Expectation Propagation for the Clutter Problem – Theory and Implementation

Matěj Korvas

June 11, 2013

Abstract

These lecture notes describe the algorithm of Expectation Propagation as applied to the Clutter problem, touching on the underlying theory. Important is also the practical part where I document my implementation of the algorithm.

1 Introduction

Expectation Propagation (EP for short) was introduced in [3] as an iterated version of the previously known Assumed-Density Filtering approximate inference algorithm. In the work [3], the author also shows how EP is applied to the clutter problem.

In the next section, we describe the EP algorithm in general, in Section 3, we formulate the clutter problem and derive formulas used in EP to solve it, and the final Section 4 discusses our implementation of EP applied to the clutter problem.

2 Expectation Propagation

Expectation Propagation is an approximate inference algorithm for graphical probabilistic models that factorise as follows:

$$p(\mathbf{z}, \mathbf{e}) = \prod_i f_i(\mathbf{z}, \mathbf{e}) \tag{1}$$

where $\mathbf{z}$ is the vector of latent variables, $\mathbf{e}$ is the vector of observed variables (evidence), and f_i are factors that depend on a non-empty subset of $\mathbf{z}$ and a subset of $\mathbf{e}$. This factorisation naturally emerges in experiments with i.i.d. observations where f_0 is the prior on $\mathbf{z}$ and f_i the posterior for the i-th observation for $i = 1, \ldots$

EP approximates the factors f_i with factors $\tilde{f}_i$ that belong to a convenient probability distribution family. The approximation aims to minimise the KL-divergence between a distribution computed using the exact factor f_i, and a distribution using the approximate factor $\tilde{f}_i$. If the approximating distribution family is chosen from the exponential family (which it typically is), minimising the KL-divergence is reduced to *matching moments*, i.e. setting a few moments of the estimating distribution (its sufficient statistics) to the values of corresponding moments of the distribution approximated. Choosing the family from the exponential family has also other benefits, including the fact that this family is closed under the operation of product (this property being assumed in the algorithm), and that Minka [3] proved the existence of a fixed point for the solution provided the family is exponential.

Choosing the approximating family is the first thing done in the algorithm. Next, approximated factors $\tilde{f}_i$ and their product $\mathcal{Q} = \prod_i \tilde{f}_i$ are initialised to uniform. The algorithm then proceeds in iterations, iteratively updating all the approximating factors in each of the outer iterations. When convergence is reached, the normalisation coefficient, an estimate of $p(\mathbf{e})$, is computed. A more detailed exposition of the algorithm follows.

1. Initialisation

All the approximate factors are initialised to uniform, meaning the initial approximation is non-informative. The product $\mathcal{Q}$ of the factors is computed accordingly. Typically, all the factors, as well as their product, are initialised to constant 1.

Factors that already belong to the chosen family can also be computed during initialisation, as such factors are always best approximated by themselves, not needing to be updated iteratively.

2. Outer loop

Following four steps are repeated until convergence.

2.1. Choose a factor $\tilde{f}_i$

Choose a factor to approximate.

2.2. Compute the cavity distribution $\mathcal{Q}^{\setminus i}$

When updating the factor $\tilde{f}_i$, we would ideally want to minimise the KL-divergence between the true distribution and the resulting approximative distribution:

$$\arg\min_{\tilde{f}_i} \mathrm{KL}(p \,||\, \prod_i \tilde{f}_i). \tag{2}$$

However, there we would need to compute moments of p in order to optimise for this KL-divergence. If we were able to do that, we would not need to use approximate inference in the first place, so let us assume this is intractable. In that case, we have to substitute p with an approximation. The approximation used in EP is the following:

$$\hat{p} = \frac{1}{Z_i} f_i \mathcal{Q}^{\setminus i} \tag{3}$$

where

$$\mathcal{Q}^{\setminus i} \propto \prod_{j \neq i} \tilde{f}_j \quad (= \mathcal{Q}/\tilde{f}_i). \tag{4}$$

Here, $\mathcal{Q}^{\setminus i}$ is called the *cavity distribution*, as it is a distribution over $\mathbf{z}$ obtained by multiplying all the approximate factors but the i-th one (thus creating the cavity in the distribution) and normalising (in order to make it a distribution). $\hat{p}$ is defined as a product of the *exact* factor f_i with the rest of the factors *approximated*, normalised to 1, and the cavity distribution needs to be computed in order to express $\hat{p}$.

2.3. Compute the approximative distribution $\mathcal{Q}_{\text{new}}$

Whereas the previous step was concerned with computing the cavity distribution, computing the normalisation coefficient Z_i (as $\int_{\mathbf{Z}} f_i(\mathbf{z})\mathcal{Q}^{\setminus i}(\mathbf{z})\mathrm{d}\mathbf{z}$) and the approximative posterior distribution $\hat{p}$ is reserved for this step.

Having computed $\hat{p}$, we can minimise the KL-divergence to an updated $\mathcal{Q}_{\text{new}}$ restricted to be in the approximating family $\mathcal{F}$:

$$\underset{\mathcal{Q}_{\text{new}} \in \mathcal{F}}{\arg\min}\, \mathrm{KL}(\hat{p} \,||\, \mathcal{Q}_{\text{new}}). \tag{5}$$

As mentioned earlier, this minimisation is achieved by matching moments of $\mathcal{Q}_{\text{new}}$ to those of $\hat{p}$.

2.4. Update the factor

We can see the relation of the f_i, which we wish to approximate, to $\mathcal{Q}_{\text{new}}$ by combining formulas (3) and (5):

$$\mathcal{Q}_{\text{new}} \approx \hat{p} = \frac{1}{Z_i} f_i \mathcal{Q}^{\setminus i}. \tag{6}$$

From here, we easily obtain the formula for the approximation of f_i:

$$f_i \approx \tilde{f}_i = Z_i \frac{\mathcal{Q}_{\text{new}}}{\mathcal{Q}^{\setminus i}}. \tag{7}$$

Thanks to the right hand side of Eq. (7) consisting of a division of distributions from the approximating family (and a coefficient), $\tilde{f}_i$ will also be from that family (provided it is closed under division). Now, the approximate factor gets updated according to Eq. (7), and the outer loop is repeated.

3. Evaluate the normalisation constant

After the algorithm has converged to a set of factors $\{\tilde{f}_i\}$, an approximate posterior $p(\mathbf{z}, \mathbf{e})$ can be computed as a product of the factors, according to the assumptions. If we are interested in $p(\mathbf{e})$, the model evidence, it can be computed now as

$$p(\mathbf{e}) \approx \int_{\mathbf{Z}} \prod_i \tilde{f}_i(\mathbf{z})\mathrm{d}\mathbf{z}. \tag{8}$$

3 The Clutter Problem

Specification of the Problem

In the Clutter problem, we assume a sequence of d-dimensional i.i.d. observations being generated either from a normal distribution with an unknown mean with some probability, or from the "clutter" distribution. The model is specified by the following formulas:

$$W_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(w_0) \tag{9}$$

$$\mathbf{x}_i \mid \boldsymbol{\mu} \overset{\text{ind.}}{\sim} W_i\, \mathcal{N}(\mathbf{0}_d, a\mathbf{I}_d) + (1 - W_i)\, \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) \tag{10}$$

The w_0 parameter determines the *proportion of clutter*, W_i select for each observation whether it was generated from the distribution of interest, $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d)$, or the clutter, and finally, $\boldsymbol{\mu}$ is the unknown mean of the distribution we are trying to estimate. When learning the model, we will not learn W_i explicitly for each i, but rather treat the observations as identically distributed with the same proportion of clutter w_0:

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = w_0 \,\mathcal{N}(\mathbf{0}_d, a\mathbf{I}_d) + (1 - w_0)\,\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d). \tag{11}$$

Finally, we adopt a broad Gaussian prior on $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_d, b\mathbf{I}_d). \tag{12}$$

This problem fits nicely the assumptions for EP:

1. It is intractable to do exact inference to find the value of $\boldsymbol{\mu}$. This is due to the fact that in the Bayesian network, the node for $\boldsymbol{\mu}$ has $(N + 1)$ independent parent nodes, a prior and the N likelihood factors, of which the N likelihood factors have 2 Gaussian components each. This results in the posterior for $\boldsymbol{\mu}$ consisting of 2^N N-dimensional Gaussians, corresponding to the 2^N subsets of observations that could have been generated from the true distribution (as opposed to the clutter).

2. The posterior is a product of factors that depend on a non-empty subset of the latent variables (which is $\{\boldsymbol{\mu}\}$ in this case) and a subset of the observed variables (either $\{\mathbf{x}_i\}$ for the likelihood factors, or $\emptyset$ for the prior) – exactly as required.

Instantiating the general Eq. (1) for the Clutter problem, we get the following:

$$p((\boldsymbol{\mu}), (\mathbf{x}_1, \ldots, \mathbf{x}_N)) = p(\boldsymbol{\mu}) \cdot \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\mu}). \tag{13}$$

In Eq. (13), the generic f_0 is instantiated as the prior $p(\boldsymbol{\mu})$, and the generic $f_i, i = 1, \ldots$ as the likelihood $p(\mathbf{x}_i \mid \boldsymbol{\mu})$. In the following, we may use one or the other notation, whichever is more convenient.

We choose to approximate the factors, and hence also their product, by (unnormalised) spherical Gaussians, with one stipulation: the factors approximating the likelihoods may have their σ^2 parameter negative. This is an inherent property of the algorithm, and we discuss it later in Section 4. Still, each factor $\tilde{f}_i$ can be represented by the triple $\langle \tilde{s}_i, \tilde{\mathbf{m}}_i, \tilde{v}_i \rangle$, describing its scale ($\int_{\mathbf{Z}} \tilde{f}_i(\mathbf{z})\mathrm{d}\mathbf{z}$), mean, and variance, respectively:

$$\tilde{f}_i = \tilde{s}_i \,\mathcal{N}(\tilde{\mathbf{m}}_i, \tilde{v}_i\mathbf{I}_d). \tag{14}$$

Besides that, also the approximate posterior has the same form, and we shall denote its parameters as follows:

$$\mathcal{Q} = \mathcal{N}(\mathbf{m}, v\mathbf{I}_d). \tag{15}$$

Note that $\mathcal{Q}$ is an approximating *distribution*, i.e. it is normalised to 1.

Since $\tilde{f}_0$, the prior, already is a spherical Gaussian, its parameters can be set as part of initialisation:

$$\tilde{s}_0 = 1 \qquad\qquad \tilde{\mathbf{m}}_0 = \mathbf{0}_d \qquad\qquad \tilde{v}_0 = b. \tag{16}$$

This factor is exact and need not be updated anymore.

What remains is expressing the formulas (4), (5), (7) for a factor $\tilde{f}_i, i = 1, \ldots,$ and (8). The following sections are concerned with this.

Update formula for the cavity distribution

The general formula is as follows:

$$\mathcal{Q}^{\setminus i} \propto \mathcal{Q}/\tilde{f}_i. \tag{4 – repeated}$$

After substituting the values of $\mathcal{Q}$ and $\tilde{f}_i$, represented as shown in Eqs. (15) and (14), respectively, we obtain the following:

$$\mathcal{Q}^{\setminus i} \propto \frac{\mathcal{N}(\mathbf{m}, v\mathbf{I}_d)}{\tilde{s}_i\,\mathcal{N}(\tilde{\mathbf{m}}_i, \tilde{v}_i\mathbf{I}_d)}. \tag{17}$$

The parameters of $\mathcal{Q}^{\setminus i}$ can be computed using the formula for the ratio of Gaussians,

$$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)/\,\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = C\,\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{18}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \\
C &= \sqrt{\frac{|\boldsymbol{\Sigma}|\,|\boldsymbol{\Sigma}_2|}{(2\pi)^d|\boldsymbol{\Sigma}_1|}}\,\exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}.
\end{aligned} \tag{19}$$

As the result, we can express $\mathcal{Q}^{\setminus i}$ in terms of its parameters $\mathbf{m}^{\setminus \mathbf{i}}$ (mean) and $v^{\setminus i}$ ($v^{\setminus i}\mathbf{I}_d$ being the variance-covariance matrix) as follows:

$$\mathbf{m}^{\setminus \mathbf{i}} = v^{\setminus i}(\mathbf{m}v^{-1} - \tilde{\mathbf{m}}_i\tilde{v}_i^{-1}) \qquad\qquad v^{\setminus i} = \left(v^{-1} - \tilde{v}_i^{-1}\right)^{-1}. \tag{20}$$

Update formula for $\mathcal{Q}$

In computing $\mathcal{Q}_{\text{new}}$ according to Eq. (5), we have to compute $\hat{p}$ and then its first and second moment in order to arrive at the spherical normal distribution minimising the KL-divergence to $\hat{p}$. In the definition of $\hat{p}$ in Eq. (3), the quantity Z_i is yet to be computed. It is the normalisation constant of $f_i\mathcal{Q}^{\setminus i}$, i.e.:

$$Z_i = \int_{\mathbb{R}^d} f_i(\boldsymbol{\mu})\mathcal{Q}^{\setminus i}(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}. \tag{21}$$

Parameters of $\mathcal{Q}^{\setminus i}$ were obtained in the previous step, and f_i was defined as the likelihood for $\mathbf{x}_i$ (cf. Eq. (11)):

$$f_i(\boldsymbol{\mu}) = w_0\,\mathcal{N}(\mathbf{x}_i; \mathbf{0}_d, a\mathbf{I}_d) + (1 - w_0)\,\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \mathbf{I}_d). \tag{22}$$

Substituting into Eq. (21), we get

$$Z_i = \int_{\mathbb{R}^d}[w_0\,\mathcal{N}(\mathbf{x}_i; \mathbf{0}_d, a\mathbf{I}_d) + (1 - w_0)\,\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \mathbf{I}_d)]\;\mathcal{N}(\boldsymbol{\mu}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\boldsymbol{\mu} \tag{23}$$

$$\begin{aligned}
&= \int_{\mathbb{R}^d} w_0\,\mathcal{N}(\mathbf{x}_i; \mathbf{0}_d, a\mathbf{I}_d)\,\mathcal{N}(\boldsymbol{\mu}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\boldsymbol{\mu} \\
&\quad + \int_{\mathbb{R}^d}(1 - w_0)\,\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \mathbf{I}_d)\,\mathcal{N}(\boldsymbol{\mu}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\boldsymbol{\mu}
\end{aligned} \tag{24}$$

$$= w_0\,\mathcal{N}(\mathbf{x}_i; \mathbf{0}_d, a\mathbf{I}_d) + (1 - w_0)\int_{\mathbb{R}^d}\mathcal{N}(\mathbf{x}_i - \boldsymbol{\mu}; \mathbf{0}_d, \mathbf{I}_d)\,\mathcal{N}(\boldsymbol{\mu}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\boldsymbol{\mu} \tag{25}$$

$$= w_0\,\mathcal{N}(\mathbf{x}_i; \mathbf{0}_d, a\mathbf{I}_d) + (1 - w_0)\,\mathcal{N}\Big(\mathbf{x}_i; \mathbf{m}^{\setminus \mathbf{i}}, (v^{\setminus i} + 1)\mathbf{I}_d\Big) \tag{26}$$

where, going from (25) to (26), we used the result [1] about convolution of Gaussians.

The mean value and variance of $\hat{p}$ can be derived for a general form of the factor f_i. Hence, we will simplify the next derivations by rewriting $\hat{p}$ in the following form:

$$\hat{p}(\boldsymbol{\mu}) = \frac{1}{Z(\mathbf{m}, \boldsymbol{\Sigma})} f(\boldsymbol{\mu})\, \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, \boldsymbol{\Sigma}) \tag{27}$$

where

$$Z(\mathbf{m}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} f(\boldsymbol{\mu})\, \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\mu}. \tag{28}$$

The two moments will be found from derivatives of Z:

$$\frac{\mathrm{d}Z(\mathbf{m}, \boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}} = \int_{\mathbb{R}^d} \frac{\mathrm{d}}{\mathrm{d}\mathbf{m}} \left(f(\boldsymbol{\mu})\, \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, \boldsymbol{\Sigma})\right) \mathrm{d}\boldsymbol{\mu} \tag{29}$$

$$= \int_{\mathbb{R}^d} \frac{f(\boldsymbol{\mu})}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m})\right\} \left((\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1}\right) \mathrm{d}\boldsymbol{\mu} \tag{30}$$

$$= \int_{\mathbb{R}^d} Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \left((\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1}\right) \mathrm{d}\boldsymbol{\mu} \tag{31}$$

$$= Z(\mathbf{m}, \boldsymbol{\Sigma}) \left(\int_{\mathbb{R}^d} \boldsymbol{\mu} \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} - \int_{\mathbb{R}^d} \mathbf{m} \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \tag{32}$$

$$= Z(\mathbf{m}, \boldsymbol{\Sigma}) \cdot \left(\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}] - \mathbf{m}\right)^T \boldsymbol{\Sigma}^{-1}, \tag{33}$$

$$\frac{\mathrm{d}Z(\mathbf{m}, \boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}} = \int_{\mathbb{R}^d} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Sigma}} \left(f(\boldsymbol{\mu})\, \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, \boldsymbol{\Sigma})\right) \mathrm{d}\boldsymbol{\mu} \tag{34}$$

$$\begin{aligned} &= \int_{\mathbb{R}^d} \frac{f(\boldsymbol{\mu})}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m})\right\} \left(\frac{1}{2} \boldsymbol{\Sigma}^{-T} (\boldsymbol{\mu} - \mathbf{m})(\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-T}\right) \\ &\qquad - \frac{1}{2} \frac{f(\boldsymbol{\mu})}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \boldsymbol{\Sigma}^{-T} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m})\right\} \mathrm{d}\boldsymbol{\mu} \end{aligned} \tag{35}$$

$$\begin{aligned} &= \frac{1}{2} \boldsymbol{\Sigma}^{-1} \left[\int_{\mathbb{R}^d} \boldsymbol{\mu}\boldsymbol{\mu}^T Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} - \int_{\mathbb{R}^d} \boldsymbol{\mu} \mathbf{m}^T Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} \right. \\ &\qquad \left. - \int_{\mathbb{R}^d} \mathbf{m} \boldsymbol{\mu}^T Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} + \int_{\mathbb{R}^d} \mathbf{m}\mathbf{m}^T Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu}\right] \boldsymbol{\Sigma}^{-1} \\ &\qquad - \frac{1}{2} \int_{\mathbb{R}^d} \boldsymbol{\Sigma}^{-1} Z(\mathbf{m}, \boldsymbol{\Sigma}) \hat{p}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} \end{aligned} \tag{36}$$

$$= Z(\mathbf{m}, \boldsymbol{\Sigma}) \cdot \left\{\frac{1}{2} \boldsymbol{\Sigma}^{-1} \left[\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}\boldsymbol{\mu}^T] - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbf{m}^T - \mathbf{m}\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T + \mathbf{m}\mathbf{m}^T\right] \boldsymbol{\Sigma}^{-1} - \frac{1}{2} \boldsymbol{\Sigma}^{-1}\right\} \tag{37}$$

where $\mathbf{X}^{-T}$ is a shorthand for $\left(\mathbf{X}^{-1}\right)^T$ $\left(= \left(\mathbf{X}^T\right)^{-1}\right)$ and we applied matrix calculus results from [2].

Eqs. (33) and (37) give us formulas for the moments we are interested in. However, they include the term $Z(\mathbf{m}, \boldsymbol{\Sigma})$, which is an inconvenient integral to compute. Taking the derivative of the $\log$ instead will get us rid of this term:

$$\begin{aligned}
\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}} &= \frac{1}{Z(\mathbf{m},\boldsymbol{\Sigma})}\frac{\mathrm{d}Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}} = \left(\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}] - \mathbf{m}\right)^T \boldsymbol{\Sigma}^{-1} \qquad (38)\\
\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}} &= \frac{1}{Z(\mathbf{m},\boldsymbol{\Sigma})}\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}}\\
&= \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left[\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}\boldsymbol{\mu}^T] - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbf{m}^T - \mathbf{m}\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T + \mathbf{m}\mathbf{m}^T\right]\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1} \qquad (39)
\end{aligned}$$

The first and second moment are now obtained easily from Eqs. (38) and (39) by shuffling them a bit. We get:

$$\begin{aligned}
\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}] &= \mathbf{m} + \boldsymbol{\Sigma}\left(\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}}\right)^T \qquad (40)\\
\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}\boldsymbol{\mu}^T] - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T &= \boldsymbol{\Sigma}\left(2\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\Sigma} - \Big[-\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbf{m}^T - \mathbf{m}\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T + \mathbf{m}\mathbf{m}\Big]\\
&\quad - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T \qquad (41)\\
&= 2\boldsymbol{\Sigma}\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\\
&\quad - \Big[\left(\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}] - \mathbf{m}\right)\left(\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}] - \mathbf{m}\right)^T - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T\Big] - \mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]\mathbb{E}_{\hat{p}}[\boldsymbol{\mu}]^T \qquad (42)\\
&= 2\boldsymbol{\Sigma}\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}}\boldsymbol{\Sigma} + \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\left(\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}}\right)^T\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}}\boldsymbol{\Sigma}^T \qquad (43)\\
&= \boldsymbol{\Sigma}\left[2\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\boldsymbol{\Sigma}} - \left(\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}}\right)^T\frac{\mathrm{d}\log Z(\mathbf{m},\boldsymbol{\Sigma})}{\mathrm{d}\mathbf{m}}\right]\boldsymbol{\Sigma} + \boldsymbol{\Sigma} \qquad (44)
\end{aligned}$$

Now, what remains to be computed in order to arrive at the KL-divergence minimiser are the derivatives of $\log Z$:

$$\frac{\mathrm{d}\log Z_i}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} = \frac{\mathrm{d}\log\left[w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)\right]}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} \tag{45}$$

$$= \frac{1}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)} \cdot \frac{(1-w_0)\mathrm{d}\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} \tag{46}$$

$$= \frac{1}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)} \cdot \frac{\frac{1-w_0}{\sqrt{(2\pi)^d|(v^{\setminus i}+1)\mathbf{I}_d|}}\mathrm{d}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\setminus i})^T((v^{\setminus i}+1)\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\setminus i})\right)}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} \tag{47}$$

$$= \frac{\frac{1-w_0}{\sqrt{(2\pi)^d|(v^{\setminus i}+1)\mathbf{I}_d|}}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\setminus i})^T((v^{\setminus i}+1)\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\setminus i})\right)}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)} \cdot \frac{\mathrm{d}\left(-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\setminus i})^T((v^{\setminus i}+1)\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\setminus i})\right)}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} \tag{48}$$

$$= \frac{(1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)} \cdot \Big(\mathbf{x}_i-\mathbf{m}^{\setminus\mathbf{i}}\Big)^T \Big((v^{\setminus i}+1)\mathbf{I}_d\Big)^{-1} \tag{49}$$

Let us simplify the expression by introducing r as the probability of $\mathbf{x}_i$ not being generated from the clutter, and realising that multiplication by the last term is equivalent to division by $(v^{\setminus i}+1)$:

$$r := \frac{(1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\setminus\mathbf{i}},(v^{\setminus i}+1)\mathbf{I}_d\big)} \tag{50}$$

$$\frac{\mathrm{d}\log Z_i}{\mathrm{d}\mathbf{m}^{\setminus\mathbf{i}}} = r\frac{\big(\mathbf{x}_i-\mathbf{m}^{\setminus\mathbf{i}}\big)^T}{v^{\setminus i}+1} \tag{51}$$

The derivative of $\log Z_i$ by the variance parameter is obtained similarly (let $\mathbf{\Sigma}$ denote the second parameter of Z_i, which has the value $v^{\backslash i}\mathbf{I}_d$):

$$\frac{\mathrm{d}\log Z_i}{\mathrm{d}\boldsymbol{\Sigma}} = \frac{\mathrm{d}\log\left[w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)\right]}{\mathrm{d}\boldsymbol{\Sigma}} \tag{52}$$

$$= \frac{1}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)} \cdot \frac{(1-w_0)\mathrm{d}\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)}{\mathrm{d}\boldsymbol{\Sigma}} \tag{53}$$

$$\begin{aligned}= &\frac{(1-w_0)(2\pi)^{-d/2}}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)} \\ &\cdot \frac{\mathrm{d}\left[|\boldsymbol{\Sigma}+\mathbf{I}_d|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\backslash i})^T(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\backslash i})\right)\right]}{\mathrm{d}\boldsymbol{\Sigma}}\end{aligned} \tag{54}$$

$$\begin{aligned}= &\frac{(1-w_0)(2\pi)^{-d/2}}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)} \\ &\cdot \Bigg\{|\boldsymbol{\Sigma}+\mathbf{I}_d|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})\right] \\ &\quad\cdot\left(\frac{1}{2}(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-T}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-T}\right) \\ &\quad-\frac{1}{2}\,|\boldsymbol{\Sigma}+\mathbf{I}_d|^{-1/2}\,(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-T}\exp\left[-\frac{1}{2}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T(\boldsymbol{\Sigma}+\mathbf{I}_d)^{-1}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})\right]\Bigg\}\end{aligned} \tag{55}$$

$$\begin{aligned}= &\frac{(1-w_0)}{w_0\,\mathcal{N}(\mathbf{x}_i;\mathbf{0}_d,a\mathbf{I}_d) + (1-w_0)\,\mathcal{N}\big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\big)} \\ &\cdot\left[\mathcal{N}\Big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\Big)\cdot\frac{(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T}{2(v^{\backslash i}+1)^2} - \mathcal{N}\Big(\mathbf{x}_i;\mathbf{m}^{\backslash\mathbf{i}},\boldsymbol{\Sigma}+\mathbf{I}_d\Big)\cdot\frac{\mathbf{I}_d}{2(v^{\backslash i}+1)}\right]\end{aligned} \tag{56}$$

$$= \frac{r}{2(v^{\backslash i}+1)^2}\cdot\Big[(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T-(v^{\backslash i}+1)\mathbf{I}_d\Big] \tag{57}$$

Substituting into Eqs. (40) and (44), we finally arrive at the new parameters of $\mathcal{Q}$, $\mathbf{m}_{\text{new}}$ (mean) and $\boldsymbol{\Sigma}_{\text{new}}$ (variance):

$$\mathbf{m}_{\text{new}} = \mathbf{m}^{\backslash\mathbf{i}} + \boldsymbol{\Sigma}\left(\frac{\mathrm{d}\log Z_i}{\mathrm{d}\mathbf{m}^{\backslash\mathbf{i}}}\right)^T = \mathbf{m}^{\backslash\mathbf{i}} + \boldsymbol{\Sigma}r\frac{\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}}}{v^{\backslash i}+1} = \mathbf{m}^{\backslash\mathbf{i}} + r\frac{v^{\backslash i}}{v^{\backslash i}+1}(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}}) \tag{58}$$

$$\boldsymbol{\Sigma}_{\text{new}} = \boldsymbol{\Sigma}\left\{2\frac{\mathrm{d}\log Z_i}{\mathrm{d}\boldsymbol{\Sigma}} - \left(\frac{\mathrm{d}\log Z_i}{\mathrm{d}\mathbf{m}^{\backslash\mathbf{i}}}\right)^T\frac{\mathrm{d}\log Z_i}{\mathrm{d}\mathbf{m}^{\backslash\mathbf{i}}}\right\}\boldsymbol{\Sigma}+\boldsymbol{\Sigma} \tag{59}$$

$$\begin{aligned}= \boldsymbol{\Sigma}\Bigg\{&\frac{r}{(v^{\backslash i}+1)^2}\cdot\Big[(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T-(v^{\backslash i}+1)\mathbf{I}_d\Big] \\ &-\left(r\frac{\big(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}}\big)^T}{v^{\backslash i}+1}\right)^T r\frac{\big(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}}\big)^T}{v^{\backslash i}+1}\Bigg\}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\end{aligned} \tag{60}$$

$$= r\left(\frac{v^{\backslash i}}{v^{\backslash i}+1}\right)^2\cdot\Big[(1-r)(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})(\mathbf{x}_i-\mathbf{m}^{\backslash\mathbf{i}})^T-(v^{\backslash i}+1)\mathbf{I}_d\Big]+\boldsymbol{\Sigma} \tag{61}$$

where we again used the symbol $\mathbf{\Sigma}$ to denote $v^{\backslash i}\mathbf{I}_d$.

However, this $\mathbf{\Sigma}_{\text{new}}$ is generally not a variance matrix of a spherical normal, which is the form we assume for the posterior distribution. Hence, we need to find the KL-divergence minimiser of a spherical normal from a normal with the general covariance $\mathbf{\Sigma}$.[1] Let us solve this problem now, denoting the general multivariate normal with $\mathcal{Q}$ and the spherical one with $\mathcal{S}$, and assuming the mean $\mathbf{0}_d$ for both, WLOG:

$$\begin{aligned}
\arg\min_v \mathrm{KL}(\mathcal{Q}\,||\,\mathcal{S}) &= \arg\min_v \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \log \frac{\mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma})}{\mathcal{N}(\mathbf{x};\mathbf{0}_d,v\mathbf{I}_d)} \mathrm{d}\mathbf{x} && (62)\\
&= \arg\min_v \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \log \frac{\sqrt{|v\mathbf{I}_d|}\exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right\}}{\sqrt{|\mathbf{\Sigma}|}\exp\left\{-\frac{1}{2}\mathbf{x}^T v^{-1}\mathbf{I}_d\mathbf{x}\right\}} \mathrm{d}\mathbf{x} && (63)\\
&= \arg\min_v \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \left[\frac{d}{2}\log v - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{1}{2}\mathbf{x}^T v^{-1}\mathbf{I}_d\mathbf{x}\right] \mathrm{d}\mathbf{x} && (64)\\
&= \arg\min_v \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \big(d\log v + \mathbf{x}^T\mathbf{x}/v\big) \mathrm{d}\mathbf{x} && (65)\\
&=: \arg\min_v G(v) && (66)
\end{aligned}$$

Because the function we minimise here is smooth for $v > 0$, we shall find the minimum by setting the derivative equal to zero:

$$\begin{aligned}
0 &= \frac{\mathrm{d}G}{\mathrm{d}v}(v^*) && (67)\\
&= \left(\frac{\mathrm{d}}{\mathrm{d}v}\int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \big(d\log v + \mathbf{x}^T\mathbf{x}/v\big) \mathrm{d}\mathbf{x}\right)(v^*) && (68)\\
&= \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}}{\mathrm{d}v} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \big(d\log v + \mathbf{x}^T\mathbf{x}/v\big) \mathrm{d}\mathbf{x}\right)(v^*) && (69)\\
&= \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x};\mathbf{0}_d,\mathbf{\Sigma}) \big(d/v - \mathbf{x}^T\mathbf{x}/v^2\big) \mathrm{d}\mathbf{x}\Big|_{v=v^*} && (70)\\
&= \frac{d}{v} - \frac{1}{v^2}\mathbb{E}_{\mathcal{Q}}[\mathbf{x}^T\mathbf{x}]\Big|_{v=v^*} && (71)
\end{aligned}$$

Using the following identity for the product of a quadratic form with a Gaussian density function,

$$\int_{\mathbb{R}^d} (\mathbf{x}-\mathbf{x_0})^T\mathbf{F}^{-1}(\mathbf{x}-\mathbf{x_0})\,\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Sigma})\mathrm{d}\mathbf{x} = (\mathbf{x_0}-\boldsymbol{\mu})^T\mathbf{F}^{-1}(\mathbf{x_0}-\boldsymbol{\mu}) + \mathrm{Tr}[\mathbf{F}^{-1}\mathbf{\Sigma}], \qquad (72)$$

we can express the minimiser v^* from Eq. (71):

$$v^* = \mathrm{Tr}[\mathbf{\Sigma}]/d. \qquad (73)$$

This result can be combined with Eq. (61) to give us the updated variance of the spherical Gaussian

[1] This Σ will be the $\mathbf{\Sigma}_{\text{new}}$ as given by Eq. (61).

posterior:

$$v_{\text{new}} = \operatorname{Tr}\left[r\left(\frac{v^{\setminus i}}{v^{\setminus i}+1}\right)^2 \cdot \Big[(1-r)(\mathbf{x}_i - \mathbf{m}^{\setminus \mathbf{i}})(\mathbf{x}_i - \mathbf{m}^{\setminus \mathbf{i}})^T - (v^{\setminus i}+1)\mathbf{I}_d\Big] + v^{\setminus i}\mathbf{I}_d \right] /d \tag{74}$$

$$= r\left(\frac{v^{\setminus i}}{v^{\setminus i}+1}\right)^2 \cdot \Big[(1-r)\operatorname{Tr}\Big[(\mathbf{x}_i - \mathbf{m}^{\setminus \mathbf{i}})(\mathbf{x}_i - \mathbf{m}^{\setminus \mathbf{i}})^T\Big]/d - (v^{\setminus i}+1)\Big] + v^{\setminus i} \tag{75}$$

$$= v^{\setminus i} - r\frac{(v^{\setminus i})^2}{v^{\setminus i}+1} + \frac{r(1-r)}{d}\left(\frac{v^{\setminus i}}{v^{\setminus i}+1}\right)^2 ||\mathbf{x}_i - \mathbf{m}^{\setminus \mathbf{i}}||^2 \tag{76}$$

Eqs. (26), (58) and (76) give us the updated parameters for $\mathcal{Q}$, which was the objective of this step.

Update formula for $\tilde{f}_i$

The updated $\tilde{f}_i$ is now computed according to Eq. (7) using the formula for a ratio of Gaussians, yielding:

$$\tilde{v}_i = \Big((v_{\text{new}})^{-1} - (v^{\setminus i})^{-1}\Big)^{-1} \tag{77}$$

$$\tilde{\mathbf{m}}_i = \tilde{v}_i \Big((v_{\text{new}})^{-1}\mathbf{m}_{\text{new}} - (v^{\setminus i})^{-1}\mathbf{m}^{\setminus \mathbf{i}}\Big). \tag{78}$$

We could compute the normalisation constant $\tilde{s}_i$ using the appropriate formula for a ratio of Gaussians, but we can get it in a simpler way in terms of a convolution of Gaussians if we reorganise Eq. (7) slightly:

$$\tilde{f}_i = Z_i \frac{\mathcal{Q}_{\text{new}}}{\mathcal{Q}^{\setminus i}} \tag{7 – repeated}$$

$$Z_i\mathcal{Q}_{\text{new}} = \tilde{s}_i\,\mathcal{N}(\tilde{\mathbf{m}}_{\mathbf{i}}, \tilde{v}_i\mathbf{I}_d)\cdot\mathcal{Q}^{\setminus i} \tag{79}$$

$$Z_i = \int_{\mathbb{R}^d} Z_i\,\mathcal{N}(\mathbf{x}; \mathbf{m}_{\text{new}}, v_{\text{new}}\mathbf{I}_d)\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \tilde{s}_i\,\mathcal{N}(\mathbf{x}; \tilde{\mathbf{m}}_{\mathbf{i}}, \tilde{v}_i\mathbf{I}_d)\,\mathcal{N}(\mathbf{x}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\mathbf{x} \tag{80}$$

$$= \tilde{s}_i \int_{\mathbb{R}^d} \mathcal{N}(\tilde{\mathbf{m}}_{\mathbf{i}} - \mathbf{x}; \mathbf{0}_d, \tilde{v}_i\mathbf{I}_d)\,\mathcal{N}(\mathbf{x}; \mathbf{m}^{\setminus \mathbf{i}}, v^{\setminus i}\mathbf{I}_d)\mathrm{d}\mathbf{x} \tag{81}$$

$$= \tilde{s}_i\,\mathcal{N}(\tilde{\mathbf{m}}_i; \mathbf{m}^{\setminus \mathbf{i}}, (\tilde{v}_i + v^{\setminus i})\mathbf{I}_d) \tag{82}$$

Note how we moved from Eq. (79) to (80) – the former asserts the equality of measures (unnormalised distributions), hence their integral over the whole sample space must equal too, as asserted in Eq. (80).

From the equality of (80) and (82), we easily obtain the value of $\tilde{s}_i$ as

$$\tilde{s}_i = \frac{Z_i}{\mathcal{N}(\tilde{\mathbf{m}}_i; \mathbf{m}^{\setminus \mathbf{i}}, (\tilde{v}_i + v^{\setminus i})\mathbf{I}_d)}. \tag{83}$$

Formula for the normalisation constant

According to Eq. (8), here we need to evaluate the normalisation constant of a product of $(N+1)$ (spherical) Gaussians $\tilde{f}_i, i = 0, \ldots, N$. The general formulas for a product of a number of Gaussians are the following:

$$\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{Z}\,\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{84}$$

where

$$\mathbf{\Sigma} = \left(\sum_{i=1}^{N} \mathbf{\Sigma}_i^{-1} \right)^{-1} \tag{85}$$

$$\boldsymbol{\mu} = \mathbf{\Sigma} \left(\sum_{i=1}^{N} \mathbf{\Sigma}_i^{-1} \boldsymbol{\mu}_i \right) \tag{86}$$

$$Z = \frac{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}}{\prod_{i=1}^{N} (2\pi)^{d/2} |\mathbf{\Sigma}_i|^{1/2}} \exp\left\{ \frac{1}{2} \left(\boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu} - \sum_{i=1}^{N} \boldsymbol{\mu}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu} \right) \right\}. \tag{87}$$

In the case of spherical Gaussians, using the notation introduced for the Clutter problem, the normalisation constant Z is expressed as follows:

$$Z = (2\pi v)^{d/2} \exp(B/2) \prod_{i=0}^{N} \tilde{s}_i (2\pi \tilde{v}_i)^{-d/2} \tag{88}$$

where

$$B = \mathbf{m}^T v^{-1} \mathbf{m} - \sum_{i=0}^{N} \tilde{\mathbf{m}}_i^T \tilde{v}_i^{-1} \tilde{\mathbf{m}}_i. \tag{89}$$

Interpreting the results

Here we briefly summarise what is the actual result of running the algorithm described up to here. After the posterior parameters $\mathbf{m}$ and v stopped changing in the EP updates, and the normalisation constant Z has been evaluated according to Eq. (88), the approximated posterior for $\boldsymbol{\mu}$ is given by

$$p(\boldsymbol{\mu}) = \frac{1}{Z} \mathcal{N}(\boldsymbol{\mu}; \mathbf{m}, v\mathbf{I}_d). \tag{90}$$

4 Implementation

I have written a Python script implementing EP for the clutter problem that generates a random $\boldsymbol{\mu}$ and data from the implied distribution, and then infers the posterior on $\boldsymbol{\mu}$ from the cluttered observations. In the following two sections, I summarise the script's features and point out some interesting aspects of the implementation.

Script Features

Configuration. The clutter problem has a number of parameters, including:

- number of observations N
- proportion of clutter w
- number of dimensions d

- variance of the prior on $\boldsymbol{\mu}$, $b\mathbf{I}_d$ (b being the parameter)
- parameters of the clutter: $\mathbf{m}_c$ (the mean; fixed at $\mathbf{0}_d$ in the preceding text), and a ($a\mathbf{I}_d$ being the variance).

Two other parameters are involved, related to the execution of the algorithm:

- maximum number of iterations
- tolerance.

All the parameters are set by overwriting their values near the beginning of the script. The prior on $\boldsymbol{\mu}$ is treated as its true distribution, meaning the value of $\boldsymbol{\mu}$ is sampled from the prior in the first phase of computation.

The script can be asked to run in a debugging mode or in an interactive mode (any combination thereof). This is also set by modifying the script (variables `DEBUG` and `INTERACTIVE`).

Interactivity. When asked to run in the interactive mode, the script pauses after each iteration of the inner loop (iterating over observed points) and brings up a window with a plot of the current situation (see Fig. 1), waiting for the user to hit return until it continues. The plot shows the following items:

- points that have been used to update the posterior estimate ("used points")
- the point used in the last inner loop iteration ("last point")
- points that were not used to update the posterior estimate due to the factor variance being negative ("skipped points"; cf. the next section for explanation)
- the true mean ("true x")
- PDF of the distribution from which the samples were drawn ("original distribution")
- the cavity distribution PDF ("cavity")
- last point's posterior PDF or the last updated factor normalised to 1 ("last point's posterior")
- the estimated posterior PDF ("x posterior")
- PDF of the estimated posterior weighted with the clutter distribution ("cluttered x posterior").

If $d > 1$ was set in the configuration, only the first dimension is shown in this plot.

After hitting return, next iterations of the inner loop are performed until another valid point (having a valid variance for its factor) is encountered. Then, the plot is updated. Subsequent iterations can be skipped by pressing Ctrl-D (end of input) instead of return.

When convergence is reached, a plot of all normalised factors is shown until the user hits return once again. In the case of $d > 1$, only projection to the first dimension is plotted. An example of this plot is shown in Figure 2.

Output. The script dumps all variables' values that capture the current state of the estimation, once at the beginning and once at the end. The key for their names as shown on the output follows:

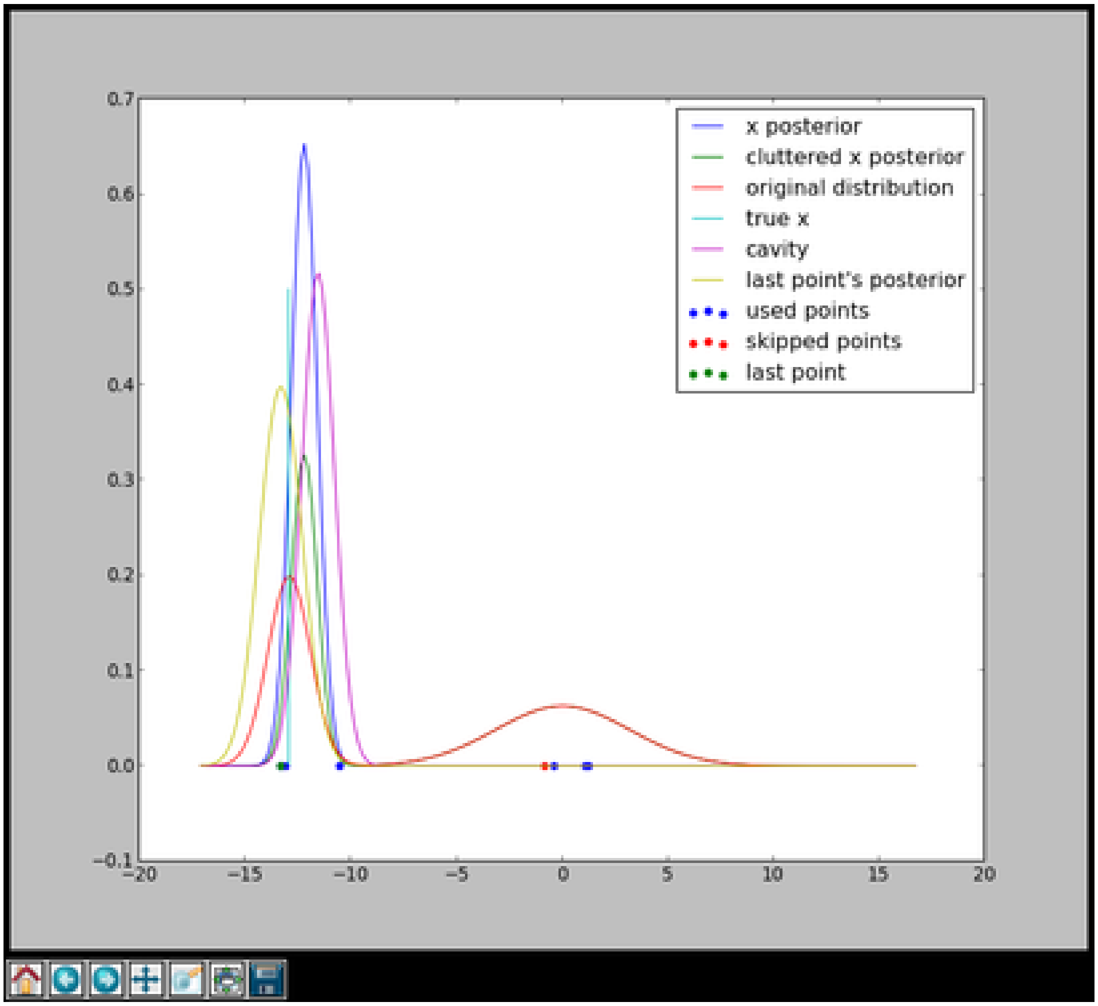

Figure 1: An example plot drawn in the interactive mode

script		notes	script		notes
`x`	…	$\boldsymbol{\mu}$	`mx`	…	$\tilde{\mathbf{m}}$
`y`	…	$\mathbf{x}$	`vx`	…	$\tilde{v}$
`vs`	…	$\tilde{v}_i, i = 1, \ldots, N$	`B`	…	B (as used in Eq. (89))
`ms`	…	$\tilde{\mathbf{m}}_i, i = 1, \ldots, N$	`Z`	…	Z (as used in Eq. (88))
`ss`	…	$\tilde{s}_i, i = 1, \ldots, N$			

Table 2: Key to variable names dumped by the script

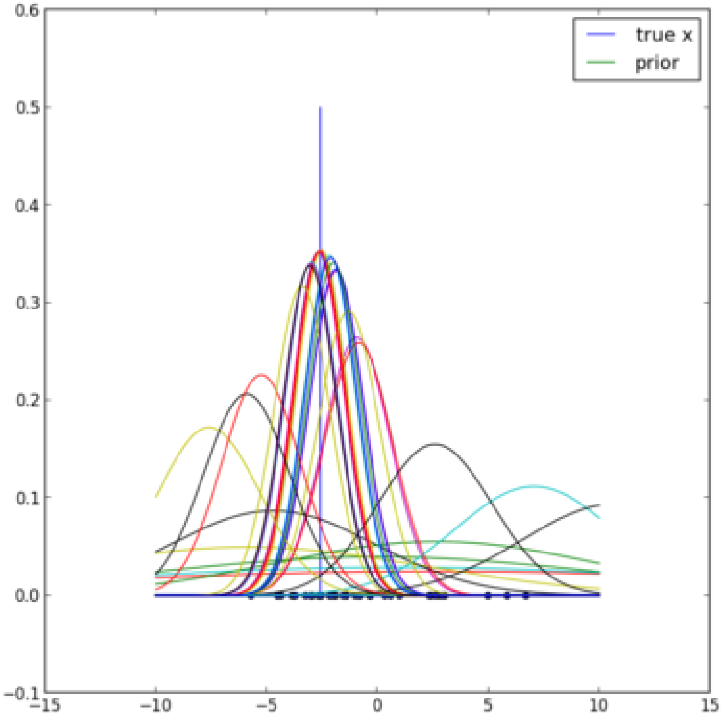

Figure 2: Example plot of factors after convergence

Interesting Points

This subsection points out some problems I encountered during writing the script, most of which have been solved but some not yet. Some other aspects of the implementation are also discussed here, which are not actually problems.

The first problem encountered is inherent to initialisation of the factors that are supposed to not influence the initial estimate of the posterior. For that reason, they are set to constant 1 over the whole space of $\mathbb{R}^d$, while being assumed to be in the form

$$\tilde{f}_i = \tilde{s}_i \, \mathcal{N}(\tilde{\mathbf{m}}_i, \tilde{v}_i \mathbf{I}_d). \tag{91}$$

However it is mathematically unsustainable, these factors are typically represented as having arbitrary $\mathbf{m}_i$, unit $\tilde{s}_i$ and infinite $\tilde{v}_i$. That is also the way they have been implemented in the script.

Next, factors in the inner loop of EP can be iterated in an arbitrary order. I have not done any randomisation to check this actually holds true, and so the order of iterating the factors is fixed (they are iterated sequentially from the first one to the last one).

Several problems arose related to the actual computations. Some of them come from different sources providing slightly different formulas. These alternative versions of formulas are sometimes apparently equivalent (I was not able to prove or disprove they are), and sometimes differ due to a typo. They are always commented on in the code.

More severe problems, though, are caused by some optimising factors not actually being in the required form (of a spherical Gaussian) due to the estimated negative variance. These are called *skipped points* (cf. Fig. 1) as the iteration where their likelihood factor gets updated is skipped to keep all factors in the assumed form. This way of dealing with negative variances was suggested at the lectures.

The other extreme to a variance being negative is its being infinite, which also happens in the computations. I have tried three alternative ways to deal with infinite variances:

1. Leave the variance infinite.

2. Set the variance to a very large number (e.g., 2^{512}).

3. Skip the point with an infinite variance.

Of these three strategies, the first one seemed to work best. All of them are left in the code commented out. Infinite variances arise from dividing two Gaussians with the same variance, where the resulting variance should be computed as $1/0$. As the variances of the two Gaussians go closer, variance of the resulting Gaussian goes to infinity. This is illustrated in Fig. 3, which shows a sequence of functions obtained as

$$f(x) = \mathcal{N}(x; \mu_1, \sigma^2) / \mathcal{N}(x; \mu_2, k \cdot \sigma^2) \tag{92}$$

with k going to 1 from above. The smaller k (subject to $k > 1$), the larger curve we see in Fig. 3.

Related to the two extreme values obtained for variance estimates are the estimates of the scale coefficients $\tilde{s}_i$ which often acquire the value of infinity. However, they are only used in the end to evaluate the normalisation constant. The only problem is that infinite scale coefficients would lead to an infinite normalisation coefficient, therefore Z is computed only from factors having their scale coefficient finite.

Last problem to be discussed is the normalisation coefficient. It is often very small, and sometimes even evaluated as zero due to rounding error. Precision of this value would be enhanced if it was computed in log-scale. However, this is one of many enhancements that are not implemented in the script.

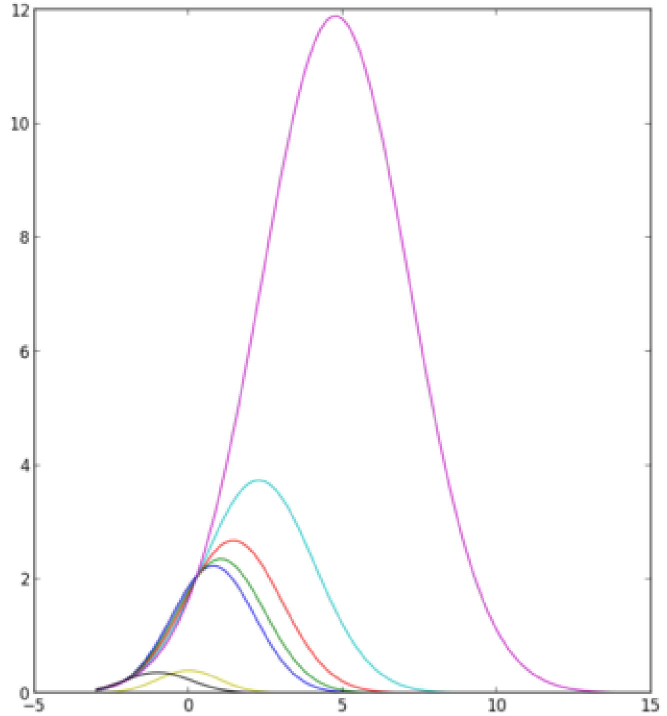

Figure 3: Dividing Gaussians with similar variance

References

[1] BROMILEY, P. Products and convolutions of Gaussian distributions. Internal Report 2003-003, TINA Vision, 2003.

[2] FACKLER, P. L. Notes on matrix calculus, 2005.

[3] MINKA, T. P. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence* (2001), Morgan Kaufmann Publishers Inc., pp. 362–369.