A Minimum Divergence Recipe for VBEM

Niko Brümmer

October 14, 2010

We derive a new way to perform minimum-divergence model updates (M-step) in the VBEM algorithm. As example we work out some details of how to apply it to Patrick Kenny's (Odyssey 2010) heavy-tailed PLDA speaker recognition model.

1 VBEM in a nutshell

In VBEM, we work with L, a *lowerbound* to the model log-likelihood:

$$L = \int Q(h) \log \frac{P(\mathbf{x}, h|\boldsymbol{\lambda})}{Q(h)} \, dh \leq \log P(\mathbf{x}|\boldsymbol{\lambda}) \tag{1.1}$$

Here $\mathbf{x}$ is observed training data, $\boldsymbol{\lambda}$ is the to-be-optimized model, h represents all of the hidden variables and Q is any probability density over h. The VBEM algorithm alternates between the E-step, which finds Q subject to constraints, to maximize L while $\boldsymbol{\lambda}$ is fixed; and the M-step, which for a fixed Q, finds $\boldsymbol{\lambda}$ to maximize L. Here we discuss only the M-step, which we augment with a minimum-divergence strategy.

2 M-step with minimum divergence

2.1 Model assumptions

We work with a model of the form $\boldsymbol{\lambda} = (\mathbf{V}, \boldsymbol{\Pi})$, such that $P(\mathbf{x}, h|\boldsymbol{\lambda}) = P(\mathbf{x}|h, \mathbf{V})P(h|\boldsymbol{\Pi})$. We also assume that the model is *overparametrized*, in the following sense. We assume:

- For any prior parameter $\boldsymbol{\Pi}$, there is a *standard form*, denoted $\bar{\boldsymbol{\Pi}}$. For example, if $\boldsymbol{\Pi} = (a, b)$ are the parameters of a gamma distribution, then $\bar{\boldsymbol{\Pi}} = (a, a)$.

- There is a change of variables, $\tilde{h} = \phi(h)$, such that if h is distributed with parameter $\mathbf{\Pi}$, then $\tilde{h}$ is distributed with parameter $\bar{\mathbf{\Pi}}$.
- For any such transformation ϕ, and any $\mathbf{V}$, there exists a $\tilde{\mathbf{V}}$, such that:

  $$P\big(\mathbf{x}|\phi^{-1}(\tilde{h}), \mathbf{V}) = P(\mathbf{x}|\tilde{h}, \tilde{\mathbf{V}}) \tag{2.1}$$

  for every $\mathbf{x}$.

The goal is to maximize L, w.r.t. both $\mathbf{V}$ and $\mathbf{\Pi}$, subject to the constraint that $\mathbf{\Pi}$ ends up in standard form.

2.2 Lower bound decomposition

We decompose L as:

$$\begin{aligned} L &= \int Q(h) \log \frac{P(\mathbf{x}|h, \mathbf{V}) P(h|\mathbf{\Pi})}{Q(h)} \, dh && (2.2) \\ &= \int Q(h) \log P(\mathbf{x}|h, \mathbf{V}) \, dh - \int Q(h) \log \frac{Q(h)}{P(h|\mathbf{\Pi})} \, dh && (2.3) \\ &= \int Q(h) \log P(\mathbf{x}|h, \mathbf{V}) \, dh - D\big(Q(h) \| P(h|\mathbf{\Pi})\big) && (2.4) \end{aligned}$$

where the second term is the KL-divergence from the posterior to the prior for the hidden variables.

2.3 M-step recipe

We assume Q has been chosen by means of the E-step and is given here. The M-step can be broken down into the following sub-steps:

1. Minimize the second term (divergence) of L w.r.t. $\mathbf{\Pi}$, to find $\mathbf{\Pi}_{md}$.
2. Find an invertible transformation, ϕ, so that if $\tilde{h} = \phi(h)$, then

   $$P(\tilde{h}|\phi, \mathbf{\Pi}_{md}) = \tfrac{1}{J} P\big(\phi^{-1}(\tilde{h})\big|\mathbf{\Pi}_{md}\big) = P(\tilde{h}|\bar{\mathbf{\Pi}}_{md}) \tag{2.5}$$

   where $J = |\det(\mathbf{J})|$, is the absolute value of the determinant of the Jacobian of the transformation ϕ. Also transform Q:

   $$\tilde{Q}(\tilde{h}) = \tfrac{1}{J} Q\big(\phi^{-1}(\tilde{h})\big) \tag{2.6}$$

   If we work with a conjugate prior, then Q is of the same form as the prior and transforms in a similar way. Notice that since this step is just a change variables of the divergence integral (with $dh = \frac{1}{J} d\tilde{h}$), it keeps the value of the divergence term unchanged.

3. Now also apply the same change of variables to the first term of L, and then plug in (2.6) and (2.1):

$$\begin{aligned} &\int Q(h)\log P(\mathbf{x}|h,\mathbf{V})\,dh \\ &= \int \tfrac{1}{J}Q\big(\phi^{-1}(\tilde{h})\big)\log P\big(\mathbf{x}\big|\phi^{-1}(\tilde{h}),\mathbf{V}\big)\,d\tilde{h} \\ &= \int \tilde{Q}(\tilde{h})\log P(\mathbf{x}|\tilde{h},\tilde{\mathbf{V}})\,d\tilde{h} \end{aligned} \tag{2.7}$$

   Note that it is enough to know that $\tilde{\mathbf{V}}$ exists—we do not have to also compute $\tilde{\mathbf{V}}$ to satisfy the above equality. We next proceed to maximize the last form w.r.t. $\tilde{\mathbf{V}}$, to give $\mathbf{V}_{ml}$. The information about the change of variables is automatically transferred to $\mathbf{V}_{ml}$ via the transformed posterior—we do not explicitly have to adapt $\mathbf{V}$ to absorb the change of variables as in other variants of minimum divergence.

2.4 M-step result

The end result of the M-step is now $\boldsymbol{\lambda}^* = (\mathbf{V}_{ml}, \bar{\mathbf{\Pi}})$, for which we have:

$$L(Q,\mathbf{V},\mathbf{\Pi}) \leq L(Q,\mathbf{V},\mathbf{\Pi}_{md}) = L(\tilde{Q},\tilde{\mathbf{V}},\bar{\mathbf{\Pi}}_{md}) \leq L(\tilde{Q},\mathbf{V}_{ml},\bar{\mathbf{\Pi}}_{md}) \tag{2.8}$$

as can be seen by following the above steps in order from left to right.

The model $\boldsymbol{\lambda}^*$ and the posterior $\tilde{Q}$ are now both in standard form and ready for another application of the E-step.

3 Example: hierarchical hidden variables

Here we work out some details of how to apply this minimum divergence recipe for Patrick Kenny's (Odyssey 2010) heavy-tailed PLDA speaker recognition model. The recipe has more steps than explained above, because of the hierarchical nature of the model.

3.1 Model

In our notation, the model for a single speaker can be described as:

$$P(\mathbf{x}|\mathbf{y},\boldsymbol{\lambda}) = P(\mathbf{x}|\mathbf{V}\mathbf{y}+\mathbf{m}) \tag{3.1}$$

$$P(\mathbf{y}|s,\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{y}|\mathbf{0},\tfrac{1}{s}\mathbf{I}) \tag{3.2}$$

$$P(s|\boldsymbol{\lambda}) = G(s|a,a) \tag{3.3}$$

where G denotes the *gamma distribution* and $\mathcal{N}$ the multivariate normal distribution, of which details are given in the appendix.

Here we consider just the heavy-tailed distribution of the speaker variable $\mathbf{y}$, which is generated with the help of the hidden 'speaker-scale' variable s. Note $P(\mathbf{x}|\mathbf{V}\mathbf{y}+\mathbf{m})$ may also be a heavy-tailed distribution, and it can be treated similarly, with additional 'channel-scale' hidden variables.

3.2 Lower bound decomposition

The lower bound is a sum over K speakers of the form $L = \sum_{i=1}^{K} L_i$. The contribution due to a single speaker is:

$$\begin{aligned} L_i = & \int Q_i(\mathbf{y}) P(\mathbf{x}_i|\mathbf{V}\mathbf{y}+\mathbf{m})\, \mathbf{dy} \\ & - \int Q_i(s) \left[\int Q_i(\mathbf{y}) \log \frac{Q_i(\mathbf{y})}{\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \frac{1}{s}\boldsymbol{\Sigma})}\, \mathbf{dy} \right] ds \\ & - \int Q_i(s) \log \frac{Q_i(s)}{G(s|a,b)}\, ds \end{aligned} \tag{3.4}$$

where we have plugged in the priors with general (non-standard) parameters. Note that there are three terms here, rather than just two as we had above.

The VB posteriors are of the same form as the priors. The given posterior parameters, are subscripted, which distinguishes them from the to-be-optimized prior parameters. The posteriors are:

$$Q_i(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mu_i, \boldsymbol{\Sigma}_i), \qquad Q_i(s) = G(s|a_i, b_i) \tag{3.5}$$

L is now expressed accordingly as:

$$L = \mathcal{O}_1 - \mathcal{O}_2 - \mathcal{O}_3 \tag{3.6}$$

where:

$$\mathcal{O}_1 = \sum_{i=1}^{K} \int Q_i(\mathbf{y}) P(\mathbf{x}_i|\mathbf{V}\mathbf{y}+\mathbf{m})\, \mathbf{dy} \tag{3.7}$$

$$\mathcal{O}_2 = \sum_{i=1}^{K} \int Q_i(s) D\big(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \| \mathcal{N}(\boldsymbol{\mu}, \tfrac{1}{s}\boldsymbol{\Sigma})\big)\, ds \tag{3.8}$$

$$\mathcal{O}_3 = \sum_{i=1}^{K} D\big(G(a_i, b_i) \| G(a, b)\big) \tag{3.9}$$

3.3 M-step recipe

The M-step optimizes the three components of L in reverse order:

1. We expand $\mathcal{O}_3$ using the formula for gamma divergence (see appendix):

$$\begin{aligned} \mathcal{O}_3 &= \sum_{i=1}^{K} D\big(G(a_i, b_i) \| G(a, b)\big) \\ &= \sum_{i=1}^{K} \log \frac{\Gamma(a)}{\Gamma(a_i)} + a \log \frac{b_i}{b} + \psi(a_i)(a_i - a) + a_i \frac{b - b_i}{b_i} \end{aligned} \tag{3.10}$$

   To minimize, we differentiate first w.r.t. b:

$$\frac{\partial \mathcal{O}_3}{\partial b} = -K\frac{a}{b} + \sum_{i=1}^{K} \frac{a_i}{b_i} \tag{3.11}$$

   which is zeroed at:

$$\frac{a}{b} = \bar{s} = \frac{1}{K} \sum_{i=1}^{K} \frac{a_i}{b_i} \tag{3.12}$$

   Now plug $b = \frac{a}{\bar{s}}$ into $\mathcal{O}_3$ and differentiate w.r.t. a:

$$\frac{\partial \mathcal{O}_3}{\partial a} = \sum_{i=1}^{K} \psi(a) + \log b_i - \log a - 1 + \log \bar{s} - \psi(a_i) + \frac{a_i}{b_i \bar{s}} \tag{3.13}$$

   which is zeroed by solving for a_{md} in:

$$\psi(a_{md}) - \log(a_{md}) = -\log(\bar{s}) + \frac{1}{K} \sum_{i=1}^{K} \psi(a_i) - \log(b_i) \tag{3.14}$$

   after which we recover $b_{md} = \frac{a_{md}}{\bar{s}}$. Keep a_{md} as the new model parameter and use b_{md} below.

2. Next we do a change of variables $\tilde{s} = \alpha s$, so that if $s \sim G(a_{md}, b_{md})$, then $\tilde{s} \sim G(a_{md}, a_{md})$. To do this (see appendix), let $\alpha = \frac{b_{md}}{a_{md}} = \frac{1}{\bar{s}}$. Observe also $s = \bar{s}\tilde{s}$. Transform each $Q_i(s)$:

$$\tilde{Q}_i(\tilde{s}) = \tfrac{1}{\alpha} Q_i\big(\tfrac{1}{\alpha}\tilde{s}\big) = \tfrac{1}{\alpha} G(\tfrac{1}{\alpha}\tilde{s} | a_i, b_i) = G(\tilde{s} | a_i, \tfrac{b_i}{\alpha_i}) = G(\tilde{s} | \tilde{a}_i, \tilde{b}_i) \tag{3.15}$$

   where $\tilde{a}_i = a_i$ and $\tilde{b}_i = \bar{s} b_i$.

3. Notice that in the form $\frac{1}{s}\mathbf{\Sigma}$ a change of scale in s can be absorbed into $\mathbf{\Sigma}$, which satisfies our required modelling assumption. As explained above, we don't need to explicitly do this update to $\mathbf{\Sigma}$ (although it is trivial to do). We can instead proceed directly to find the optimum $\mathbf{\Sigma}$ by minimizing the expected divergence $\mathcal{O}_2$, now expressed in terms of $\tilde{Q}_i(\tilde{s})$:

$$\begin{aligned}\mathcal{O}_2 &= \sum_{i=1}^{K} \int \tilde{Q}_i(\tilde{s}) D\big(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \| \mathcal{N}(\boldsymbol{\mu}, \tfrac{1}{\tilde{s}}\mathbf{\Sigma})\big)\, d\tilde{s} \\ &= \sum_{i=1}^{K} \Big\langle D\big(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \| \mathcal{N}(\boldsymbol{\mu}, \tfrac{1}{\tilde{s}_i}\mathbf{\Sigma})\big) \Big\rangle \\ &= \sum_{i-1}^{K} \Big\langle -\tfrac{N}{2} - \tfrac{1}{2}\log|\tilde{s}_i \mathbf{P}\mathbf{\Sigma}_i| + \tfrac{1}{2}\operatorname{tr}\Big(\tilde{s}_i\mathbf{P}\big(\mathbf{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'\big)\Big)\Big\rangle\end{aligned} \tag{3.16}$$

where we defined the prior precision matrix $\mathbf{P} = \mathbf{\Sigma}^{-1}$. Differentiating w.r.t. $\boldsymbol{\mu}$ and equating to zero gives:

$$\boldsymbol{\mu}_{md} = \frac{\sum_{i=1}^{K} \langle \tilde{s}_i \rangle \boldsymbol{\mu}_i}{\sum_{i=1}^{K} \langle \tilde{s}_i \rangle} = \frac{1}{K} \sum_{i=1}^{K} \langle \tilde{s}_i \rangle \boldsymbol{\mu}_i \tag{3.17}$$

where[1] $\langle \tilde{s}_i \rangle = \frac{\tilde{a}_i}{\tilde{b}_i}$. Now (using the new value of $\boldsymbol{\mu}$) let:

$$\mathbf{C}_i = \langle s_i \rangle \big(\mathbf{\Sigma}_i + (\boldsymbol{\mu}_{md} - \boldsymbol{\mu}_i)(\boldsymbol{\mu}_{md} - \boldsymbol{\mu}_i)'\big) \tag{3.18}$$

Then differentiating $\mathcal{O}_2$ w.r.t $\mathbf{P}$ gives:

$$\begin{aligned} d\mathcal{O}_2 &= \sum_{k=1}^{K} -\tfrac{1}{2} d\log|\mathbf{P}| + \tfrac{1}{2} d\operatorname{tr}(\mathbf{P}\mathbf{C}_i) \\ &= -\tfrac{1}{2}K\operatorname{tr}(\mathbf{P}^{-1}d\mathbf{P}) + \tfrac{1}{2}\sum_{i=1}^{K}\operatorname{tr}(\mathbf{C}_i d\mathbf{P})\end{aligned} \tag{3.19}$$

which is zeroed at:

$$\mathbf{\Sigma}_{md} = \mathbf{P}^{-1} = \tfrac{1}{K} \sum_{i=1}^{K} \mathbf{C}_i \tag{3.20}$$

[1] Here $\langle \tilde{s}_i \rangle$ denotes posterior expectation. Do not confuse with prior expectation, $\langle \tilde{s} \rangle = 1$.

4. Now do a change of variables, $\tilde{\mathbf{y}} = \mathbf{J}\mathbf{y} + \mathbf{k}$, so that if $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{md}, \boldsymbol{\Sigma}_{md})$, then $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We need (see appendix) $\mathbf{J}^{-1} = \text{chol}(\boldsymbol{\Sigma}_{md})$ and $\mathbf{k} = -\mathbf{J}\boldsymbol{\mu}_{md}$. Then use these to transform the posteriors, so that

$$\tilde{Q}_i(\tilde{\mathbf{y}}) = \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i) \tag{3.21}$$

where

$$\tilde{\boldsymbol{\mu}}_i = \mathbf{J}\boldsymbol{\mu}_i + \mathbf{k} = \mathbf{J}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{md}) \tag{3.22}$$

$$\tilde{\boldsymbol{\Sigma}}_i = \mathbf{J}\boldsymbol{\Sigma}_i\mathbf{J}' \tag{3.23}$$

5. Finally, maximize $\mathcal{O}_1$ w.r.t. $(\mathbf{V}, \mathbf{m})$, where we plug in the transformed posteriors:

$$(\mathbf{V}_{ml}, \mathbf{m}_{ml}) = \arg\max_{\mathbf{V},\boldsymbol{\mu}} \int \tilde{Q}(\tilde{\mathbf{y}}) P(\mathbf{x}_i|\mathbf{V}\tilde{\mathbf{y}} + \mathbf{m})\, \mathbf{d}\tilde{\mathbf{y}} \tag{3.24}$$

The end-result of this M-step is $(a_{md}, \mathbf{V}_{ml}, \mathbf{m}_{ml})$.

4 Appendix

4.1 Density transformations

gamma: For a transformation $\tilde{s} = \phi(s) = \alpha s$, where $\alpha > 0$, the Jacobian determinant is $J = \alpha$ and the *gamma density*, G, transforms as:

$$P(s|a,b) = G(s|a,b) = \frac{b^a}{\Gamma(a)} s^{a-1} e^{-bs} \tag{4.1}$$

$$P(\tilde{s}|\phi, a, b) = \tfrac{1}{\alpha} G\big(\tfrac{\tilde{s}}{\alpha}\big|a, b\big) = G\big(\tilde{s}\big|a, \tfrac{b}{\alpha}\big) \tag{4.2}$$

To massage the distribution of $\tilde{s}$ to have a standard distribution, with $\langle \tilde{s} \rangle = \frac{a}{b} = 1$, we need $a = \frac{b}{\alpha}$, or $\alpha = \frac{b}{a}$.

normal: For a transformation $\tilde{\mathbf{y}} = \phi(\mathbf{y}) = \mathbf{J}\mathbf{y} + \mathbf{k}$, the Jacobian is $\mathbf{J}$, and $J = |\det \mathbf{J}|$. The inverse transform is $\mathbf{y} = \phi^{-1}(\tilde{\mathbf{y}}) = \mathbf{J}^{-1}(\tilde{\mathbf{y}} - \mathbf{k})$. The *multivariate normal density* transforms as:

$$\begin{aligned} P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\big(-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\big) \end{aligned} \tag{4.3}$$

$$\begin{aligned} P(\tilde{\mathbf{y}}|\phi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \tfrac{1}{J} \mathcal{N}\big(\phi^{-1}(\tilde{\mathbf{y}})\big|\boldsymbol{\mu}, \boldsymbol{\Sigma}\big) \\ &= \mathcal{N}\big(\tilde{\mathbf{y}}\big|\mathbf{J}\boldsymbol{\mu} + \mathbf{k}, \mathbf{J}\boldsymbol{\Sigma}\mathbf{J}'\big) \end{aligned} \tag{4.4}$$

To massage $\tilde{\mathbf{y}}$ to have standard $N(\mathbf{0}, \mathbf{I})$ distribution, we need[2]: $\mathbf{J}^{-1} = \mathrm{chol}(\boldsymbol{\Sigma})$, and $\mathbf{k} = -\mathbf{J}\boldsymbol{\mu}$.

4.2 KL-divergences

normal: The KL-divergence between two N-dimensional normal distributions is:

$$\begin{aligned} &D\big(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \| \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) \\ &= -\tfrac{N}{2} - \tfrac{1}{2}\log|\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0| + \tfrac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\Sigma}_0 + (\boldsymbol{\mu}_0 - \boldsymbol{\mu})(\boldsymbol{\mu}_0 - \boldsymbol{\mu})'\big)\Big) \end{aligned} \tag{4.5}$$

gamma: The KL-divergence between two gamma distributions is:

$$\begin{aligned} &D\big(G(a_0, b_0) \| G(a, b)\big) \\ &= \log\frac{\Gamma(a)}{\Gamma(a_0)} + a\log\frac{b_0}{b} + \psi(a_0)(a_0 - a) + a_0\frac{b - b_0}{b_0} \end{aligned} \tag{4.6}$$

[2] We use the Cholesky transform definition: $\mathrm{chol}(\boldsymbol{\Sigma})\,\mathrm{chol}(\boldsymbol{\Sigma})' = \boldsymbol{\Sigma}$. Watch out, MATLAB's function chol() returns the *transpose* of this definition!