of Bayesian statistical estimation[1], whose central identifying feature is the treatment of parameters in probability distributions (e.g. the *mean* μ and *covariance* $\mathbf{\Sigma}$ in the case of multivariate Gaussian distributions) as random variables with their own distribution referred to as the *prior distribution* or just "prior". A convention often used (and followed by us here) in generic discussions of Bayesian theory is to represent the collection of these parameters (a.k.a. the "parameter vector") by θ and the *prior* by $\pi(\theta)$. For a set X of data vectors x_i where:

$$X \quad = \quad \{x_i | i = 1, 2, \ldots, n\}, \quad x_i \in \mathbb{R}^d$$

we represent the specific parametric probability model form (e.g. multivariate Gaussian) by $P(X|\theta)$ and we can also use these definitions to introduce the joint distribution:

$$P(X, \theta) \quad = \quad P(X|\theta)\, \pi(\theta).$$

The *marginal probability* (or probability density) of X is obtained by integrating $P(X, \theta)$ over the domain Θ of θ:

$$P(X) \quad = \quad \int_\Theta P(X, \theta)\, d\theta$$

A distribution classically more interesting to Bayesian statisticians is the *posterior* $P(\theta|X)$ over the parameters θ:

$$P(\theta|X) \quad = \quad \frac{P(X|\theta)\, \pi(\theta)}{P(X)}.$$

For example, the *maximum a posteriori* (MAP) estimate for θ, $\hat{\theta}$ is the one that maximizes this.

$$\hat{\theta}(X) \quad = \quad \arg\max_\theta P(\theta|X)$$

The *predictive distribution* over new (unseen) data x given the past (previously observed) data X can also be expressed in terms of $P(\theta|X)$:

$$P(x|X) \quad = \quad \int_\Theta P(x|\theta)\, P(\theta|X)\, d\theta.$$

Our interest in $P(X)$ is due to its usefulness in model selection, or more specifically model *order/structure* selection. By this we mean choosing things like the number of clusters in a clustering problem or the number of features and particular feature subset in feature selection, a problem that arrises in many contexts in both supervised and unsupervised

[1] Readers desiring to learn more about this subject can consult one of the many excellent texts. Three used by us are [4, 3, 2]

learning. The set of model structures for a given problem is always countable and frequently finite. For this reason it is often convienient to think of them as being indexed by the integers. The value of $P(X)$ in model selection is its *regularizing* property. By this we refer to the fact that it effects a natural trade-off between the complexity (e.g. the number of clusters) and goodness of fit. When $P(X)$ is considered as a function of X for a fixed model structure it is referred to as a probability, whereas, when it is considered as a function of the model structure for a fixed X, it is called a *likelihood*. Because the parameters θ have been integrated out, $P(X)$ is referred to as the *marginal likelihood* in this context. This connection with the more familiar likelihood function $P(X|\theta)$ would be made clearer by writing $P(X)$ as $P(X|k)$ where k is a variable/parameter representing model structure. We retain the $P(X)$ form, however, to keep the notation simpler. Two examples of the use of $P(X)$ for model-structure selection in unsupervised learning are [7, 8].

The quantity $-\log P(X)$ has been referred to by Rissanen in some work (see [5]) as the *stochastic complexity* of X with respect to the *model class* consisting of all models of a given type/structure (i.e. indexed by all values of θ) and the associated prior $\pi(\theta)$[2].

1.2. Specific Distributions

In the following equations the notation $|\mathbf{\Sigma}|$ denotes the determinant of the matrix $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^T$ denotes its transpose. Vectors are assumed to be single-column matrices and when transposed, therefore, are single-row matrices (e.g. μ and μ^T).

The result we will obtain is a closed form for the following integral

$$P(X) \;=\; \int\int P(X,\mu,\mathbf{\Sigma})\, d\mu\, d\mathbf{\Sigma} \tag{1}$$

where

$$P(X,\mu,\mathbf{\Sigma}) \;=\; P(X|\mu,\mathbf{\Sigma})\,\pi(\mu|\mathbf{\Sigma})\,\pi(\mathbf{\Sigma}) \tag{2}$$

and

$$P(X|\mu,\mathbf{\Sigma}) \;=\; N(X|\mu,\mathbf{\Sigma}) \;\equiv\; (2\pi)^{-nd/2}|\mathbf{\Sigma}|^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu)\right], \tag{3}$$

where $\mu \in \mathbb{R}^d$ is the mean vector and $\mathbf{\Sigma}$ is the symmetric, real-valued $d \times d$ covariance matrix. Its inverse $\mathbf{\Sigma}^{-1}$, which appears frequently, is referred to as the "precision matrix". For the associated priors we use so-called *conjugate* priors for the multi-variate Gaussian. The term "conjugate" refers to priors for which the general functional form (structure) of the *posterior* is the same as that of the prior. An important implication of this is that they produce closed forms when integrated with $P(X|\mu,\mathbf{\Sigma})$. The prior for μ is

$$\pi(\mu|\mathbf{\Sigma}) \;=\; N(\mu|\mu_0,\mathbf{\Sigma}/\beta) \;\equiv\; \left(\frac{\beta}{2\pi}\right)^{d/2}|\mathbf{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\mu-\mu_0)^T\beta\mathbf{\Sigma}^{-1}(\mu-\mu_0)\right]. \tag{4}$$

[2] In more recent work[6] he has used a somewhat different definition of *stochastic complexity*

where μ_0 and β are hyperparameters. The prior for the covariance-matrix parameter $\mathbf{\Sigma}$ is the following *inverse Wishart* distribution[3] $\mathcal{W}^{-1}(\mathbf{\Sigma}|\mathbf{\Psi},\alpha)$.

$$\pi(\mathbf{\Sigma}) \;=\; \mathcal{W}^{-1}(\mathbf{\Sigma}|\mathbf{\Psi},\alpha) \;\equiv\; \frac{|\mathbf{\Psi}|^{\alpha/2}\,|\mathbf{\Sigma}|^{-\frac{1}{2}(\alpha+d+1)}}{2^{\frac{1}{2}\alpha d}\Gamma_d\left(\frac{\alpha}{2}\right)} \exp\left[-\frac{1}{2}\mathrm{trace}(\mathbf{\Psi}\mathbf{\Sigma}^{-1})\right] \tag{5}$$

where $\mathbf{\Psi}$ (a real-valued $d \times d$ matrix) and α (a real-valued scalar) are hyperparameters and

$$\Gamma_d\left(\frac{\alpha}{2}\right) \;\equiv\; \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((\alpha+1-i)/2).$$

2. Program to be followed to Obtain an Expression for $P(X)$

The integral in (1) can be rewritten using (2) as follows:

$$\int\left[\int P(X|\mu,\mathbf{\Sigma})\,\pi(\mu|\mathbf{\Sigma})\,d\mu\right]\pi(\mathbf{\Sigma})\,d\mathbf{\Sigma} \tag{6}$$

We will perform the integration in (6) in two steps. We first perform the inner integral (over μ), which yields $P(X|\mathbf{\Sigma})$, followed by the outer integral (over $\mathbf{\Sigma}$).

3. An Expression for $P(X|\mathbf{\Sigma})$

The goal here is to obtain an expression for:

$$P(X|\mathbf{\Sigma}) \;=\; \int P(X,\mu|\mathbf{\Sigma})d\mu \;=\; \int P(X|\mu,\mathbf{\Sigma})\,\pi(\mu|\mathbf{\Sigma})d\mu, \tag{7}$$

where $P(X|\mu,\mathbf{\Sigma})$ and $\pi(\mu|\mathbf{\Sigma})$ are given by (3) and (4) respectively. Thus:

$$\begin{aligned} P(X|\mu,\mathbf{\Sigma})\,\pi(\mu|\mathbf{\Sigma}) \;&=\; (2\pi)^{-(n+1)d/2}\beta^{d/2}|\mathbf{\Sigma}|^{-(n+1)/2}\;\times \\ &\exp\left\{\left[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu)\right]+\left[-\frac{1}{2}(\mu-\mu_0)^T\beta\mathbf{\Sigma}^{-1}(\mu-\mu_0)\right]\right\} \end{aligned} \tag{8}$$

Next we manipulate the terms in "[...]" in the argument of the exp{...}. Starting with the leftmost (and omitting the $-1/2$ factor for the moment) and expanding we obtain:

$$(\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu) \;=\; \mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i-\mu^T\mathbf{\Sigma}^{-1}\mathbf{x}_i-\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mu+\mu^T\mathbf{\Sigma}^{-1}\mu. \tag{9}$$

Next we sum this over i and use the definition of the sample mean $\bar{\mathbf{x}} \equiv (1/n)\sum_i \mathbf{x}_i$ to obtain:

$$\sum_i(\mathbf{x}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\mu) \;=\; \sum_i \mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x}_i-n\mu^T\mathbf{\Sigma}^{-1}\bar{\mathbf{x}}-n\bar{x}^T\mathbf{\Sigma}^{-1}\mu+n\mu^T\mathbf{\Sigma}^{-1}\mu \tag{10}$$

[3] See Equation (1) on p.268, in Section 7.7.1 "The Inverted Wishart Distribution" of [1].

Next we add and subtract $\bar{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}$ and regroup to obtain:

$$\sum_i(\mathbf{x}_i-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mu) \;=\; \left(\sum_i \mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - n\bar{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right) \;+\; n(\mu-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mu-\bar{\mathbf{x}}). \quad (11)$$

This can be further manipulated, using the following result. This result can be easily seen by expanding the right hand side of (12) and cancelling terms.

$$\sum_i \mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - n\bar{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} \;=\; \sum_i(\mathbf{x}_i-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) \quad (12)$$

It can also be shown that:

$$\sum_i(\mathbf{x}_i-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}) \;=\; n\,\mathrm{trace}[\mathbf{S}\boldsymbol{\Sigma}^{-1}] \;=\; n\,\mathrm{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}], \quad (13)$$

where S is the *sample covariance matrix*[4]. of the data $X \equiv \{\mathbf{x}_i | i = 1, 2, \ldots, n\}$:

$$\mathbf{S} \equiv \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(x_i-\bar{\mathbf{x}})^T.$$

Finally, substituting (13) into (12) and then (12) into (11) yields the following for this first term:

$$\sum_i(\mathbf{x}_i-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mu) \;=\; n\,\mathrm{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}] \;+\; n(\mu-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mu-\bar{\mathbf{x}}) \quad (14)$$

At this point, we have the following for the argument of $\exp\{\ldots\}$ in (8):

$$-\frac{1}{2}\left\{\, n\,\mathrm{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}] \;+\; n(\mu-\bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mu-\bar{\mathbf{x}}) \;+\; (\mu-\mu_0)^T\beta\boldsymbol{\Sigma}^{-1}(\mu-\mu_0)\,\right\} \quad (15)$$

We can now apply the result of Appendix A to combine the two μ-dependent terms. This will give a new Gaussian with covariance matrix

$$\boldsymbol{\Sigma}_\mu \;=\; (n+\beta)^{-1}\boldsymbol{\Sigma}$$

and mean

$$\mu_\mu \;=\; \frac{n\bar{\mathbf{x}} + \beta\mu_0}{n+\beta}$$

.

Using this result, we can rewrite (8) as follows.

$$P(X|\mu,\boldsymbol{\Sigma})\,\pi(\mu|\boldsymbol{\Sigma}) = \left\{\left[(2\pi)^{d/2}(n+\beta)^{-d/2}|\boldsymbol{\Sigma}|^{1/2}\right]N(\mu|\mu_\mu,\boldsymbol{\Sigma}_\mu)\right\}\left[(2\pi)^{-(n+1)d/2}\beta^{d/2}|\boldsymbol{\Sigma}|^{-(n+1)/2}\right] \;\times$$
$$\exp\left\{-\frac{1}{2}\left[n\,\mathrm{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}] + n\bar{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}} + \beta\mu_0^T\boldsymbol{\Sigma}^{-1}\mu_0 - (n\bar{\mathbf{x}}^T+\beta\mu_0^T)\frac{\boldsymbol{\Sigma}^{-1}}{n+\beta}(n\bar{\mathbf{x}}+\beta\mu_0)\right]\right\}, \quad (16)$$

[4]The definition of this varies. For example one frequently sees $\mathbf{S} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(x_i-\bar{\mathbf{x}})^T$.

where $N(\mu_\mu, \mathbf{\Sigma}_\mu)$ is a Gaussian over μ. The term immediately to the left of it in "[...]" is the associated normalizing factor, which we simultaneously multiplied and divided by. The term in "[...]" immediately to the right of $N(\mu_\mu, \mathbf{\Sigma}_\mu)$ is the term multiplying "exp{...}" on the left in (8).

Expanding the argument of exp{...} in (16) and collecting terms yields:

$$-\frac{1}{2}\left[n\,\mathrm{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}] + \left(\frac{n\beta}{n+\beta}\right)(\bar{\mathbf{x}}-\mu_0)^T\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}-\mu_0)\right]$$

Combining the terms multiplying the exp{...} on the left yields:

$$(2\pi)^{-nd/2}\left(\frac{\beta}{n+\beta}\right)^{d/2}|\mathbf{\Sigma}|^{-n/2}$$

Next, integrating over μ results in $N(\mu|\mu_\mu, \mathbf{\Sigma}_\mu) \to 1$, leaving:

$$\begin{aligned} P(X|\mathbf{\Sigma}) &= (2\pi)^{-nd/2}\left(\frac{\beta}{n+\beta}\right)^{d/2}|\mathbf{\Sigma}|^{-n/2} \times \\ &\quad \exp\left\{-\frac{1}{2}\left[n\,\mathrm{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}] + \left(\frac{n\beta}{n+\beta}\right)(\bar{\mathbf{x}}-\mu_0)^T\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}-\mu_0)\right]\right\} \end{aligned} \quad (17)$$

which can also be written as

$$P(X|\mathbf{\Sigma}) = N\left[\bar{\mathbf{x}}\,\middle|\,\mu_0, \left(\frac{n+\beta}{n\beta}\right)\mathbf{\Sigma}\right](2\pi)^{-(n-1)d/2}|\mathbf{\Sigma}|^{-(n-1)/2}n^{-d/2}\exp\left\{-\frac{1}{2}n\,\mathrm{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}]\right\} \quad (18)$$

4. An Expression for $P(X)$

Next we perform the outer integration in (6).

$$P(X) = \int P(X|\mathbf{\Sigma})\,\pi(\mathbf{\Sigma})\,d\mathbf{\Sigma}, \quad (19)$$

where $P(X|\mathbf{\Sigma})$ is given by (17) and the prior on the covariance parameter $\mathbf{\Sigma}$ is given by (5). Multiplying (5) by (17) yields:

$$\begin{aligned} P(X|\mathbf{\Sigma})\,\pi(\mathbf{\Sigma}) &= \left(\tfrac{\beta}{n+\beta}\right)^{d/2}|\mathbf{\Psi}|^{\alpha/2}|\mathbf{\Sigma}|^{-\frac{1}{2}(\alpha+n+d+1)}\left[(2\pi)^{nd/2}2^{\frac{1}{2}\alpha d}\Gamma_d\left(\tfrac{\alpha}{2}\right)\right]^{-1} \times \\ &\quad \exp\left\{-\tfrac{1}{2}\left[n\,\mathrm{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}] + \mathrm{trace}(\mathbf{\Psi}\mathbf{\Sigma}^{-1}) + \left(\tfrac{n\beta}{n+\beta}\right)(\bar{\mathbf{x}}-\mu_0)^T\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}-\mu_0)\right]\right\} \end{aligned} \quad (20)$$

This can be manipulated into the following form

$$\begin{aligned} P(X|\mathbf{\Sigma})\,\pi(\mathbf{\Sigma}) &= \pi^{-\frac{1}{2}nd}\left(\tfrac{\beta}{n+\beta}\right)^{d/2}|\mathbf{\Psi}|^{\alpha/2}\left[\frac{\Gamma_d\left(\frac{\alpha+n}{2}\right)}{\Gamma_d\left(\frac{\alpha}{2}\right)}\right] \times \\ &\quad \left|\mathbf{\Psi} + nS + \tfrac{n\beta}{n+\beta}(\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)^T\right|^{-\frac{1}{2}(n+\alpha)} \times \\ &\quad \mathcal{W}^{-1}\left(\mathbf{\Sigma}\,\middle|\,\mathbf{\Psi} + nS + \tfrac{n\beta}{n+\beta}(\bar{\mathbf{x}}-\mu_0)(\bar{\mathbf{x}}-\mu_0)^T,\, n+\alpha\right) \end{aligned} \quad (21)$$

When we integrate this over $\mathbf{\Sigma}$, the $\mathcal{W}^{-1}$ term integrates to 1 leaving:

$$P(X) = \pi^{-\frac{1}{2}nd} \left(\frac{\beta}{n+\beta}\right)^{d/2} |\mathbf{\Psi}|^{\alpha/2} \left[\frac{\Gamma_d\left(\frac{\alpha+n}{2}\right)}{\Gamma_d\left(\frac{\alpha}{2}\right)}\right] \left|\mathbf{\Psi} + nS + \frac{n\beta}{n+\beta}(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)^T\right|^{-\frac{1}{2}(n+\alpha)}, \tag{22}$$

which is the result we seek.

Acknowledgement: We thank Shivakumar Vaithyanathan for helpful discussions and for checking some of the mathematics. One of us (Dom) also thanks David Heckerman and Steffen Lauritzen for an interesting and useful discussion on the usage of the term "likelihood".

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York, second edition, 1984.

[2] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York; Berlin, 1985. ISBN: 0-387-96098-8.

[3] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley, New York, 1994. ISBN 0-471-92416-4.

[4] Peter M. Lee. *Bayesian Statistics: An Introduction.* Arnold, second edition, 1997. ISBN: 0-471-19481-6.

[5] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, 1989.

[6] J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans Inf Th*, 42(1):40–47, January 1996.

[7] Shivakumar Vaithyanathan and Byron Dom. Model selection in unsupervised learning with applications to document clustering. In I. Brakto and S. Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning*, Bled, Slovenia, June 1999. Morgan Kaufman.

[8] Shivakumar Vaithyanathan and Byron Dom. Hierarchical unsupervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford University; Stanford, CA, June 2000.

A. The product of two multivariate Gaussians

Here we obtain an expression for the product of two multivariate Gaussians with mean vectors μ_1 and μ_2 and precision matrices P_1 and P_2. First we combine the two exponential arguments by completing the square:

$$(\mathbf{x}-\mu_1)^T P_1(\mathbf{x}-\mu_1) \;+\; (\mathbf{x}-\mu_2)^T P_2(\mathbf{x}-\mu_2) \tag{23}$$

Expand this:

$$\mathbf{x}^T(P_1+P_2)\mathbf{x} \;-\; (\mu_1^T P_1+\mu_2^T P_2)\mathbf{x} \;-\; \mathbf{x}^T(P_1\mu_1+P_2\mu_2) \;+\; (\mu_1^T P_1\mu_1+\mu_2^T P_1\mu_2)$$

This suggests a new Gaussian with precision matrix $P_{12}=P_1+P_2$. Following that thought, the first three terms (the ones involving $\mathbf{x}$) can be rewritten as:

$$\mathbf{x}^T(P_1+P_2)\mathbf{x} - (\mu_1^T P_1+\mu_2^T P_2)(P_1+P_2)^{-1}(P_1+P_2)\mathbf{x} - \mathbf{x}^T(P_1+P_2)(P_1+P_2)^{-1}(P_1\mu_1+P_2\mu_2)$$

This, in turn, suggests that the mean vector of the new Gaussian will be

$$\mu_{12} \;=\; (P_1+P_2)^{-1}(P_1\mu_1+P_2\mu_2)$$

Using these values for μ_1 and P_{12}, to have an expression for

$$(\mathbf{x}-\mu_{12})^T P_{12}(\mathbf{x}-\mu_{12})$$

we need a constant (non-$\mathbf{x}$-dependent) term of:

$$\begin{aligned}\mu_{12}^T P_{12}\mu_{12} \;=\;& (\mu_1^T P_1+\mu_2^T P_2)(P_1+P_2)^{-1}(P_1+P_2)(P_1+P_2)^{-1}(P_1\mu_1+P_2\mu_2)\\ & (\mu_1^T P_1+\mu_2^T P_2)(P_1+P_2)^{-1}(P_1\mu_1+P_2\mu_2)\end{aligned}$$

We can now rewrite (23) as:

$$(\mathbf{x}-\mu_{12})^T P_{12}(\mathbf{x}-\mu_{12}) \;+\; (\mu_1^T P_1\mu_1+\mu_2^T P_1\mu_2) \;-\; \mu_{12}^T P_{12}\mu_{12}.$$

The complete product of the two Gaussians is:

$$\begin{aligned}&\left|\frac{P_{12}}{2\pi}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_{12})^T P_{12}(\mathbf{x}-\mu_{12})\right] \;\times\\ &\quad\left\{\left|\frac{2\pi}{P_{12}}\right|^{\frac{1}{2}}\left|\frac{P_1}{2\pi}\right|^{\frac{1}{2}}\left|\frac{P_2}{2\pi}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\mu_1^T P_1\mu_1+\mu_2^T P_1\mu_2-\mu_{12}^T P_{12}\mu_{12})\right]\right\}.\end{aligned}$$