Analysing gene expression data using Gaussian Processes

Lorenz Wernisch

Complex gene regulatory mechanisms ensure the proper functioning of biological cells. New high-throughput experimental techniques, such as microarrays, provide a snapshot of gene expression levels of thousands of genes at the same time. If repeated on a sample of synchronized cells, time-series profiles of gene activity can be obtained. The aim is to reconstruct the complex gene regulatory network underlying these profiles. Genes often influence each other in a nonlinear fashion and with intricate interaction patterns. Linear models are often unsuited to capture such relationships. Gaussian processes (GPs), on the other hand, are ideal for representing nonlinear relationships. A particular attraction is the automatic relevance determination (ARD) effect, removing unused inputs and resulting in sparse gene networks.

Prediction with Gaussian processes

For the purpose of regressing gene expression data, a combination of a linear covariance part with a squared exponential proves useful. Gene regulatory functional relationships rarely show sharp jumps and the smoothness of squared exponentials is usually not a problem. Especially when working with logarithmic values many relationships are actually almost linear and the inclusion of a linear part seems advisable. An automatic relevance determination (ARD) procedure is then able to select the simpler linear regression if the nonlinear component is neglectable.

More specifically, for each of the d-dimensional input values $x = (x_1, \ldots, x_N)$ the output value is $t_i = f(x_i)$. The joint distribution of the output $t = (t_1, \ldots, t_N)'$ is a multivariate Gaussian $N(0, K)$, where K is given by

$$K_{pq} = \beta_0 + C_L(x_p, x_q) + C_G(x_p, x_q) + \sigma_\epsilon^2 I(p = q) \tag{1}$$

Here the linear covariance part is

$$C_L(x_p, x_q) = x_p' B^{-1} x_q$$

with the diagonal matrix $B = \mathrm{diag}(\beta_1, \ldots, \beta_d)$ and the squared exponential (Gaussian) covariance part is

$$C_G(x_p, x_q) = \alpha_0 \exp(-\frac{1}{2}(x_p - x_q)' A^{-1} (x_p - x_q))$$

with the diagonal matrix $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$. Once training inputs $x_1, \ldots, x_N$ with known target values t_i are given the output distribution of $f(x^*)$ at

a new input point x^* can be calculated from the joint covariance function

$$\tilde{K} = \begin{pmatrix} K & k(x^*) \\ k(x^*)' & k(x^*, x^*) \end{pmatrix}$$

for all inputs, where

$$k(x^*) = (\beta_0 + C_L(x^*, x_q) + C_G(x^*, x_q))_{q=1}^N$$

and

$$k(x^*, x^*) = \beta_0 + x^{*\prime} B^{-1} x^* + \alpha_0 + \sigma_\epsilon^2$$

The conditional distribution of $f(x^*)$ is Gaussian $N(\mu(x^*), \sigma^2(x^*))$ with

$$\begin{aligned} \mu(x^*) &= k(x^*)' K^{-1} t \\ \sigma^2(x^*) &= k(x^*, x^*) - k(x^*)' K^{-1} k(x^*) \end{aligned} \tag{2}$$

A GP is specified by the parameters $\theta = (\beta_0, \beta_1, \ldots, \beta_d, \alpha_0, \alpha_1, \ldots, \alpha_d, \sigma_\epsilon^2)$. Fitting θ for given input data x to output values $t = (t_1, \ldots, t_N)'$ is achieved by optimising the log likelihood

$$\log p(t \mid x, \theta) = -\frac{1}{2}\big(t' K(x, \theta) t - \log |K(x, \theta)| - n \log 2\pi$$

where $K(x, \theta)$ is given in (1). Using partial derivatives we find that a conjugate gradient method is very efficient in learning GPs for the problems below.

Optimizing parameters $\beta_1, \ldots, \beta_d$ for the linear covariance part amounts to a likelihood type II optimization of the precision of regression factors with Gaussian priors when marginalising them out. This has the effect of removing unused linear input dimensions. Similarly, relevance parameters $\alpha_1, \ldots, \alpha_d$ going to 0 indicate that an input is not needed in the nonlinear part. We will make use of this effect in the reconstruction of gene networks below.

GPs for time-series data

Time-series data consist of T values $x_1, \ldots, x_T$ with $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})$ d-dimensional time sections. We assume that they are generated by a process

$$x_t = f(x_{t-1}) + \epsilon, \quad t = 2, \ldots, T$$

where $f = (f^{(1)}, \ldots, f^{(d)})$ is a (nonlinear) smooth function vector and $\epsilon \sim N(0, \sigma_\epsilon^2)$ a Gaussian noise. We will represent each regression function $f^{(i)}$ by a GP with parameters $\theta^{(i)}$.

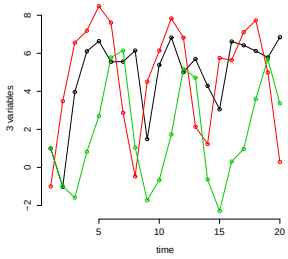
In gene regulatory networks a variable usually depends on very few other variables. Optimizing hyperparameters $\theta^{(i)}$ in a GP when regressing a variable $x^{(i)}$ on $x^{(1)}, \ldots, x^{(d)}$ amounts to determining the relevant input variables

or input genes in our case. The expectation is that the ARD results in a sparse network. In the following we chose lognormal priors on all the parameters θ. The choise of expectation and variance of the lognormal priors has some influence on the success of the regression and usually need some adjustment by hand.

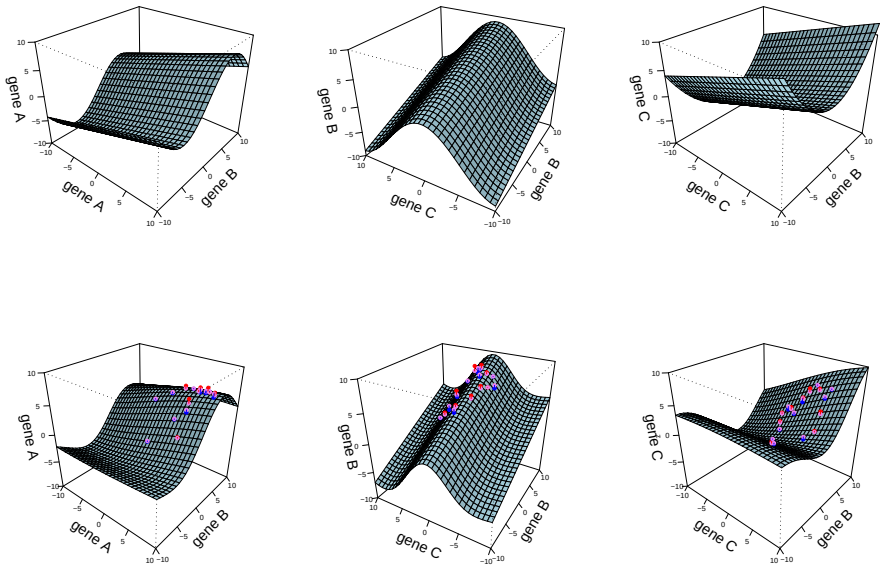
Simulated data

We simulated data from an artificial dynamic network on 3 variables connected by mixed linear and nonlinear relationships ($\epsilon \sim N(0, 0.5)$, start values (1,-1,1)). Also shown are the time profiles of the 3 variables.

$$x_{t+1} = 0.35x_t + 5\sin(0.3y_t) + \epsilon_1$$
$$y_{t+1} = 0.4y_t + 5\cos(0.3z) + \epsilon_2$$
$$z_{t+1} = 0.35z_t + 5\sin(0.3y_t) + \epsilon_3$$



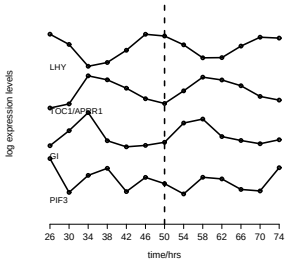
The following plots compare the true functions (top row) with the reconstruction by GPs (bottom row) for the 3 variables:

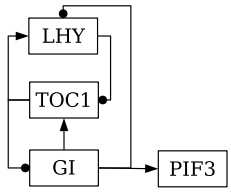


Also ARD removed unused inputs successfully and the corresponding hyperparameters were all 0.

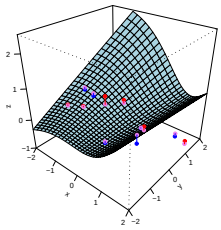
Circadian clock in *Arabidopsis thaliana*

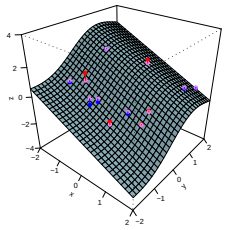
The daily up- and down-regulation of genes in the plant *A. thaliana* is entrained by rhythmic light conditions but continues after light is held constant. A core set of genes, in particular LHY and TOC1 are responsible for sustaining oscillation under constant light conditions. Time-series data of gene expression every 4 hours over two days have been obtained under such conditions[1]. The profiles for some core genes are shown in the following figure on the left hand side.





The right hand side shows the network as inferred by using GPs. Inhibition (bullets) or activation (arrows) are inferred from mean slopes when fitting the linear part only. A comparatively sparse network that agrees with what is believed to be the correct network. In particular, the two interlinked feedback loops are thought to drive the circadian clock. The next plots show some typical nonlinear dependencies for genes LHY and GI as reconstructed by the GPs:





GPs for state space models

Often not all relevant variables are observable. Profiles of hidden variables can in principle be reconstructed by applying the above estimation of GPs in an iterated fashion to a state space inferred by an extended Kalman filter (EKF) approach. We show in the appendix how to calculate expectation and variance of a multivariate GP given an uncertain input. This can be used in a straightforward manner by a standard EKF to filter hidden states assuming nonlinear transitions. Experiments on simulated data show, however, that careful initialization (for example, by a factor analysis) is crucial.

[1]Data provide by Kieron Edwards and Andrew Millar, University of Edinburgh

Appendix

A Prediction with uncertain input

Quiñonero-Candela et al. [2003] show how to calculate a Gaussian approximation of a Gaussian process with uncertain input. The sections below extend their approach to the case of a Gaussian process combining a Gaussian term with a linear term and to the multivariate case where several Gaussian processes are used in parallel.

A Gaussian process is specified by the parameters $\theta = (\beta_0, \beta_1, \ldots, \beta_d)$, $\alpha_0, \alpha_1, \ldots, \alpha_d, \sigma_\epsilon^2)$ and the N input vectors and target values as above, denote these data by D. Assume now that the input variable x^* is uncertain with a Gaussian distribution $N(u, S)$. The predictive distribution for $t^* = f(x^*)$

$$p(t^* \mid u, S, D) = \int p(t^* \mid x^*, D) p_G(x^* \mid u, S)\, dx^*$$

is analytically intractable. Here

$$p(t^* \mid x^*, D) = \frac{1}{\sigma(x^*)\sqrt{2\pi}} \exp(-\frac{(t^* - \mu(x^*))^2}{2\sigma^2(x^*)})$$

according to the Gaussian process defined by D via relations (2). What we can calculate exactly though is the mean and variance of the resulting distribution. This is enough, for example, to define an exact version of an extended Kalman filter.

In the following we also assume that we not only have one Gaussian process but two. Given are N_1 inputs $x_{1,1}, \ldots, x_{N_1,1}$ for the first and N_2 inputs $x_{1,2}, \ldots, x_{N_2,2}$ for the second process. Target vectors are $t^{(1)}$ and $t^{(2)}$, and parameters θ_1 and θ_2. The corresponding predictive distribution of $\tilde{t} = ({t^{(1)}}', {t^{(2)}}')'$ is a bivariate Gaussian with mean vector $\tilde{\mu}(x) = (\mu_1(x)', \mu_2(x)')'$ and covariance

$$\tilde{\Sigma}(x^*) = \begin{pmatrix} \sigma_1^2(x^*) & 0 \\ 0 & \sigma_2^2(x^*) \end{pmatrix}$$

The covariance of 0 is a reflection of the independence of the underlying Gaussian processes. We assume the probability distributions over the function spaces of the two processes are independent. The two Gaussian process are each defined by their individual parameters $\theta_1 = (\beta_{0,1}, B_1, \alpha_{0,1}, A_1, \sigma_{\epsilon,1})$ and $\theta_2 = (\beta_{0,2}, B_2, \alpha_{0,2}, A_2, \sigma_{\epsilon,2})$.

The Gaussian approximation is based on the laws of iterated expectation and variance:

$$\begin{aligned} E(\tilde{t}^*) &= E_{x^*}(E(\tilde{t}^* \mid x^*)) = E_{x^*}(\tilde{\mu}(x^*)) \\ \mathrm{var}(\tilde{t}^*) &= E_{x^*}(\mathrm{var}(\tilde{t}^* \mid x^*)) + \mathrm{var}_{x^*}(E(\tilde{t}^* \mid x^*)) = E_{x^*}(\tilde{\Sigma}(x^*)) + \mathrm{var}_{x^*}(\tilde{\mu}(x^*)) \end{aligned} \tag{3}$$

A.1 Computing the mean

Setting $\gamma = K^{-1}t$ the mean of a Gaussian process is

$$\mu(x) = \sum_j \gamma_j \left(\beta_0 + x' B^{-1} x_j + C_G(x, x_j)\right)$$

Consequently, the expectation of the mean for uncertain $x^* \sim N(u, S)$ is calculated as

$$E_{x^*}(\mu(x^*)) = \sum_j \gamma_j \left(\beta_0 + u' B^{-1} x_j + \int C_G(x^*, x_j)\, p_G(x^* \mid u, S)\, dx^*\right) \quad (4)$$

Fortunately, due to the choice of $C_G(x^*, x_j)$, the latter integral is easily solved analytically using identity (21) for the combination of Gaussians

$$\begin{aligned} l_j &= \int C_G(x^*, x_j)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_0 (2\pi)^{d/2} |A|^{1/2} \int p_G(x^* \mid x_j, A)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_0 |A^{-1}S + I|^{-1/2} \exp\left(-\frac{1}{2}(u - x_j)'(A + S)^{-1}(u - x_j)\right) \end{aligned} \quad (5)$$

Since A is diagonal, it might be advantageous to write $(A+S)^{-1} = A^{-1}(SA^{-1}+I)^{-1}$ and use, say, a Cholesky decomposition of $(SA^{-1} + I)$ for both, the inverse and the determinant.

A.2 Computing the variance, first component

The first component $E_{x^*}(\tilde{\Sigma}(x^*))$ of the variance in equation (3) is the expectation of

$$\begin{aligned} \sigma^2(x^*) &= k(x^*, x^*) - k(x^*)' K^{-1} k(x^*) \\ &= \beta_0 + \alpha_0 + x^{*\prime} B^{-1} x^* + \sigma_\epsilon^2 \\ &\quad - \sum_{ij} (\beta_0 + x^{*\prime} B^{-1} x_i + C_G(x^*, x_i)) K_{ij}^{-1} (\beta_0 + x^{*\prime} B^{-1} x_j + C_G(x^*, x_j)) \end{aligned} \quad (6)$$

Hence,

$$\begin{aligned} E_{x^*}(\sigma^2(x^*)) &= \int \sigma^2(x^*)\, p_G(x^* \mid u, S)\, dx^* \\ &= \beta_0 + \alpha_0 + \sigma_\epsilon^2 + \mathrm{tr}(B^{-1}(S + uu')) - \sum_{ij} K_{ij}^{-1} L_{ij} \end{aligned} \quad (7)$$

where

$$L_{ij} = E_{x^*}((\beta_0 + x^{*\prime} B^{-1} x_i + C_G(x^*, x_i))(\beta_0 + x^{*\prime} B^{-1} x_j + C_G(x^*, x_j))) \quad (8)$$

We used that $x'B^{-1}x = \mathrm{tr}(x'B^{-1}x) = \mathrm{tr}(B^{-1}xx')$ and that $E_{x^*}(x^*x^{*\prime}) = S+uu'$. Since B is diagonal we simply have $\mathrm{tr}(B^{-1}(S+uu')) = \sum_i \beta_i^{-1}(S_{ii}+u_i^2)$. Before we can calculate L_{ij} we need a couple of integrals.

$$\begin{aligned} E_j &= \int x^* C_G(x^*, x_j)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_0 (2\pi)^{d/2} |A|^{1/2} \int x^* p_G(x^* \mid x_j, A)\, p_G(x^* \mid u, S)\, dx^* \\ &= l_j (A^{-1} + S^{-1})^{-1} (A^{-1}x_j + S^{-1}u) \end{aligned} \tag{9}$$

If we substitute A for both, A and B, x_i, x_j for a, b in (17) and (21) and if we further substitute $A/2$ for D, $x_d = (x_i + x_j)/2$ for d, S for C and u for c in (22) and use $|A||2A|^{-1/2}|A/2 + S|^{-1/2} = |2A^{-1}S + I|^{-1/2}$ we obtain

$$\begin{aligned} I_{ij} &= \int C_G(x^*, x_i)\, C_G(x^*, x_j)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_0^2 (2\pi)^d |A| \int p_G(x^* \mid x_i, A)\, p_G(x^* \mid x_j, A)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_0^2 |2A^{-1}S + I|^{-1/2} \\ &\quad \exp\left(-\frac{1}{2}(x_d - u)'(A/2 + S)^{-1}(x_d - u) - \frac{1}{2}(x_i - x_j)'(2A)^{-1}(x_i - x_j) \right) \end{aligned} \tag{10}$$

We can now finish (7) by specifying L_{ij} (using $(x'B^{-1}x_i)'x'B^{-1}x_i = x_i'B^{-1}xx'B^{-1}x_j$).

$$\begin{aligned} L_{ij} = \beta_0^2 &+ \beta_0\, u'B^{-1}(x_i + x_j) + x_i'B^{-1}(S + uu')B^{-1}x_j \\ &+ \beta_0(l_i + l_j) + x_i B^{-1} E_j + x_j B^{-1} E_i + I_{ij} \end{aligned} \tag{11}$$

A.3 Computing the variance, second component

The second component $\mathrm{var}_{x^*}(\tilde{\mu}(x^*))$ of the variance can be split as

$$\mathrm{var}_{x^*}(\tilde{\mu}(x^*)) = E_{x^*}(\tilde{\mu}(x^*)\tilde{\mu}(x^*)') - E_{x^*}(\tilde{\mu}(x^*))E_{x^*}(\tilde{\mu}(x^*))'$$

We already calculated $E_{x^*}(\tilde{\mu}(x^*))$ in section A.1. Continuing with the notation of that section and using results from the previous section A.2, we have

$$E_{x^*}(\mu(x^*)^2) = \sum_{ij} \gamma_i \gamma_j L_{ij}$$

What is left is the expectation of the covariance between $\mu_1(x^*)$ and $\mu_2(x^*)$

$$\mu_1(x)\mu_2(x) = \sum_{ij} \gamma_{i,1}\gamma_{j,2}\, (\beta_{0,1} + x'B_1^{-1}x_{i,1} + C_{G1}(x, x_{i,1}))(\beta_{0,2} + x'B_2^{-1}x_{j,2} + C_{G2}(x, x_{j,2})) \tag{12}$$

that is

$$E_{x^*}(\mu_1(x)\mu_2(x)) = \sum_{ij} \gamma_{i,1}\gamma_{j,2}\, L_{ij}^{(12)} \tag{13}$$

Similarly as in section A.2 it is convenient to calculate a couple of integrals first. We denote with $E_{i,1}$ and $E_{j,2}$ versions of (9), and with $l_{j,1}$ and $l_{j,2}$ versions of (5) for the respective Gaussian processes. Since A_1 and A_2 are diagonal matrices it is computationally advantagous to combine them into a diagonal matrix D as in (17):

$$\begin{aligned} D^{(12)} &= (A_1^{-1} + A_2^{-1})^{-1} \\ d_{ij}^{(12)} &= D^{(12)}(A_1^{-1}x_{i,1} + A_2^{-1}x_{j,2}) \end{aligned} \tag{14}$$

If we now use (21) involving $\tilde{z}_H$ and substitute S for C and u for c in (22) and observe

$$|A_1|^{1/2}|A_2|^{1/2}|A_1 + A_2|^{-1/2}||D^{(12)} + S|^{-1/2} = |(D^{(12)})^{-1}S + I|^{-1/2}$$

we obtain

$$\begin{aligned} I_{ij}^{(12)} &= \int C_{G1}(x^*, x_{i,1})\, C_{G2}(x^*, x_{j,2})\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_{0,1}\alpha_{0,2}(2\pi)^d|A_1|^{1/2}|A_2|^{1/2} \int p_G(x^* \mid x_{i,1}, A_1)\, p_G(x^* \mid x_{j,2}, A_2)\, p_G(x^* \mid u, S)\, dx^* \\ &= \alpha_{0,1}\alpha_{0,2}\, |(A_1^{-1} + A_2^{-1})S + I|^{-1/2} \\ &\quad \exp\left(-\frac{1}{2}(d_{ij}^{(12)} - u)'(D^{(12)} + S)^{-1}(d_{ij}^{(12)} - u) - \frac{1}{2}(x_{i,1} - x_{j,2})'(A_1 + A_2)^{-1}(x_{i,1} - x_{j,2})\right) \end{aligned} \tag{15}$$

We can now finish (13) by specifying $L_{ij}^{(12)}$.

$$\begin{aligned} L_{ij}^{(12)} &= \beta_{0,1}\beta_{0,2} + \beta_{0,2}u'B_1^{-1}x_{i,1} + \beta_{0,1}u'B_2^{-1}x_{j,2} + x_{i,1}'B_1^{-1}(S + uu')B_2^{-1}x_{j,2} \\ &\quad + \beta_{0,1}l_{j,2} + \beta_{0,2}l_{i,1} + x_{i,1}B_1^{-1}E_{j,2} + x_{j,2}B_2^{-1}E_{i,1} + I_{ij}^{(12)} \end{aligned} \tag{16}$$

B Integrating products of Gaussians

Following Quiñonero-Candela et al. [2003] we note that two Gaussians $N(a, A)$ and $N(b, B)$ can be combined as follows ($p_G(x \mid a, B)$ a Gaussian density on x with mean a and covariance matrix B):

$$\begin{aligned} p_G(x \mid a, A)\, p_G(x \mid b, B) &= z_D\, p_G(x \mid d, D) \\ D &= (A^{-1} + B^{-1})^{-1} \\ d &= D(A^{-1}a + B^{-1}b) \\ z_D &= \frac{|D|^{1/2}}{(2\pi)^{d/2}|A|^{1/2}|B|^{1/2}} \exp\left(-\frac{1}{2}(a'A^{-1}a + b'B^{-1}b - d'D^{-1}d)\right) \end{aligned} \tag{17}$$

Consequently, the integral of two Gaussians is

$$\int p_G(x \mid a, A)\, p_G(x \mid b, B)\, dx = z_D$$

The normalizing constant z_D can be simplified using the following matrix identities (variants of Woodbury's inversion formula)

$$\begin{aligned}
(A+B)^{-1} &= A^{-1} - A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = A^{-1} - A^{-1}DA^{-1} \\
&= B^{-1} + B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1} = B^{-1} + B^{-1}DB^{-1} \\
(A+B)^{-1} &= A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1} = A^{-1}DB^{-1} \\
&= B^{-1}(A^{-1}+B^{-1})^{-1}A^{-1} = B^{-1}DA^{-1}
\end{aligned} \tag{18}$$

We then have

$$|D|^{1/2}|A|^{-1/2}|B|^{-1/2} = |A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}|^{1/2} = |A+B|^{-1/2} \tag{19}$$

and (noting that covariance matrices are symmetric)

$$\begin{aligned}
-d'D^{-1}d &= -(A^{-1}a + B^{-1}b)'DD^{-1}D(A^{-1}a + B^{-1}b) \\
&= -a'A^{-1}DA^{-1}a - b'B^{-1}DB^{-1}b - a'A^{-1}DB^{-1}b - b'B^{-1}DA^{-1}a \\
&= a'(A+B)^{-1}a - a'A^{-1}a + b'(A+B)^{-1}b - b'B^{-1}b \\
&\quad - a'(A+B)^{-1}b - b'(A+B)^{-1}a \\
&= (a-b)'(A+B)^{-1}(a-b) - a'A^{-1}a - b'B^{-1}b
\end{aligned} \tag{20}$$

and the following simplification of the above integral

$$\begin{aligned}
z_D &= \int p_G(x \mid a, A)\, p_G(x \mid b, B)\, dx \\
&= (2\pi)^{-d/2}|A+B|^{-1/2} \exp\left(-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)\right)
\end{aligned} \tag{21}$$

Extending the result to the product of three Gaussians $N(a, A)$, $N(b, B)$

and $N(c, C)$ is straighforward:

$$
\begin{aligned}
p_G(x \mid a, A)\, p_G(x \mid b, B)\, p_G(x \mid c, C) &= z_D\, p_G(x \mid d, D) p_G(x \mid c, C) \\
&= z_D \tilde{z}_H p_G(x \mid h, H) = z_H p_G(x \mid h, H) \\
H &= (D^{-1} + C^{-1})^{-1} = (A^{-1} + B^{-1} + C^{-1})^{-1} \\
h &= H(D^{-1}d + C^{-1}c) = H(A^{-1}a + B^{-1}b + C^{-1}c) \\
\tilde{z}_H &= \frac{|H|^{1/2}}{(2\pi)^{d/2}|D|^{1/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(d'D^{-1}d + c'C^{-1}c - h'H^{-1}h)\right) \\
&= (2\pi)^{-d/2}|D + C|^{-1/2} \exp\left(-\frac{1}{2}(d - c)'(D + C)^{-1}(d - c)\right) \\
z_H &= \frac{|H|^{1/2}}{(2\pi)^{d}|A|^{1/2}|B|^{1/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(a'A^{-1}a + b'B^{-1}b + c'C^{-1}c - h'H^{-1}h)\right)
\end{aligned}
\tag{22}
$$

References

J. Quiñonero-Candela, A. Girard, and C. E. Rassmussen. Prediction at an uncertain input for gaussian processes and relevance vector machines - application to time-series forecasting. Technical report, Informatics and Mathematical Modelling, Technical Univesity of Denmark, 2003.