On Gaussian Expectation Propagation

Ralf Herbrich

9th July 2005

Abstract

In this short note we will re-derive the Gaussian expectation propagation (Gaussian EP) algorithm as presented in Minka (2001) and demonstrate an application of Gaussian EP to approximate multi-dimensional truncated Gaussians.

1 On Gaussian Distributions

Here we will summarise some important equalities about the Gaussian density. A Gaussian density in $\mathbb{R}^n$ is defined by

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) := (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.1}$$

We will write $\mathbf{x} \sim \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ to both denote $\mathbf{x}$ has a distribution $P(\mathbf{x})$ and that the density of this distribution is given by (1.1). We will write $\mathcal{N}(\mathbf{x})$ as a shorthand for $\mathcal{N}(\mathbf{x};\mathbf{0},\mathbf{I})$. For $t \in \mathbb{R}$, we will denote the cumulative Gaussian distribution function by $\Phi(t;\mu,\sigma^2)$ which is defined by

$$\Phi\left(t;\mu,\sigma^2\right) = P_{x\sim\mathcal{N}(x;\mu,\sigma^2)}(x \leq t) = \int_{-\infty}^{t} \mathcal{N}\left(x;\mu,\sigma^2\right) dx\,. \tag{1.2}$$

Again, we write $\Phi(t)$ as a shorthand for $\Phi(t;0,1)$. We will write $\langle f(x)\rangle_{x\sim P}$ to denote the expectation of f over the random draw of x, that is $\langle f(x)\rangle_{x\sim P} := \int f(x)dP(x)$. The following results are given without proof; for a detailed derivation the reader is referred to Herbrich (2002).

Linear transformation

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \text{ and } \mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{b} \Rightarrow \mathbf{y} \sim \mathcal{N}\left(\mathbf{y};\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right).$$

Convolutions Assume

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \text{ and } \mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{y};\mathbf{A}\mathbf{x},\boldsymbol{\Gamma}).$$

Then

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\left(\mathbf{x};\boldsymbol{\Psi}\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right),\boldsymbol{\Psi}\right), \tag{1.3}$$

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{x};\mathbf{A}\boldsymbol{\mu},\boldsymbol{\Gamma}+\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right), \tag{1.4}$$

where $\boldsymbol{\Psi} := (\mathbf{A}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{A}+\boldsymbol{\Sigma}^{-1})^{-1}$.

Marginals Let us assume that a random vector $\mathbf{x}$ is composed such that

$$\mathbf{x} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{12}^{\mathrm{T}} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right).$$

Then we know

$$\begin{aligned}
\mathbf{x}_1 &\sim \mathcal{N}(\mathbf{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \,, \\
\mathbf{x}_2 &\sim \mathcal{N}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \,, \\
\mathbf{x}_1|\mathbf{x}_2 &\sim \mathcal{N}\left(\mathbf{x}_1; \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\mathrm{T}}\right) \,, \\
\mathbf{x}_2|\mathbf{x}_1 &\sim \mathcal{N}\left(\mathbf{x}_2; \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right) \,.
\end{aligned}$$

Rescaling and Symmetry

$$\begin{aligned}
\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot \mathcal{N}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})\right) \\
\Phi\left(t; \mu, \sigma^2\right) &= \Phi\left(\frac{t - \mu}{\sigma}\right) \,, \\
\Phi(t) &= 1 - \Phi(-t) \,, \\
\Phi\left(t; \mu, \sigma^2\right) &= 1 - \Phi\left(-t; -\mu, \sigma^2\right) \,.
\end{aligned}$$

2 Gaussian Density Filtering

Let us assume that we have a Gaussian belief in some parameter $\boldsymbol{\theta}$, $P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and that we are given a likelihood $P(\mathbf{x}|\boldsymbol{\theta})$ which we will view as a function $t_{\mathbf{x}}(\boldsymbol{\theta})$ of the parameter only. Then, in general, the posterior $P(\boldsymbol{\theta}|\mathbf{x})$ is no longer a Gaussian distribution,

$$P(\boldsymbol{\theta}|\mathbf{x}) = \frac{t_{\mathbf{x}}(\boldsymbol{\theta})\, P(\boldsymbol{\theta})}{\int t_{\mathbf{x}}(\tilde{\boldsymbol{\theta}}) P(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}} \,.$$

However, under certain conditions on the function $t_{\mathbf{x}}$ we can efficiently find the Gaussian approximation, $\mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\mu}}_{\mathbf{x}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{x}})$, which minimises the Kullback–Leibler divergence between the true posterior, $P(\boldsymbol{\theta}|\mathbf{x})$, and the itself. This approach is called *Gaussian density filtering (GDF)* and is a special case of the *assumed density filtering (ADF)* approach. Note that the subscript $\mathbf{x}$ indicates that the approximation is optimal for the given $\mathbf{x}$. It can be shown that

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{g}_{\mathbf{x}} \,, \qquad \hat{\boldsymbol{\Sigma}}_{\mathbf{x}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\left(\mathbf{g}_{\mathbf{x}}\mathbf{g}_{\mathbf{x}}^{\mathrm{T}} - 2\mathbf{G}_{\mathbf{x}}\right)\boldsymbol{\Sigma} \,, \tag{2.1}$$

where the vector $\mathbf{g}_{\mathbf{x}}$ and the matrix $\mathbf{G}_{\mathbf{x}}$ are given by

$$\mathbf{g}_{\mathbf{x}} := \left.\frac{\partial \log\left(Z_{\mathbf{x}}\left(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right)\right)}{\partial \tilde{\boldsymbol{\mu}}}\right|_{\tilde{\boldsymbol{\mu}}=\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}} \,, \qquad \mathbf{G}_{\mathbf{x}} := \left.\frac{\partial \log\left(Z_{\mathbf{x}}\left(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right)\right)}{\partial \tilde{\boldsymbol{\Sigma}}}\right|_{\tilde{\boldsymbol{\mu}}=\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}} \,,$$

and the function $Z_{\mathbf{x}}$ is defined by

$$Z_{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \int t_{\mathbf{x}}(\boldsymbol{\theta}) \cdot \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\theta} = P(\mathbf{x}) \,.$$

3 Gaussian Expectation Propagation

Similar to the last section, let us assume that we have a Gaussian belief in some parameter $\boldsymbol{\theta}$, $P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and that we are given a likelihood $P(\mathbf{x}|\boldsymbol{\theta})$ which has now m factors, that is

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{m} t_{i,\mathbf{x}}(\boldsymbol{\theta}) \,.$$

Then, in general, the posterior $P(\boldsymbol{\theta}|\mathbf{x})$ is no longer a Gaussian distribution,

$$P(\boldsymbol{\theta}|\mathbf{x}) = \frac{\prod_{i=1}^{m} t_{i,\mathbf{x}}(\boldsymbol{\theta}) \cdot \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\underbrace{\int \prod_{i=1}^{m} t_{i,\mathbf{x}}(\tilde{\boldsymbol{\theta}}) \cdot \mathcal{N}\left(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\tilde{\boldsymbol{\theta}}}_{Z_{\mathbf{x}} = P(\mathbf{x})}}$$

Moreover, we cannot hope to efficiently find the best approximation in the Kullback–Leibler divergence between the true posterior and the Gaussian approximation as this requires to have an efficient way to compute the derivatives of the normalisation constant w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which is a sum of products and is thus subject to the *curse of dimensionality*. We can still make progress if we assume that it is possible to efficiently incorporate a *single* factor $t_{i,\mathbf{x}}$. This algorithm is known as the *Gaussian expectation propagation (Gaussian EP)* algorithm which was systematically introduced in Minka (2001).

Approximation Model In its most general form, suppose the the ith factor in the likelihood is some function of a low-dimensional projection of $\boldsymbol{\theta}$, that is

$$t_{i,\mathbf{x}}(\boldsymbol{\theta}) = h\left(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\theta}\right) .$$

Then we use the following m functions f_i in place of the m factors $t_{i,\mathbf{x}}$ [1]

$$f_i(\boldsymbol{\theta}) := s_i \exp\left(-\frac{1}{2}\left(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\theta} - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Pi}_i \left(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\theta} - \boldsymbol{\mu}_i\right)\right) ,$$

and define $f_0(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The approximation, $\hat{P}(\boldsymbol{\theta}|\mathbf{x})$, of the posterior, $P(\boldsymbol{\theta}|\mathbf{x})$, is assumed to have the same *functional* form, that is,

$$\hat{P}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\prod_{i=0}^{m} f_i(\boldsymbol{\theta})}{\underbrace{\int \prod_{i=0}^{m} f_i(\tilde{\boldsymbol{\theta}})\, d\tilde{\boldsymbol{\theta}}}_{\hat{Z}_{\mathbf{x}} = \hat{P}(\mathbf{x})}} = \mathcal{N}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right) . \tag{3.1}$$

Note that due to the projection the function $Z_i := \int t_{i,\mathbf{x}}(\boldsymbol{\theta}) \cdot \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{\theta}$, the vector $\mathbf{g}_i := \partial \log(Z_i(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}))/\partial\tilde{\boldsymbol{\mu}}$ and the matrix $\mathbf{G}_i := \partial \log(Z_i(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}))/\partial\tilde{\boldsymbol{\Sigma}}$ have the following functional form

$$\begin{aligned}
Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \int h(y)\, \mathcal{N}\left(y; \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i\right) dy , && (3.2)\\
\mathbf{g}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathbf{A}_i \left[\boldsymbol{\alpha}_i\left(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i\right)\right] ,\\
\mathbf{g}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - 2\mathbf{G}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathbf{A}_i \left[\boldsymbol{\Gamma}_i\left(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i\right)\right] \mathbf{A}_i^{\mathrm{T}} , && (3.3)
\end{aligned}$$

where $\boldsymbol{\alpha}_i$ is vector valued function and $\boldsymbol{\Gamma}_i$ is a matrix valued function for the ith factor.

Algorithmic Overview At the beginning, we assume that $\boldsymbol{\mu}_i = \mathbf{0}$, $\boldsymbol{\Pi}_i = \mathbf{0}$ and $s_i = 1$ which implies that f_i is the constant unit function and thus $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ and $\hat{Z}_{\mathbf{x}} = 1$. The idea of the Gaussian EP algorithm is to pick a factor, say t_j, and improve the corresponding approximation f_j via adjustments to the parameters s_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Pi}_j$. In order to perform this improvement, the EP algorithm:

[1] Note that we have not used the Gaussian density $\mathcal{N}(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\theta}; \boldsymbol{\mu}_i, \boldsymbol{\Pi}_i^{-1})$ in the approximation f_i as these approximations are not required to be densities, that is, $\boldsymbol{\Pi}_i$ does not have to be a positive-semidefinite matrix and f_i does not need to integrate to unity over $\boldsymbol{\theta}$.

1. Computes the parameters $\boldsymbol{\mu}_{\setminus j}$, $\boldsymbol{\Sigma}_{\setminus j}$ of the Gaussian approximation of the posterior *without* the factor t_j but keeping all other factors to their current value,

$$P_{\setminus j}(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right) = \frac{\prod_{i\neq j} f_i(\boldsymbol{\theta})}{\int \prod_{i\neq j} f_i\left(\tilde{\boldsymbol{\theta}}\right) d\tilde{\boldsymbol{\theta}}} .$$

2. Employs the Gaussian density filtering approximation outlined in Section 2 to obtain the new Gaussian approximation of the posterior, $\hat{P}(\boldsymbol{\theta}|\mathbf{x})$, where

$$\hat{P}(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right) \approx \frac{t_{j,\mathbf{x}}(\boldsymbol{\theta}) \cdot \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right)}{\int t_{j,\mathbf{x}}(\tilde{\boldsymbol{\theta}}) \cdot \mathcal{N}(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j})\, d\tilde{\boldsymbol{\theta}}} .$$

   By assumption, this can be done efficiently for every single factor $t_{j,\mathbf{x}}$.

3. Updates the parameters s_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Pi}_j$ of the factor f_j such that

$$\mathcal{N}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right) = \frac{f_j(\boldsymbol{\theta}) \cdot \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right)}{\int f_j(\tilde{\boldsymbol{\theta}}) \cdot \mathcal{N}\left(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right) d\tilde{\boldsymbol{\theta}}} , , \tag{3.4}$$

   and, at the same time,

$$\int f_j(\tilde{\boldsymbol{\theta}}) \cdot \mathcal{N}\left(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right) d\tilde{\boldsymbol{\theta}} \;=\; \int t_{j,\mathbf{x}}(\tilde{\boldsymbol{\theta}}) \cdot \mathcal{N}\left(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right) d\tilde{\boldsymbol{\theta}} . \tag{3.5}$$

   Note that (3.4) alone is not sufficient to update all parameters because s_j appears both in the numerator and denominator of the l.h.s. of (3.4).

Central Relations In order to derive the remove and update equations for the jth function, we use (1.3) in (3.4),

$$\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} = \mathbf{A}_j \boldsymbol{\Pi}_j \boldsymbol{\mu}_j + \boldsymbol{\Sigma}_{\setminus j}^{-1}\boldsymbol{\mu}_{\setminus j} , \qquad \hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{A}_j \boldsymbol{\Pi}_j \mathbf{A}_j^{\mathrm{T}} + \boldsymbol{\Sigma}_{\setminus j}^{-1} . \tag{3.6}$$

Moreover, by virtue of (1.4) we have

$$\int f_j\left(\tilde{\boldsymbol{\theta}}\right) \cdot \mathcal{N}\left(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Sigma}_{\setminus j}\right) d\tilde{\boldsymbol{\theta}} = s_j (2\pi)^{\frac{n}{2}} \left|\boldsymbol{\Pi}_j\right|^{-\frac{1}{2}} \mathcal{N}\left(\boldsymbol{\mu}_j; \mathbf{A}_j^{\mathrm{T}}\boldsymbol{\mu}_{\setminus j}, \boldsymbol{\Pi}_j^{-1} + \mathbf{A}_j^{\mathrm{T}}\boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j\right) \tag{3.7}$$

In the following we will derive efficient and numerically stable removal and update equations using the shorthand notations

$$\mathbf{U}_j := \hat{\boldsymbol{\Sigma}}\mathbf{A}_j , \qquad \mathbf{C}_j := \mathbf{A}_j^{\mathrm{T}}\mathbf{U}_j , \qquad \mathbf{m}_j := \mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\mu}} , \qquad \mathbf{D}_j := \mathbf{C}_j \boldsymbol{\Pi}_j , \tag{3.8}$$

$$\mathbf{E}_j := \left(\mathbf{I} - \mathbf{D}_j\right)^{-1} , \qquad \mathbf{F}_j := \left(\mathbf{I} - \mathbf{D}_j^{\mathrm{T}}\right)^{-1} = \mathbf{I} + \boldsymbol{\Pi}_j \mathbf{E}_j \mathbf{C}_j , \tag{3.9}$$

where the expression for $\mathbf{F}_j$ follows from the Woodbury formula. The full algorithm is given in Algorithm 1 on the following page. In the case of rank 1 update, that is, $\mathbf{A}_i = \mathbf{a}_i$ the EP algorithm can be done without ever computing an inverse and is given in Algorithm 2.

Remove Equations In order to remove the jth function, we use the Woodbury formula and exploit the symmetry of $\hat{\boldsymbol{\Sigma}}$ to get

$$\begin{aligned}
\boldsymbol{\Sigma}_{\setminus j} &= \left(\hat{\boldsymbol{\Sigma}}^{-1} - \mathbf{A}_j \boldsymbol{\Pi}_j \mathbf{A}_j^{\mathrm{T}}\right)^{-1} \\
&= \hat{\boldsymbol{\Sigma}} + \left(\hat{\boldsymbol{\Sigma}}\mathbf{A}_j\right) \boldsymbol{\Pi}_j \left(\mathbf{I} - \mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\mathbf{A}_j \boldsymbol{\Pi}_j\right)^{-1} \left(\mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\right) , && (3.10) \\
\boldsymbol{\mu}_{\setminus j} &= \boldsymbol{\Sigma}_{\setminus j} \left(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} - \mathbf{A}_j \boldsymbol{\Pi}_j \boldsymbol{\mu}_j\right) \\
&= \hat{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j \boldsymbol{\Pi}_j \left(\mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_j\right) && (3.11) \\
&= \hat{\boldsymbol{\mu}} + \left(\hat{\boldsymbol{\Sigma}}\mathbf{A}_j \boldsymbol{\Pi}_j\right) \left(\mathbf{I} - \mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\mathbf{A}_j \boldsymbol{\Pi}_j\right)^{-1} \left(\mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_j\right) . && (3.12)
\end{aligned}$$

Algorithm 1 General Gaussian EP algorithm

Require: Prior mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
Require: A set of m matrices $\mathbf{A}_i$ and a function for which we can efficiently evaluate $Z_i(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i)$, $\boldsymbol{\alpha}_i(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i)$ and $\boldsymbol{\Gamma}(\mathbf{A}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_i)$ (see (3.2)–(3.3)).
Require: A termination criterion.

{Initialisation}
Set $\boldsymbol{\mu}_i = \mathbf{0}$, $\boldsymbol{\Pi}_i = \mathbf{0}$ and $s_i = 1$ for $i \in \{1, \ldots, m\}$. Set $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ and $\hat{Z} = 1$.
repeat
Pick an index $j \in \{1, \ldots, m\}$.
{Pre-computations for the jth factor}
Compute $\mathbf{U}_j = \hat{\boldsymbol{\Sigma}}\mathbf{A}_j$, $\mathbf{C}_j = \mathbf{A}_j^{\mathrm{T}}\mathbf{U}_j$, $\mathbf{m}_j = \mathbf{A}_j^{\mathrm{T}}\hat{\boldsymbol{\mu}}$ and $\mathbf{D}_j = \mathbf{C}_j\boldsymbol{\Pi}_j$.
Compute $\mathbf{E}_j = (\mathbf{I} - \mathbf{D}_j)^{-1}$ and $\mathbf{F}_j = \mathbf{I} + \boldsymbol{\Pi}_j\mathbf{E}_j\mathbf{C}_j$.
Compute $\boldsymbol{\phi}_j = \mathbf{m}_j + \mathbf{D}_j\mathbf{E}_j(\mathbf{m}_j - \boldsymbol{\mu}_j)$ and $\boldsymbol{\Psi}_j = \mathbf{E}_j\mathbf{C}_j$.
Compute $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j)$ and $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}_j(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j)$.
{ADF update}
Compute $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + \mathbf{U}_j\left(\boldsymbol{\Pi}_j\mathbf{E}_j(\mathbf{m}_j - \boldsymbol{\mu}_j) + \mathbf{F}_j\boldsymbol{\alpha}_j\right)$ and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}} + \mathbf{U}_j\left(\boldsymbol{\Pi}_j - \mathbf{F}_j\boldsymbol{\Gamma}_j\right)\mathbf{E}_j\mathbf{U}_j^{\mathrm{T}}$.
{Factor update}
Compute $\boldsymbol{\Pi}_j \leftarrow (\boldsymbol{\Gamma}_j^{-1} - \boldsymbol{\Psi}_j)^{-1}$ and $\boldsymbol{\mu}_j \leftarrow \boldsymbol{\Gamma}_j^{-1}\boldsymbol{\alpha}_j + \boldsymbol{\phi}_j$.
Compute $s_j = Z_j \cdot \exp(\frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\alpha}_j)/\sqrt{|\mathbf{I} - \boldsymbol{\Gamma}_j\boldsymbol{\Psi}_j|}$.
until termination criterion is fulfilled
Compute $\hat{Z} = (\prod_{i=1}^m s_i) \cdot \sqrt{|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|} \cdot \exp(-\frac{1}{2}(\sum_{i=1}^m \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Pi}_i\boldsymbol{\mu}_i + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}))$
return Mean $\hat{\boldsymbol{\mu}}$, covariance $\hat{\boldsymbol{\Sigma}}$ and normalisation constant $\hat{Z}$.

where (3.11) follows by inserting $\mathbf{A}_j\boldsymbol{\Pi}_j\mathbf{A}_j^{\mathrm{T}} + \boldsymbol{\Sigma}_{\setminus j}^{-1}$ for $\hat{\boldsymbol{\Sigma}}^{-1}$ (see (3.6)) and (3.12) follows by left-multiplying (3.10) with $\mathbf{A}_j$and inserting it into (3.11). Thus, using the notation in (3.8) the removal equation can also be written as

$$\boldsymbol{\mu}_{\setminus j} = \hat{\boldsymbol{\mu}} + \mathbf{U}_j\boldsymbol{\Pi}_j\mathbf{E}_j\left(\mathbf{m}_j - \boldsymbol{\mu}_j\right), \quad \boldsymbol{\Sigma}_{\setminus j} = \hat{\boldsymbol{\Sigma}} + \mathbf{U}_j\boldsymbol{\Pi}_j\mathbf{E}_j\mathbf{U}_j^{\mathrm{T}}.$$

We notice that all further equations based on $\boldsymbol{\mu}_{\setminus j}$ and $\boldsymbol{\Sigma}_{\setminus j}$ only depend on $\mathbf{A}_j^{\mathrm{T}}\boldsymbol{\mu}_{\setminus j}$ and $\mathbf{A}_j^{\mathrm{T}}\boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j$. Using (3.10) and (3.12) and the shorthand notations in (3.8) and (3.9) these two quantities are given by

$$\boldsymbol{\phi}_j := \mathbf{A}_j^{\mathrm{T}}\boldsymbol{\mu}_{\setminus j} = \mathbf{m}_j + \mathbf{D}_j\mathbf{E}_j\left(\mathbf{m}_j - \boldsymbol{\mu}_j\right), \quad \boldsymbol{\Psi}_j := \mathbf{A}_j^{\mathrm{T}}\boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j = \mathbf{E}_j\mathbf{C}_j. \tag{3.13}$$

GDF Update Equations According to (2.1), the update equations after removing the jth factor are straightforward and are given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\text{new}} &= \boldsymbol{\mu}_{\setminus j} + \boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j\left[\boldsymbol{\alpha}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right)\right], \\ \hat{\boldsymbol{\Sigma}}_{\text{new}} &= \boldsymbol{\Sigma}_{\setminus j} - \boldsymbol{\Sigma}_{\setminus j}\mathbf{A}_j\left[\boldsymbol{\Gamma}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right)\right]\mathbf{A}_j^{\mathrm{T}}\boldsymbol{\Sigma}_{\setminus j}, \end{aligned}\tag{3.14}$$

where $\boldsymbol{\phi}_j$ and $\boldsymbol{\Psi}_j$ are given by (3.13). Inserting (3.10) and (3.12) for $\boldsymbol{\mu}_{\setminus j}$ and $\boldsymbol{\Sigma}_{\setminus j}$ and using the notation introduced in (3.8) and (3.9) we see that

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\text{new}} &= \hat{\boldsymbol{\mu}} + \mathbf{U}_j\left[\boldsymbol{\Pi}_j\mathbf{E}_j\left(\mathbf{m}_j - \boldsymbol{\mu}_j\right) + \mathbf{F}_j \cdot \boldsymbol{\alpha}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right)\right] \\ \hat{\boldsymbol{\Sigma}}_{\text{new}} &= \hat{\boldsymbol{\Sigma}} + \mathbf{U}_j\left[\boldsymbol{\Pi}_j - \mathbf{F}_j \cdot \boldsymbol{\Gamma}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right)\right]\mathbf{E}_j\mathbf{U}_j^{\mathrm{T}}. \end{aligned}$$

Factor Update Equations We can use the Woodbury formula in (3.14) to derive an inverse of the new covariance matrix $\hat{\boldsymbol{\Sigma}}_{\text{new}}$,

$$\hat{\boldsymbol{\Sigma}}_{\text{new}}^{-1} = \boldsymbol{\Sigma}_{\setminus j}^{-1} + \mathbf{A}_j\left(\boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) - \boldsymbol{\Psi}_j\right)^{-1}\mathbf{A}_j^{\mathrm{T}}.$$

Algorithm 2 Rank 1 Gaussian EP algorithm

Require: Prior mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
Require: A set of m vectors $\mathbf{a}_i$ and a function for which we can efficiently evaluate $Z_i(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}_i)$, $\alpha_i(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}_i)$ and $\gamma_i(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{\mu}, \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}_i)$ (see (3.2)–(3.3)).
Require: A termination criterion.

{Initialisation}
Set $\mu_i = 0$, $\pi_i = 0$ and $s_i = 1$ for $i \in \{1, \ldots, m\}$. Set $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$.
repeat
Pick an index $j \in \{1, \ldots, m\}$.
{Pre-computations for the jth factor}
Compute $\mathbf{u}_j = \hat{\boldsymbol{\Sigma}}\mathbf{a}_j$, $c_j = \mathbf{a}_j^{\mathrm{T}}\mathbf{u}_j$, $m_j = \mathbf{a}_j^{\mathrm{T}}\hat{\boldsymbol{\mu}}$, $d_j = \pi_j c_j$ and $e_j = 1/(1-d_j)$.
Compute $\phi_j = m_j + d_j e_j \left(m_j - \mu_j\right)$ and $\psi_j = e_j c_j$.
Compute $\alpha_j = \alpha_j(\phi_j, \psi_j)$ and $\gamma_j = \gamma_j(\phi_j, \psi_j)$.
{ADF update}
Update $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + e_j\left(\pi_j\left(m_j - \mu_j\right) + \alpha_j\right) \cdot \mathbf{u}_j$ and $\hat{\boldsymbol{\Sigma}} \leftarrow \hat{\boldsymbol{\Sigma}} + e_j^2\left(\pi_j\left(1 - d_j\right) - \gamma_j\right) \cdot \mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}$.
{Factor update}
Update $\pi_j \leftarrow 1/(\gamma_j^{-1} - \psi_j)$ and $\mu_j \leftarrow \alpha_j/\gamma_j + \phi_j$.
Update $s_j \leftarrow Z_j(\phi_j, \psi_j) \cdot \exp(\alpha_j^2/(2\gamma_j))/\sqrt{1 - \psi_j\gamma_j}$.
until termination criterion is fulfilled
Compute $\hat{Z} = (\prod_{i=1}^m s_i) \cdot \sqrt{|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|} \cdot \exp(-\frac{1}{2}(\sum_{i=1}^m \pi_i\mu_i^2 + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}))$
return Mean $\hat{\boldsymbol{\mu}}$, covariance $\hat{\boldsymbol{\Sigma}}$ and normalisation constant $\hat{Z}$.

Furthermore, we can express $\hat{\boldsymbol{\Sigma}}_{\mathrm{new}}^{-1}\hat{\boldsymbol{\mu}}_{\mathrm{new}}$ similarly,

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{new}}^{-1}\hat{\boldsymbol{\mu}}_{\mathrm{new}} = \boldsymbol{\Sigma}_{\setminus j}^{-1}\boldsymbol{\mu}_{\setminus j} + \mathbf{A}_j\left[\left[\boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) - \boldsymbol{\Psi}_j\right]^{-1}\left[\boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \boldsymbol{\alpha}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) + \boldsymbol{\phi}_j\right]\right].$$

Now, we exploit (3.6) and (3.7) to obtain the update equation for the jth factor

$$\begin{aligned}
\boldsymbol{\Pi}_{j,\mathrm{new}} &= \left(\boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) - \boldsymbol{\Psi}_j\right)^{-1}, \\
\boldsymbol{\mu}_{j,\mathrm{new}} &= \boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \boldsymbol{\alpha}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) + \boldsymbol{\phi}_j, \\
s_{j,\mathrm{new}} &= Z_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \left|\mathbf{I} - \boldsymbol{\Gamma}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \boldsymbol{\Psi}_j\right|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \boldsymbol{\Gamma}_j^{-1}\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right) \cdot \boldsymbol{\alpha}_j\left(\boldsymbol{\phi}_j, \boldsymbol{\Psi}_j\right)\right).
\end{aligned}$$

Approximate Normalisation The normalisation $\hat{Z}_{\mathbf{x}} = \hat{P}(\mathbf{x})$ of the approximation of $P(\mathbf{x})$ can easily be computed from (3.1) using $\boldsymbol{\theta} = \mathbf{0}$ resulting in

$$\hat{Z}_{\mathbf{x}} = \left(\prod_{i=1}^m s_i\right) \cdot \sqrt{\frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}\left(\sum_{i=1}^m \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Pi}_i\boldsymbol{\mu}_i + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right)\right). \tag{3.15}$$

4 Rectified Truncated Gaussians

4.1 One Dimensional Rectified Truncated Gaussians

We say that x is distributed according to a *rectified doubly truncated Gaussian* (for short *rectified Gaussian*), $x \sim \mathcal{R}(x; \mu, \sigma^2, l, u)$, if the density of x is given by

$$\begin{aligned} \mathcal{R}\left(x; \mu, \sigma^2, l, u\right) &= \mathbb{I}_{x\in(l,u)} \cdot \frac{\mathcal{N}\left(x; \mu, \sigma^2\right)}{\Phi\left(u; \mu, \sigma^2\right) - \Phi\left(l; \mu, \sigma^2\right)} \\ &= \mathbb{I}_{x\in(l,u)} \cdot \frac{\mathcal{N}\left(\frac{x-\mu}{\sigma}\right)}{\sigma \cdot \left(\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)\right)} . \end{aligned} \tag{4.1}$$

We will write $\mathcal{R}(x; \mu, \sigma^2, l)$ to denote $\lim_{u\to+\infty} \mathcal{R}(x; \mu, \sigma^2, l, u)$; this distribution is sometimes referred to simply as a rectified Gaussian. Note that the class of rectified Gaussian contains the Gaussian family as a special case, that is,

$$\lim_{l\to-\infty} \mathcal{R}\left(x; \mu, \sigma^2, l\right) = \mathcal{N}\left(x; \mu, \sigma^2\right) .$$

We have the following properties for the mean and variance of the double rectified Gaussian (see Figure 4.1).

Proposition 1 (Rectified Gaussian Mean and Variance). *The mean and variance of the rectified Gaussian are given by*

$$\langle x \rangle_{x\sim\mathcal{R}} = \mu + \sigma \cdot v\left(\frac{\mu}{\sigma}, \frac{l}{\sigma}, \frac{u}{\sigma}\right) , \tag{4.2}$$

$$\left\langle x^2 \right\rangle_{x\sim\mathcal{R}} - \left(\langle x \rangle_{x\sim\mathcal{R}}\right)^2 = \sigma^2 \cdot \left(1 - w\left(\frac{\mu}{\sigma}, \frac{l}{\sigma}, \frac{u}{\sigma}\right)\right) , \tag{4.3}$$

where the functions v and w are given by

$$v(t, l, u) := \frac{\mathcal{N}(l-t) - \mathcal{N}(u-t)}{\Phi(u-t) - \Phi(l-t)} , \tag{4.4}$$

$$w(t, l, u) := v^2(t, l, u) + \frac{(u-t)\cdot\mathcal{N}(u-t) - (l-t)\cdot\mathcal{N}(l-t)}{\Phi(u-t) - \Phi(l-t)} . \tag{4.5}$$

Proof. In order to derive the first and second moment of a rectified Gaussian we exploit the symmetry of the function Φ in its first two arguments, that is

$$\Phi\left(\frac{u-\mu}{\sigma}\right) = \Phi\left(u; \mu, \sigma^2\right) = 1 - \Phi\left(\mu; u, \sigma^2\right) . \tag{4.6}$$

The exact equation for the first moment is then obtained by considering the first derivative of $\Phi(\frac{u-\mu}{\sigma}) - \Phi(\frac{l-\mu}{\sigma})$ with respect to μ; the exact equation for the second moment follows from considering the second derivative with respect to μ.

First moment Following the symmetry shown in (4.6) we note

$$\frac{\partial\left(\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)\right)}{\partial\mu} = \int_l^u \frac{\partial\mathcal{N}\left(x; \mu, \sigma^2\right)}{\partial\mu}\, dx = \sigma^{-2} \langle x - \mu \rangle_{x\sim\mathcal{R}} .$$

At the same time

$$\begin{aligned} \frac{\partial\left(\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)\right)}{\partial\mu} &= \frac{\partial\left(\Phi\left(\mu; l, \sigma^2\right) - \Phi\left(\mu; u, \sigma^2\right)\right)}{\partial\mu} \\ &= \mathcal{N}\left(\mu; l, \sigma^2\right) - \mathcal{N}\left(\mu; u, \sigma^2\right) . \end{aligned}$$

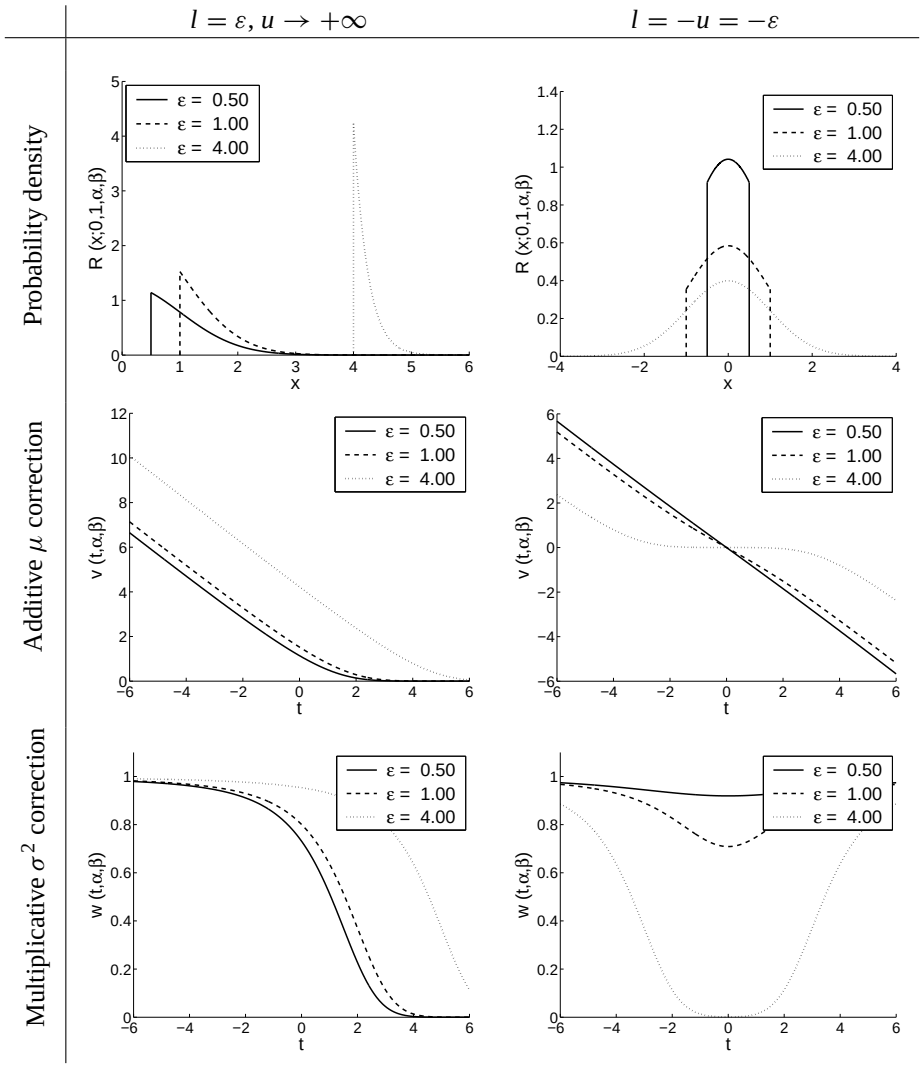

Figure 4.1: The probability density, additive μ correction and multiplicative σ^2 correction as given in (4.1)–(4.3). **(First column)** The special case of a single-sided rectified Gaussian, i.e. $u \to +\infty$. **(Second column)** The special case of a symmetrical rectified Gaussian, i.e. $l = -\beta = -\varepsilon$.

Since the first line must equal the second line we conclude that

$$\begin{aligned} \langle x - \mu \rangle_{x \sim \mathcal{R}} &= \sigma^2 \cdot \frac{\mathcal{N}\left(\mu; l, \sigma^2\right) - \mathcal{N}\left(\mu; u, \sigma^2\right)}{\Phi\left(u; \mu, \sigma^2\right) - \Phi\left(l; \mu, \sigma^2\right)}, \\ \langle x \rangle_{x \sim \mathcal{R}} &= \mu + \sigma^2 \cdot \frac{\mathcal{N}\left(\mu; l, \sigma^2\right) - \mathcal{N}\left(\mu; u, \sigma^2\right)}{\Phi\left(u; \mu, \sigma^2\right) - \Phi\left(l; \mu, \sigma^2\right)}. \end{aligned}$$

This expression is (4.2) when using the function v as defined in (4.4).

Second Moment In order to derive the second moment we appeal again to (4.6) and see that

$$\begin{aligned}\frac{\partial^2\left(\Phi\left(\frac{u-\mu}{\sigma}\right)-\Phi\left(\frac{l-\mu}{\sigma}\right)\right)}{\partial\mu^2} &= \int_l^u \frac{\partial\left(\frac{x-\mu}{\sigma^2}\right)\cdot\mathcal{N}\left(x;\mu,\sigma^2\right)}{\partial\mu}dx \\ &= \sigma^{-4}\left\langle(x-\mu)^2-\sigma^2\right\rangle_{x\sim\mathcal{R}}\,.\end{aligned}$$

At the same time we have

$$\begin{aligned}\frac{\partial^2\left(\Phi\left(\frac{u-\mu}{\sigma}\right)-\Phi\left(\frac{l-\mu}{\sigma}\right)\right)}{\partial\mu^2} &= \frac{\partial\left(\mathcal{N}\left(\mu;l,\sigma^2\right)-\mathcal{N}\left(\mu;u,\sigma^2\right)\right)}{\partial\mu} \\ &= \sigma^{-2}\left[(\mu-u)\,\mathcal{N}\left(\mu;u,\sigma^2\right)-(\mu-l)\,\mathcal{N}\left(\mu;l,\sigma^2\right)\right].\end{aligned}$$

Combining these two lines we obtain

$$\left\langle x^2-2x\mu+\mu^2-\sigma^2\right\rangle_{x\sim\mathcal{R}} = \sigma^2\cdot\frac{(\mu-u)\cdot\mathcal{N}\left(\mu;u,\sigma^2\right)-(\mu-l)\cdot\mathcal{N}\left(\mu;l,\sigma^2\right)}{\Phi\left(u;\mu,\sigma^2\right)-\Phi\left(l;\mu,\sigma^2\right)}$$

$$\left\langle x^2\right\rangle_{x\sim\mathcal{R}} = \mu^2+\sigma^2\left(1-\frac{(\mu+u)\cdot\mathcal{N}\left(\mu;u,\sigma^2\right)-(\mu+l)\cdot\mathcal{N}\left(\mu;l,\sigma^2\right)}{\Phi\left(u;\mu,\sigma^2\right)-\Phi\left(l;\mu,\sigma^2\right)}\right)$$

Variance The variance of a rectified Gaussian is simply given as the difference of the second moment and the squared first moment,

$$\sigma^2\left(1-v^2\left(\frac{\mu}{\sigma},\frac{l}{\sigma},\frac{u}{\sigma}\right)+\frac{(\mu-u)\cdot\mathcal{N}\left(\mu;u,\sigma^2\right)-(\mu-l)\cdot\mathcal{N}\left(\mu;l,\sigma^2\right)}{\Phi\left(u;\mu,\sigma^2\right)-\Phi\left(l;\mu,\sigma^2\right)}\right).$$

This expression is (4.3) when using the function w as defined in (4.5).

□

For the limit of $u\to+\infty$ we note that the function v and w reduce to

$$v(t,l) := \lim_{u\to+\infty} v(t,l,u) = \frac{\mathcal{N}(t-l)}{\Phi(t-l)}\,, \tag{4.7}$$

$$w(t,l) := \lim_{u\to+\infty} w(t,l,u) = v(t,l)\cdot(v(t,l)+(t-l))\,. \tag{4.8}$$

Note that the function w is always bounded by $[0,1]$ whereas v grows roughly like $l-t$ for $t<l$ and quickly approaching zero for $t>l$. Furthermore, the w function is a smooth approximation to the indicator function $\mathbb{I}_{t\leq l}$.

4.2 Multidimensional Rectified Truncated Gaussians

We say that $\mathbf{x}$ is distributed according to a *rectified truncated Gaussian* (for short *rectified Gaussian*), $\mathbf{x}\sim\mathcal{R}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{l},\mathbf{u})$, if the density of $\mathbf{x}$ is given by

$$\mathcal{R}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{l},\mathbf{u}) = \frac{\mathbb{I}_{\mathbf{l}<\mathbf{x}<\mathbf{u}}\cdot\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\int\mathbb{I}_{\mathbf{l}<\tilde{\mathbf{x}}<\mathbf{u}}\cdot\mathcal{N}\left(\tilde{\mathbf{x}};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)d\tilde{\mathbf{x}}}\,.$$

There are no efficient analytic expressions for both the normalisation constant and any moments of this distribution. However, we can use Gaussian EP to compute both the mean, covariance and normalisation constant of this distribution. In order to see this, note that the density $\mathcal{R}$ can be written as

$$\mathcal{R}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{l},\mathbf{u}) = \frac{\prod_{i=1}^n\mathbb{I}_{l_i<x_i<u_i}\cdot\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\int\prod_{i=1}^n\mathbb{I}_{l_i<\tilde{x}_i<u_i}\cdot\mathcal{N}\left(\tilde{\mathbf{x}};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)d\tilde{\mathbf{x}}}\,.$$

Algorithm 3 Approximation algorithm for truncated Gaussians

Require: Mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ of (non-truncated) Gaussian.
Require: Lower and upper truncation points, $\mathbf{l} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^n$.
Require: Efficient method to compute $v(\cdot)$ and $w(.)$ (see (4.4) and (4.5)).
Require: A termination criterion.

{Initialisation}
Set $\mu_i = 0$, $\pi_i = 0$ and $s_i = 1$ for $i \in \{1, \ldots, n\}$. Set $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$.
repeat
  Pick an index $j \in \{1, \ldots, n\}$.
  {Pre-computations for the jth factor}
  Compute $\mathbf{t}_j = [\hat{\Sigma}_{1,j}, \hat{\Sigma}_{2,j}, \ldots, \hat{\Sigma}_{n,j}]^{\mathrm{T}}$, $d_j = \pi_j \hat{\Sigma}_{j,j}$ and $e_j = 1/(1-d_j)$.
  Compute $\phi_j = \hat{\mu}_j + d_j e_j (\hat{\mu}_j - \mu_j)$ and $\psi_j = \hat{\Sigma}_{j,j} e_j$.
  Compute $\phi'_j = \phi_j/\sqrt{\psi_j}$, $\psi'_j = \psi_j/\sqrt{\psi_j}$, $l'_j = l_j/\sqrt{\psi_j}$ and $u'_j = u_j/\sqrt{\psi_j}$.
  Compute $\alpha_j = v(\phi'_j, l'_j, u'_j)/\sqrt{\psi_j}$ and $\beta_j = w(\phi'_j, l'_j, u'_j)/\psi_j$.
  {ADF update}
  Update $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} + e_j \left(\pi_j(\hat{\mu}_j - \mu_j) + \alpha_j\right) \cdot \mathbf{t}_j$ and $\hat{\boldsymbol{\Sigma}} \leftarrow \hat{\boldsymbol{\Sigma}} + \left(\pi_j e_j - e_j^2 \beta_j\right) \cdot \mathbf{t}_j \mathbf{t}_j^{\mathrm{T}}$.
  {Factor update}
  Update $\pi_j \leftarrow 1/(\beta_j^{-1} - \psi_j)$ and $\mu_j \leftarrow \alpha_j/\beta_j + \phi_j$.
  Update $s_j \leftarrow (\Phi(u'_j - \phi'_j) - \Phi(l'_j - \phi'_j)) \cdot \exp(\alpha_j^2/(2\beta_j))/\sqrt{1-\psi_j\beta_j}$.
until termination criterion is fulfilled
Compute $\hat{Z} = (\prod_{i=1}^n s_i) \cdot \sqrt{|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|} \cdot \exp(-\frac{1}{2}(\sum_{i=1}^n \pi_i \mu_i^2 + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}))$.
return Mean $\hat{\boldsymbol{\mu}}$, covariance $\hat{\boldsymbol{\Sigma}}$ and normalisation constant $\hat{Z}$.

All that remains is to derive the exact equations for $Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\alpha_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\gamma_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\begin{aligned} \frac{\partial \log(Z_i)}{\partial \boldsymbol{\mu}} &= \alpha_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \mathbf{a}_i \,, \\ \left[\frac{\partial \log(Z_i)}{\partial \boldsymbol{\mu}}\right]\left[\frac{\partial \log(Z_i)}{\partial \boldsymbol{\mu}}\right]^{\mathrm{T}} - 2 \cdot \frac{\partial \log(Z_i)}{\partial \boldsymbol{\Sigma}} &= \gamma_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}} \,. \end{aligned}$$

To this end, we shall assume that the individual factors have the form

$$t_i(\mathbf{x}) = \mathbb{I}_{l_i < \mathbf{a}_i^{\mathrm{T}}\mathbf{x} < u_i} \,.$$

The exact form follows by setting $\mathbf{a}_i = \mathbf{e}_i$. It will be useful to use the following two shorthand notations

$$\phi_i := \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\mu} \,, \qquad \psi_i := \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a}_i \,.$$

Using (1.2) we have for $Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int t_i(\mathbf{x}) \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\mathbf{x} = \int \mathbb{I}_{l_i < y < u_i} \mathcal{N}(y; \phi_i, \psi_i) \, dy = \Phi\left(\frac{u_i - \phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i - \phi_i}{\sqrt{\psi_i}}\right) \,.$$

Let us start by considering the derivative of $\log(Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$w.r.t. $\boldsymbol{\mu}$. Using the chain rule, we have

$$\frac{\partial \log(Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}))}{\partial \boldsymbol{\mu}} = \frac{1}{Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \cdot \frac{\partial \Phi\left(\frac{u_i - \phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i - \phi_i}{\sqrt{\psi_i}}\right)}{\partial \phi_i} \cdot \frac{\partial \phi_i}{\boldsymbol{\mu}} = \frac{1}{\sqrt{\psi_i}} \cdot \frac{\mathcal{N}\left(\frac{l_i - \phi_i}{\sqrt{\psi_i}}\right) - \mathcal{N}\left(\frac{u_i - \phi_i}{\sqrt{\psi_i}}\right)}{\Phi\left(\frac{u_i - \phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i - \phi_i}{\sqrt{\psi_i}}\right)} \cdot \mathbf{a}_i \,.$$

Thus, using the function v defined in (4.4) we have

$$\alpha_i(\phi_i, \psi_i) = \frac{1}{\sqrt{\psi_i}} \cdot v\left(\frac{\phi_i}{\sqrt{\psi_i}}, \frac{l_i}{\sqrt{\psi_i}}, \frac{u_i}{\sqrt{\psi_i}}\right) \,.$$

In order to derive γ_i we will consider the derivative of $\log(Z_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$ w.r.t. $\boldsymbol{\Sigma}$. Using the chain rule, we see that

$$\begin{aligned}
\frac{\partial \log\left(Z_i\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right)}{\partial \boldsymbol{\Sigma}} &= \frac{1}{Z_i\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)} \cdot \frac{\partial \Phi\left(\frac{u_i-\phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i-\phi_i}{\sqrt{\psi_i}}\right)}{\partial \psi_i} \cdot \frac{\partial \psi_i}{\boldsymbol{\mu}} \\
&= \frac{1}{\Phi\left(\frac{u_i-\phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i-\phi_i}{\sqrt{\psi_i}}\right)} \cdot \frac{\partial \Phi\left(\frac{u_i-\phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i-\phi_i}{\sqrt{\psi_i}}\right)}{\partial \psi_i} \cdot \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}}, \\
&= \frac{1}{2\psi_i} \cdot \frac{\mathcal{N}\left(\frac{l_i-\phi_i}{\sqrt{\psi_i}}\right) \cdot \frac{l_i-\phi_i}{\sqrt{\psi_i}} - \mathcal{N}\left(\frac{u_i-\phi_i}{\sqrt{\psi_i}}\right) \cdot \frac{u_i-\phi_i}{\sqrt{\psi_i}}}{\Phi\left(\frac{u_i-\phi_i}{\sqrt{\psi_i}}\right) - \Phi\left(\frac{l_i-\phi_i}{\sqrt{\psi_i}}\right)} \cdot \mathbf{a}_i \mathbf{a}_i^{\mathrm{T}}
\end{aligned}$$

Thus, for γ_i we obtain

$$\gamma_i\left(\phi_i, \psi_i\right) = \frac{1}{\psi_i} \cdot w\left(\frac{\phi_i}{\sqrt{\psi_i}}, \frac{l_i}{\sqrt{\psi_i}}, \frac{u_i}{\sqrt{\psi_i}}\right).$$

The full algorithm is given in Algorithm 3.

References

Herbrich, R. (2002). *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph. D. thesis, Massachusetts Institute of Technology.