Minimum Distance Estimation for Gaussian Mixtures

In general,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \int \Big[p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x})\Big]^2 d\boldsymbol{x} && (1)\\
&= \arg\min_{\boldsymbol{\theta}} \left(\int \Big[p(\boldsymbol{x}|\boldsymbol{\theta})\Big]^2 d\boldsymbol{x} - 2E_{p(\boldsymbol{x})}\Big[p(\boldsymbol{x}|\boldsymbol{\theta})\Big] \right)\\
&\approx \arg\min_{\boldsymbol{\theta}} \left(\int \Big[p(\boldsymbol{x}|\boldsymbol{\theta})\Big]^2 d\boldsymbol{x} - \frac{2}{N}\sum_{r=1}^{N} p(\boldsymbol{x}_r|\boldsymbol{\theta}) \right) && (2)
\end{aligned}$$

Assuming a mixture of Normal distributions, we find

$$\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\theta}) &= \sum_{i=1}^{K} g_i\, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \text{with} \sum_{i=1}^{K} g_i = 1, \quad \text{then} && (3)\\
\Big[p(\boldsymbol{x}|\boldsymbol{\theta})\Big]^2 &= \sum_{i=1}^{K}\sum_{j=1}^{K} g_i\, g_j\, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} g_i\, g_j\, R_{ij}\, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}), \qquad \text{where} && (4)\\
\boldsymbol{\mu}_{ij} &= \boldsymbol{\Sigma}_{ij}\Big(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\Big), && (5)\\
\boldsymbol{\Sigma}_{ij} &= \Big(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\Big)^{-1}, && (6)\\
R_{ij} &= (2\pi)^{-d/2}\, |\boldsymbol{\Sigma}_{ij}|^{+1/2}\, |\boldsymbol{\Sigma}_i|^{-1/2}\, |\boldsymbol{\Sigma}_j|^{-1/2} \quad \times\\
&\qquad \exp\Big[-\frac{1}{2}\Big(\boldsymbol{\mu}_i^T\, \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^T\, \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_{ij}^T\, \boldsymbol{\Sigma}_{ij}^{-1}\, \boldsymbol{\mu}_{ij}\Big)\Big]\\
&= (2\pi)^{-d/2}\, f(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, -1)\, f(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, -1)\, f(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}, +1), \quad \text{and} && (7)\\
f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, a) &\equiv |\boldsymbol{\Sigma}|^{+a/2}\, \exp\Big(+\frac{a}{2}\, \boldsymbol{\mu}^T\, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big). && (8)
\end{aligned}$$

Note that a Normal distribution can also be written in terms of the f function defined above: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} f(\boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{\Sigma}, -1)$.

The integration can now be performed trivially and we obtain:

$$\hat{\boldsymbol{\theta}} \approx \arg\min_{\boldsymbol{\theta}} \Bigg[\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{K} g_i\, g_j\, R_{ij} - \frac{2}{N}\sum_{r=1}^{N}\sum_{i=1}^{K} g_i\, \mathcal{N}(\boldsymbol{x}_r|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}_{S(\boldsymbol{\theta})} \Bigg]. \qquad (9)$$

Now, since no g factors appear in R_{ij} nor in $\mathcal{N}(\boldsymbol{x}_r|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, the minimization with respect to them can be done explicitly. However, not all g_i's are independent, since they must sum to 1. We thus re-define $S(\boldsymbol{\theta})$ as

$$S(\boldsymbol{\theta}) \equiv \sum_{i=1}^{K}\sum_{j=1}^{K} g_i\, g_j\, R_{ij} - \frac{2}{N}\sum_{r=1}^{N}\sum_{i=1}^{K} g_i\, \mathcal{N}(\boldsymbol{x}_r|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + 2\lambda\,(1 - \sum_{i=1}^{K} g_i), \tag{10}$$

where λ is a Lagrange multiplier, and proceed with the minimization as if all g's were independent:

$$\begin{aligned}
\left.\frac{\partial S(\boldsymbol{\theta})}{\partial g_i}\right|_{\hat{\boldsymbol{\theta}}} &= 2\Big[\sum_{j=1}^{K}\hat{R}_{ij}\,\hat{g}_j - \frac{1}{N}\sum_{r=1}^{N}\mathcal{N}(\boldsymbol{x}_r|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) - \lambda\Big] = 0 \\
\Rightarrow\quad \hat{g}_i &= \sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\Big[\frac{1}{N}\sum_{r=1}^{N}\mathcal{N}(\boldsymbol{x}_r|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) + \hat{\lambda}\Big]; \qquad (11) \\
\sum_{i=1}^{K}\hat{g}_i &= \sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\Big[\frac{1}{N}\sum_{r=1}^{N}\mathcal{N}(\boldsymbol{x}_r|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) + \hat{\lambda}\Big] = 1 \\
\Rightarrow\quad \hat{\lambda} &= \frac{1 - \dfrac{1}{N}\displaystyle\sum_{r=1}^{N}\Big[\sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\,\mathcal{N}(\boldsymbol{x}_r|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\Big]}{\displaystyle\sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{R}^{-1})_{ij}}. \qquad (12)
\end{aligned}$$

The above result can be put in a more friendly format by defining the *un*-normalized version of g_i, as follows:

$$\hat{g}_i = \hat{g}_{i,\mathrm{un}} + \hat{\lambda}\sum_{j=1}^{K}(\hat{R}^{-1})_{ij}, \tag{13}$$

$$\hat{\lambda} = \left(1 - \sum_{i=1}^{K}\hat{g}_{i,\mathrm{un}}\right)\Big/\sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{R}^{-1})_{ij}, \tag{14}$$

$$\begin{aligned}
\hat{g}_{i,\mathrm{un}} &= \sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\left[\frac{1}{N}\sum_{r=1}^{N}\mathcal{N}(\boldsymbol{x}_r|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\right] \\
&= \sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\Big[\text{sample mean of }\ \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\Big] \\
&\equiv \sum_{j=1}^{K}(\hat{R}^{-1})_{ij}\Big\langle\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\Big\rangle. \qquad (15)
\end{aligned}$$

We now need some derivative results:[1]

$$
\begin{aligned}
\frac{\partial \boldsymbol{\Sigma}_{i,ab}}{\partial \boldsymbol{\Sigma}_{j,k\ell}} &= \delta_{ij}\,\delta_{ak}\,\delta_{bl}, && (16)\\
\frac{\partial (\boldsymbol{\Sigma}_i^{-1})_{ab}}{\partial \boldsymbol{\Sigma}_{j,k\ell}} &= -\,\delta_{ij}\,(\boldsymbol{\Sigma}_j^{-1})_{ak}\,(\boldsymbol{\Sigma}_j^{-1})_{\ell b}, && (17)\\
\frac{\partial \ln|\boldsymbol{\Sigma}_i|}{\partial \boldsymbol{\Sigma}_{j,k\ell}} &= \delta_{ij}\,(\boldsymbol{\Sigma}_j^{-1})_{\ell k}, && (18)\\
\frac{\partial (\boldsymbol{\mu}_i^T\,\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i)}{\partial \boldsymbol{\Sigma}_{j,k\ell}} &= -\,\delta_{ij}\,(\boldsymbol{\mu}_j^T\,\boldsymbol{\Sigma}_j^{-1})_k\,(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j)_\ell, && (19)\\
\frac{\partial \boldsymbol{\Sigma}_{ij,ab}}{\partial \boldsymbol{\Sigma}_{\ell,mn}} &= (\delta_{i\ell}+\delta_{j\ell})\,(\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_\ell^{-1})_{am}\,(\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\Sigma}_{ij})_{nb}, && (20)\\
\frac{\partial (\boldsymbol{\Sigma}_{ij}^{-1})_{ab}}{\partial \boldsymbol{\Sigma}_{\ell,mn}} &= -\,(\delta_{i\ell}+\delta_{j\ell})\,(\boldsymbol{\Sigma}_\ell^{-1})_{am}\,(\boldsymbol{\Sigma}_\ell^{-1})_{nb}, && (21)\\
\frac{\partial \ln|\boldsymbol{\Sigma}_{ij}|}{\partial \boldsymbol{\Sigma}_{\ell,mn}} &= (\delta_{i\ell}+\delta_{j\ell})\,(\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_\ell^{-1})_{nm}, && (22)\\
\frac{\partial (\boldsymbol{\mu}_{ij}^T\,\boldsymbol{\Sigma}_{ij}^{-1}\boldsymbol{\mu}_{ij})}{\partial \boldsymbol{\Sigma}_{\ell,mn}} &= -\,(\delta_{i\ell}+\delta_{j\ell})\Big[(\boldsymbol{\mu}_\ell^T\,\boldsymbol{\Sigma}_\ell^{-1})_m\,(\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_{ij})_n \\
&+ (\boldsymbol{\mu}_{ij}^T\,\boldsymbol{\Sigma}_\ell^{-1})_m\,(\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell)_n - (\boldsymbol{\mu}_{ij}^T\,\boldsymbol{\Sigma}_\ell^{-1})_m\,(\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_{ij})_n\Big], && (23)\\
\frac{\partial f(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,-1)}{\partial \boldsymbol{\mu}_k} &= -\,\delta_{ik}\,f(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,-1)\,\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k, && (24)\\
\frac{\partial f(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i,-1)}{\partial \boldsymbol{\Sigma}_{k,\ell m}} &= -\frac{1}{2}\,\delta_{ik}\,f(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,-1)\Big[(\boldsymbol{\Sigma}_k^{-1})_{\ell m} - (\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)_\ell\,(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)_m\Big], && (25)\\
\frac{\partial f(\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij},+1)}{\partial \boldsymbol{\mu}_k} &= (\delta_{ik}+\delta_{jk})\,f(\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij},+1)\,\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_{ij}, && (26)\\
\frac{\partial f(\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij},+1)}{\partial \boldsymbol{\Sigma}_{k,\ell m}} &= \frac{1}{2}\,(\delta_{ik}+\delta_{jk})\,f(\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij},+1)\Big[(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_k^{-1})_{\ell m} \\
&+ (\boldsymbol{\mu}_{ij}^T\,\boldsymbol{\Sigma}_k^{-1})_\ell\,(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_{ij})_m - (\boldsymbol{\mu}_k^T\,\boldsymbol{\Sigma}_k^{-1})_\ell\,(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_{ij})_m \\
&- (\boldsymbol{\mu}_{ij}^T\,\boldsymbol{\Sigma}_k^{-1})_\ell\,(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)_m\Big]. && (27)
\end{aligned}
$$

[1]Some of these results apply only to symmetric matrices, whereas others are more general. Since all matrices involved here are symmetric, the distinction is not important.

Using these, the minimization with respect to the mean vectors results in:

$$\left.\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_i}\right|_{\hat{\boldsymbol{\theta}}} = -\,2\,\hat{g}_i\,\hat{\boldsymbol{\Sigma}}_i^{-1}\left\{\hat{\lambda}\,\hat{\boldsymbol{\mu}}_i - \Big[\sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\,\hat{\boldsymbol{\mu}}_{ij} - \Big\langle \boldsymbol{x}\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle\Big]\right\} = 0, \tag{28}$$

$$\Rightarrow\quad \hat{\boldsymbol{\mu}}_i = \frac{1}{\hat{\lambda}}\left[\sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\,\hat{\boldsymbol{\mu}}_{ij} - \Big\langle \boldsymbol{x}\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle\right], \qquad (\text{if } \hat{\lambda}\neq 0), \tag{29}$$

$$\Rightarrow\quad \hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\Sigma}}_i\Big[\sum_{j=1}^{K}\hat{g}_j\hat{R}_{ij}\,\hat{\boldsymbol{\Sigma}}_{ij}\Big]^{-1}\left\{\Big\langle \boldsymbol{x}\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle - \sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\,\hat{\boldsymbol{\Sigma}}_{ij}\,\hat{\boldsymbol{\Sigma}}_j^{-1}\hat{\boldsymbol{\mu}}_j\right\}, \tag{30}$$

the last equation being valid when $\hat{\lambda}=0$. Also,

$$\begin{aligned}\left.\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_{i,\ell m}}\right|_{\hat{\boldsymbol{\theta}}} &= \hat{g}_k\left\{\sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\left[(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\Sigma}}_{ij}\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + (\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_m\right]\right.\\ &\left.-\,\hat{\lambda}\left[(\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + (\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i)_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i)_m\right] - \Big\langle(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_m\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle\right\}.\end{aligned} \tag{31}$$

Thus, if $\hat{\lambda}\neq 0$,

$$\begin{aligned}(\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} &= \frac{1}{\hat{\lambda}}\left\{\sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\left[(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\Sigma}}_{ij}\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + (\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_m\right]\right.\\ &\left.-\;\Big\langle(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_m\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle\right\} - (\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i)_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i)_m.\end{aligned} \tag{32}$$

However, if $\hat{\lambda}=0$,

$$\begin{aligned}&\sum_{j=1}^{K}\hat{g}_j\,\hat{R}_{ij}\left[(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\Sigma}}_{ij}\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + (\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_{ij})_m\right]\\ &\qquad = \Big\langle(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_\ell\,(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})_m\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle,\end{aligned} \tag{33}$$

from which we obtain the following two, mathematically equivalent, solutions:

$$\hat{\boldsymbol{\Sigma}}_i^{-1} = \Big\langle(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})\otimes(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle\,\boldsymbol{\Phi}_i^{-1} \quad\text{and} \tag{34}$$

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{\Phi}_i\Big\langle(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})\otimes(\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{x})\,\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\Big\rangle^{-1}. \tag{35}$$

Which of these solutions to use is a matter of computational stability. In any case, $\boldsymbol{\Phi}_i$ is a matrix whose components are defined by

$$(\boldsymbol{\Phi}_i)_{\ell m} = \sum_{j=1}^{K} \hat{g}_j\, \hat{R}_{ij} \left[(\hat{\boldsymbol{\Sigma}}_{ij} \hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + \hat{\boldsymbol{\mu}}_{ij,\ell}\, (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_{ij})_m \right]. \tag{36}$$

The Iterative Algorithm

Given *old* values for $\{\hat{g}_i, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i^{-1}\}$, $1 \le i \le K$, the *new* values should be computed in the following order:

$$\boldsymbol{\Sigma}_{ij} = \left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)^{-1}, \qquad \boldsymbol{\mu}_{ij} = \boldsymbol{\Sigma}_{ij} \left(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j\right), \tag{37}$$

$$R_{ij} = (2\pi)^{-d/2}\, f(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, -1)\, f(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, -1)\, f(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}, +1), \tag{38}$$

$$\hat{g}_{i,\mathrm{un}} = \sum_{j=1}^{K} (\hat{R}^{-1})_{ij} \left\langle \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \right\rangle, \tag{39}$$

$$\hat{\lambda} = \left(1 - \sum_{i=1}^{K} \hat{g}_{i,\mathrm{un}}\right) \Big/ \sum_{i=1}^{K} \sum_{j=1}^{K} (\hat{R}^{-1})_{ij}, \qquad \hat{g}_i = \hat{g}_{i,\mathrm{un}} + \hat{\lambda} \sum_{j=1}^{K} (\hat{R}^{-1})_{ij} \tag{40}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{\hat{\lambda}} \left[\sum_{j=1}^{K} \hat{g}_j\, \hat{R}_{ij}\, \hat{\boldsymbol{\mu}}_{ij} - \left\langle \boldsymbol{x}\, \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) \right\rangle \right], \qquad (\text{if } \hat{\lambda} \neq 0) \tag{41}$$

$$\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\Sigma}}_i \Big[\sum_{j=1}^{K} \hat{g}_j \hat{R}_{ij}\, \hat{\boldsymbol{\Sigma}}_{ij} \Big]^{-1} \left\{ \left\langle \boldsymbol{x}\, \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) \right\rangle - \sum_{j=1}^{K} \hat{g}_j\, \hat{R}_{ij}\, \hat{\boldsymbol{\Sigma}}_{ij}\, \hat{\boldsymbol{\Sigma}}_j^{-1} \hat{\boldsymbol{\mu}}_j \right\} \tag{42}$$

$$\begin{aligned} (\hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} &= \frac{1}{\hat{\lambda}} \left\{ \sum_{j=1}^{K} \hat{g}_j\, \hat{R}_{ij} \left[(\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\Sigma}}_{ij} \hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_{ij})_\ell\, (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_{ij})_m \right] \quad (\text{if } \hat{\lambda} \neq 0) \right. \\ &- \left. \left\langle (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x})_\ell\, (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x})_m\, \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) \right\rangle \right\} - (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_i)_\ell\, (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_i)_m, \end{aligned} \tag{43}$$

$$(\boldsymbol{\Phi}_i)_{\ell m} = \sum_{j=1}^{K} \hat{g}_j\, \hat{R}_{ij} \left[(\hat{\boldsymbol{\Sigma}}_{ij} \hat{\boldsymbol{\Sigma}}_i^{-1})_{\ell m} + \hat{\boldsymbol{\mu}}_{ij,\ell}\, (\hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\mu}}_{ij})_m \right], \tag{44}$$

$$\hat{\boldsymbol{\Sigma}}_i^{-1} = \left\langle (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x}) \otimes (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x})\, \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) \right\rangle \boldsymbol{\Phi}_i^{-1} \quad \text{or} \tag{45}$$

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{\Phi}_i \left\langle (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x}) \otimes (\hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x})\, \mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i) \right\rangle^{-1}, \quad (\text{if } \hat{\lambda} = 0). \tag{46}$$

As a safety procedure to improve stability, any matrix supposed to be symmetric should be numerically symmetrized. ■