EM4JFA

Niko Brümmer

October 17, 2009

1 Model

Let GMM i be represented by $\mathbf{m}_i$, the supervector of stacked component means:

$$\mathbf{m}_i = \begin{bmatrix} \mathbf{m}'_{i1} & \mathbf{m}'_{i2} & \cdots \end{bmatrix}'. \tag{1}$$

In other words, $\mathbf{m}_{ij}$ is the mean of component j of GMM i. The GMM's differ from each other in a small subspace, so that:

$$\mathbf{m}_i = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i \tag{2}$$

where $\boldsymbol{\mu}$ is a supervector of similar structure to $\mathbf{m}_i$; where $\mathbf{V}$ is a tall, thin matrix and correspondingly, $\mathbf{y}_i$ is of much smaller dimension than $\mathbf{m}_i$. If we let $\mathbf{V} = \begin{bmatrix} \mathbf{V}'_1 & \mathbf{V}'_2 & \cdots \end{bmatrix}'$, then we can rewrite (2) as:

$$\mathbf{m}_{ij} = \boldsymbol{\mu}_j + \mathbf{V}_j\mathbf{y}_i. \tag{3}$$

In summary, the *model parameter* that we want to estimate via the EM algorithm is $\tilde{\mathbf{V}} = \begin{bmatrix} \mathbf{V} & \boldsymbol{\mu} \end{bmatrix}$; and the *hidden variables* are represented by all the $\mathbf{y}_i$.

1.1 Data

We assume that every GMM i generated a sequence of feature vectors, for which we have sufficient statistics denoted by $\mathbf{S}_i = (\mathbf{N}_i, \mathbf{f}_i)$. The zero-order statistic $\mathbf{N}_i$ is a diagonal matrix of supervector size, with elements n_{ij} on the diagonal. Every n_{ij} is replicated to match the feature dimension, so that the diagonal of $\mathbf{N}_i$ is $\begin{bmatrix} n_{i1} \cdots n_{i1} & n_{i2} \cdots n_{i2} & n_{i3} \cdots & \cdots \end{bmatrix}$. The first-order statistic $\mathbf{f}_i$ is a supervector, with the same structure as $\mathbf{m}_i$.

1.2 Prior

We assume every $\mathbf{y}_i$ is drawn i.i.d. from the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, or

$$p(\mathbf{y}_i) \propto \exp\left(-\frac{1}{2}\mathbf{y}_i'\mathbf{y}_i\right). \tag{4}$$

1.3 Likelihood

Given the statistics $S_i = (\mathbf{N}_i, \mathbf{f}_i)$, we can write the likelihood for GMM $\mathbf{m}_i = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i$ as:

$$p(\mathbf{S}_i|\mathbf{m}_i) \propto \exp\left(\mathbf{f}'\mathbf{m}_i - \frac{1}{2}\mathbf{m}_i'\mathbf{N}_i\mathbf{m}_i\right) \tag{5}$$

or,

$$p(\mathbf{S}_i|\mathbf{y}_i, \tilde{\mathbf{V}}) \propto \exp\left(\mathbf{y}_i'\mathbf{V}'\bar{\mathbf{f}}_i - \frac{1}{2}\mathbf{y}_i'\mathbf{V}'\mathbf{N}_i\mathbf{V}\mathbf{y}_i + \gamma_i\right), \tag{6}$$

where we have defined

$$\bar{\mathbf{f}}_i = \mathbf{f}_i - \mathbf{N}_i\boldsymbol{\mu}, \tag{7}$$

$$\gamma_i = \boldsymbol{\mu}'\mathbf{f}_i - \frac{1}{2}\boldsymbol{\mu}'\mathbf{N}_i\boldsymbol{\mu}. \tag{8}$$

The argument of exp() in (6) is known as the *complete-data log-likelihood* and we denote it as $\lambda(\mathbf{y}_i, \tilde{\mathbf{V}})$. It will be useful to represent it in two different forms:

$$\lambda(\mathbf{y}_i, \tilde{\mathbf{V}}) = \mathbf{y}_i'\mathbf{V}'\bar{\mathbf{f}}_i - \frac{1}{2}\mathbf{y}_i'\mathbf{V}'\mathbf{N}_i\mathbf{V}\mathbf{y}_i + \gamma_i \tag{9}$$

$$= \operatorname{tr}\left(\mathbf{f}_i\tilde{\mathbf{y}}_i'\tilde{\mathbf{V}}'\right) - \frac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{Y}}\tilde{\mathbf{V}}'\mathbf{N}_i\tilde{\mathbf{V}}\right) \tag{10}$$

where

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{y}\mathbf{y}' & \mathbf{y} \\ \mathbf{y}' & 1 \end{bmatrix}, \qquad \tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix}. \tag{11}$$

We shall use both forms. If we are interested in inferring the value of $\mathbf{y}_i$ when $\tilde{\mathbf{V}}$ is given, then (9) is more convenient. However, if we want to infer the value of the parameter $\tilde{\mathbf{V}}$, when $\mathbf{y}_i$ is given, then (10) is more convenient.

1.4 Posterior

The hidden-variable posterior is derived, without solving any integrals, by recognizing that it is proportional to the joint density of the data and the hidden variable and then massaging that expression until we can recognize it as being proportional to a normal distribution. Once we have identified the parameters[1] of the normal distribution, the integral is implicitly solved, because the normalization constant for the normal distribution is known. Here we work with the complete data log-likelihood in form (9), from which we omit the term γ_i because it is independent of $\mathbf{y}_i$:

$$
\begin{aligned}
p(\mathbf{y}_i|\mathbf{S}_i, \tilde{\mathbf{V}}) &\propto p(\mathbf{y}_i, \mathbf{S}_i|\tilde{\mathbf{V}}) = p(\mathbf{S}_i|\mathbf{y}_i, \tilde{\mathbf{V}})p(\mathbf{y}_i) && (12)\\
&\propto \exp\left(\mathbf{y}_i'\mathbf{V}'\bar{\mathbf{f}}_i - \frac{1}{2}\mathbf{y}_i'\mathbf{V}'\mathbf{N}_i\mathbf{V}\mathbf{y}_i\right)\exp(-\frac{1}{2}\mathbf{y}_i'\mathbf{y}_i) && (13)\\
&\propto \exp\left(\mathbf{y}_i'\mathbf{V}'\bar{\mathbf{f}}_i - \frac{1}{2}\mathbf{y}_i'(\mathbf{I} + \mathbf{V}'\mathbf{N}_i\mathbf{V})\mathbf{y}_i\right) && (14)\\
&\propto \exp\left(\mathbf{y}_i'\mathbf{L}_i\hat{\mathbf{y}}_i - \frac{1}{2}\mathbf{y}_i'\mathbf{L}_i\mathbf{y}_i\right) && (15)\\
&\propto \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1}) = p(\mathbf{y}_i|\mathbf{S}_i, \tilde{\mathbf{V}}), && (16)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{L}_i &= \mathbf{I} + \mathbf{V}'\mathbf{N}_i\mathbf{V},\\
\mathbf{L}_i\hat{\mathbf{y}}_i &= \mathbf{V}'\bar{\mathbf{f}}_i.
\end{aligned} \tag{17}
$$

1.5 Marginal (EM Objective)

So far, we have expressions for *prior* $p(\mathbf{y}_i)$, *likelihood* $p(\mathbf{S}_i|\mathbf{y}_i, \tilde{\mathbf{V}})$ and *posterior* $p(\mathbf{y}_i|\mathbf{S}, \tilde{\mathbf{V}})$. Here we complete our tools for this model by deriving also the *marginal* $p(\mathbf{S}_i|\tilde{\mathbf{V}})$ in terms of the others. The straight-forward way would be to integrate $p(\mathbf{y}_i, \mathbf{S}_i|\tilde{\mathbf{V}})$ w.r.t. $\mathbf{y}_i$, but we have already solved this integral in computing the posterior! We can re-use this result by re-arranging Bayes' rule:

$$
p(\mathbf{S}_i|\tilde{\mathbf{V}}) = \frac{p(\mathbf{y}_i, \mathbf{S}_i|\tilde{\mathbf{V}})}{p(\mathbf{y}_i|\mathbf{S}_i, \tilde{\mathbf{V}})} \tag{18}
$$

[1] We make use of the *natural parameters* of the normal distribution, namely the precision matrix (inverse covariance), here denoted $\mathbf{L}$ and the product of the precision and the mean, here denoted $\mathbf{L}\hat{\mathbf{y}}_i$.

This formula is strange because $\mathbf{y}_i$ appears on the RHS, but not on the left. This is because it is valid for any $\mathbf{y}_i$ that has non-zero prior density[2]. This means we can insert any convenient value, which for our purposes is $\mathbf{y}_i = \mathbf{0}$. We compute the logarithm of (18) by using (9) for the numerator and (16) for the denominator:

$$\log p(\mathbf{S}_i|\tilde{\mathbf{V}}) + \text{const} \tag{19}$$

$$= \gamma_i - \log \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1}) + \text{const} \tag{20}$$

$$= \gamma_i - \frac{1}{2}\log|\mathbf{L}_i| + \frac{1}{2}\hat{\mathbf{y}}_i'\mathbf{L}_i\hat{\mathbf{y}}_i. \tag{21}$$

In the context of the EM-algorithm, we can use (21) to compute the objective function that we are maximizing:

$$\mathcal{L}(\tilde{\mathbf{V}}) = \sum_i \gamma_i - \frac{1}{2}\log|\mathbf{L}_i| + \frac{1}{2}\hat{\mathbf{y}}_i'\mathbf{L}_i\hat{\mathbf{y}}_i \tag{22}$$

$$= \sum_i \boldsymbol{\mu}'\mathbf{f}_i - \frac{1}{2}\boldsymbol{\mu}'\mathbf{N}_i\boldsymbol{\mu} - \frac{1}{2}\log|\mathbf{L}_i| + \frac{1}{2}\hat{\mathbf{y}}_i'\mathbf{V}'\bar{\mathbf{f}}_i. \tag{23}$$

2 EM algorithm

In this section we derive formulas for an EM algorithm (with minimum-divergence) for the model described in the previous section. The EM algorithm finds a maximum-likelihood (ML) estimate for the parameter $\tilde{\mathbf{V}}$ of the model. We devote subsections to the E-step, the M-step and the (minim-divergence) MD-step.

2.1 E-step: EM auxiliary

The EM auxiliary is constructed via the hidden-variable posterior (16), given an initial parameter $\tilde{\mathbf{V}}_0$, so that $p(\mathbf{y}_i|\mathbf{S}_i, \tilde{\mathbf{V}}_0) = \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1})$. The auxiliary

[2]If the prior is zero, so is the posterior and we have zero numerator and denominator. But for the present case, this will never happen, because the densities are normal and always non-zero.

is constructed as the expected value of (10):

$$\tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}}) = \left\langle \sum_i \log p(\mathbf{S}_i|\mathbf{y}_i, \tilde{\mathbf{V}}) + \text{const} \right\rangle_{\prod_i \mathcal{N}(\mathbf{y}|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1})} \tag{24}$$

$$= \sum_i \left\langle \log p(\mathbf{S}_i|\mathbf{y}_i, \tilde{\mathbf{V}}) + \text{const} \right\rangle_{\mathcal{N}(\mathbf{y}|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1})} \tag{25}$$

$$= \sum_i \left\langle \operatorname{tr}\left(\mathbf{f}_i \tilde{\mathbf{y}}_i' \tilde{\mathbf{V}}'\right) - \frac{1}{2} \operatorname{tr}\left(\tilde{\mathbf{Y}} \tilde{\mathbf{V}}' \mathbf{N}_i \tilde{\mathbf{V}}\right) \right\rangle_{\mathcal{N}(\mathbf{y}|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1})} \tag{26}$$

$$= \sum_i \operatorname{tr}(\hat{\mathbf{X}}_i \mathbf{V}') - \frac{1}{2} \operatorname{tr}\left(\hat{\mathbf{Y}}_i \tilde{\mathbf{V}}' \mathbf{N}_i \tilde{\mathbf{V}}\right) \tag{27}$$

where

$$\hat{\mathbf{X}}_i = \mathbf{f}_i \begin{bmatrix} \hat{\mathbf{y}}_i' & 1 \end{bmatrix}, \qquad \hat{\mathbf{Y}}_i = \begin{bmatrix} \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i' + \mathbf{L}_i^{-1} & \hat{\mathbf{y}}_i \\ \hat{\mathbf{y}}_i' & 1 \end{bmatrix}. \tag{28}$$

It is important to emphasize that in this context, the posterior parameters $\hat{\mathbf{y}}_i$ and $\mathbf{L}_i$ are functions of $\tilde{\mathbf{V}}_0$ rather than of $\tilde{\mathbf{V}}$. This means that when we maximize (27) directly below w.r.t. $\tilde{\mathbf{V}}$, then $\hat{\mathbf{X}}_i$ and $\hat{\mathbf{Y}}_i$ are constants.

2.2 M-step

The M-step is a procedure to choose a new parameter $\tilde{\mathbf{V}}$, so that given an initial parameter $\tilde{\mathbf{V}}_0$, there is progress in the objective: $\mathcal{L}(\tilde{\mathbf{V}}) > \mathcal{L}(\tilde{\mathbf{V}}_0)$. The EM-auxiliary $\tilde{\mathcal{Q}}$ is constructed such that this is guaranteed, provided we can ensure $\tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}}) > \tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}}_0)$. The M-step achieves this by maximizing (27) w.r.t. $\tilde{\mathbf{V}}$. If we disassemble each $\hat{\mathbf{X}}_i$ and $\tilde{\mathbf{V}}$ into stacked submatrices, analogous to the structure defined for $\mathbf{V}$, we can differentiate (27) separately w.r.t. each submatrix $\tilde{\mathbf{V}}_j$:

$$\frac{\mathsf{d}\tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}})}{\mathsf{d}\tilde{\mathbf{V}}_j} = \sum_i \hat{\mathbf{X}}_{ij}' - \left(\sum_i n_{ij} \hat{\mathbf{Y}}_i\right) \tilde{\mathbf{V}}_j' \tag{29}$$

which can be equated to the zero matrix and then solved for $\tilde{\mathbf{V}}_j$.

2.3 MD-step

The MD-step, where MD denotes *minimum-divergence* is a procedure to choose a new parameter $\tilde{\mathbf{V}}$, so that given an initial parameter $\tilde{\mathbf{V}}_0$, there is progress in the objective: $\mathcal{L}(\tilde{\mathbf{V}}) \geq \mathcal{L}(\tilde{\mathbf{V}}_0)$. This is the same agenda as the regular M-step, but it is complementary to it.

2.3.1 Minimizing divergence

Here we need to minimize the divergence from the hidden-variable posterior to the prior. The multivariate normal prior is temporarily allowed to have non-standard parameters and the minimization is w.r.t. these prior parameters. Denote by $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \cdots\end{bmatrix}$, all of the hidden variables and by $\mathbf{S} = \begin{bmatrix}\mathbf{S}_1 & \mathbf{S}_2 & \cdots\end{bmatrix}$ all of the data, then prior and posterior are respectively:

$$p(\mathbf{Y}|\bar{\mathbf{y}}, \mathbf{P}) = \prod_i \mathcal{N}(\mathbf{y}_i|\bar{\mathbf{y}}, \mathbf{P}^{-1}) \tag{30}$$

$$p(\mathbf{Y}|\mathbf{S}, \tilde{\mathbf{V}}_0) = \prod_i \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1}) \tag{31}$$

where we have introduced the new prior parameters: mean $\bar{\mathbf{y}}$ and precision $\mathbf{P}$. The divergence[3], expressed in terms of these parameters is:

$$D(\bar{\mathbf{y}}, \mathbf{P}) = \left\langle \log \frac{p(\mathbf{Y}|\mathbf{S}, \tilde{\mathbf{V}}_0)}{p(\mathbf{Y}|\bar{\mathbf{y}}, \mathbf{P})} \right\rangle_{p(\mathbf{Y}|\mathbf{S}, \tilde{\mathbf{V}}_0)} \tag{32}$$

$$= \sum_i \left\langle \log \frac{\mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{L}_i^{-1})}{\mathcal{N}(\mathbf{y}_i|\bar{\mathbf{y}}, \mathbf{P}^{-1})} \right\rangle_{\mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}, \mathbf{L}_i^{-1})} \tag{33}$$

$$= \frac{1}{2} \sum_i -\log|\mathbf{P}| + \mathrm{tr}(\mathbf{L}_i^{-1}\mathbf{P}) + (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})'\mathbf{P}(\hat{\mathbf{y}}_i - \bar{\mathbf{y}}) + \mathrm{const} \tag{34}$$

$$= \frac{1}{2}(\mathrm{tr}(\mathbf{SP})) - N \log|\mathbf{P}|) + \mathrm{const}, \tag{35}$$

where N is the number of items summed over and

$$\mathcal{S} = \sum_i \mathbf{L}_i^{-1} + (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})' \tag{36}$$

This is minimized by using matrix calculus. First, we differentiate (34) to give:

$$\frac{\mathsf{d}D(\bar{\mathbf{y}}, \mathbf{P})}{\mathsf{d}\bar{\mathbf{y}}} = \sum_i (\bar{\mathbf{y}} - \hat{\mathbf{y}}_i)'\mathbf{P}, \tag{37}$$

which can be zeroed, independently of $\mathbf{P}$, by choosing $\bar{\mathbf{y}} = \frac{1}{N}\sum_i \hat{\mathbf{y}}_i$. Next, we differentiate (35) w.r.t. $\mathbf{P}$:

$$\frac{\mathsf{d}D(\bar{\mathbf{y}}, \mathbf{P})}{\mathsf{d}\mathbf{P}} = \frac{1}{2}(\mathcal{S} - N\mathbf{P}^{-1}), \tag{38}$$

[3]The formula for divergence between normal distributions can be found e.g. here: `http://en.wikipedia.org/wiki/Multivariate_normal_distribution`.

which is zeroed at $\mathbf{P}^{-1} = \frac{1}{N}\mathcal{S}$. Note that $\mathcal{S}$ is dependent on $\bar{\mathbf{y}}$, which should therefore be computed first. In summary, the result of this step is:

$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{y}}_i, \\ \mathbf{P}^{-1} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{L}_i^{-1} + (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})'. \end{aligned} \tag{39}$$

2.3.2 Hidden variable transformation

Let $\phi(\mathbf{y}) = \mathbf{J}^{-1}(\mathbf{y} - \bar{\mathbf{y}})$, where $\mathbf{J}$ is the Cholesky decomposition of $\mathbf{P}^{-1}$, so that $\mathbf{J}\mathbf{J}' = \mathbf{P}^{-1}$. The Jacobian of ϕ is $\mathbf{J}^{-1}$ and the inverse transformation is $\phi^{-1}(\mathbf{y}) = \mathbf{J}\mathbf{y} + \bar{\mathbf{y}}$. Now verify that if $\mathbf{y} \sim \mathcal{N}(\bar{\mathbf{y}}, \mathbf{P}^{-1})$, then $\phi(\mathbf{y})$ is distributed as:

$$\frac{1}{|\mathbf{J}^{-1}|}\mathcal{N}(\phi^{-1}(\mathbf{y})|\bar{\mathbf{y}}, \mathbf{P}^{-1}) \tag{40}$$

$$= |\mathbf{J}|\frac{1}{\sqrt{|2\pi\mathbf{P}^{-1}|}}\exp\left(-\frac{1}{2}(\phi^{-1}(\mathbf{y}) - \bar{\mathbf{y}})'\mathbf{P}(\phi^{-1}(\mathbf{y}) - \bar{\mathbf{y}})\right) \tag{41}$$

$$= \frac{1}{\sqrt{|2\pi\mathbf{I}|}}\exp(-\frac{1}{2}\mathbf{y}'\mathbf{y}) \tag{42}$$

$$= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{I}). \tag{43}$$

Transformation ϕ therefore returns the prior to standard form. Now we just need to apply the transformation also to $\tilde{\mathbf{V}}$:

2.3.3 Parameter transformation

This step starts with initial parameter value $\tilde{\mathbf{V}}_1 = \begin{bmatrix}\mathbf{V}_1, \boldsymbol{\mu}_1\end{bmatrix}$, such that $\tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}}_1) \geq \tilde{\mathcal{Q}}(\tilde{\mathbf{V}}_0, \tilde{\mathbf{V}}_0)$. For example, $\tilde{\mathbf{V}}_1$ can be the output of the M-step, or we can just use $\tilde{\mathbf{V}}_1 = \tilde{\mathbf{V}}_0$. The MD-step requires that we find $\tilde{\mathbf{V}} = \begin{bmatrix}\mathbf{V} & \boldsymbol{\mu}\end{bmatrix}$, such that for every $\mathbf{y}$:

$$\boldsymbol{\mu} + \mathbf{V}\mathbf{y} = \boldsymbol{\mu}_1 + \mathbf{V}_1\phi^{-1}(\mathbf{y}) \tag{44}$$

$$= \mathbf{V}_1\mathbf{J}\mathbf{y} + (\mathbf{V}_1\bar{\mathbf{y}} + \boldsymbol{\mu}_1) \tag{45}$$

or

$$\begin{aligned} \mathbf{V} &= \mathbf{V}_1\mathbf{J}, \\ \boldsymbol{\mu} &= \mathbf{V}_1\bar{\mathbf{y}} + \boldsymbol{\mu}_1. \end{aligned} \tag{46}$$

In summary, the final result of the whole MD-step is (46) where $\mathbf{J}$ is the Cholesky decomposition of $\mathbf{P}^{-1}$ and where $\bar{\mathbf{y}}$ and $\mathbf{P}$ are defined in (39).