EM for Probabilistic LDA

Niko Brümmer

February 2010

1 Model

Let observation j of speaker i be $\mathbf{m}_{ij}$ and let it be modeled as:

$$\mathbf{m}_{ij} = \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij} + \mathbf{z}_{ij} \tag{1}$$

where

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$
$$\mathbf{x}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$
$$\mathbf{z}_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}^{-1}) \tag{4}$$

where the dimensions of $\mathbf{x}$ and $\mathbf{y}$ may be smaller than that of $\mathbf{m}$ and where $\mathbf{D}$ is a diagonal precision matrix. The *model parameter* that we want to estimate via the EM algorithm is $\boldsymbol{\lambda} = (\mathbf{V}, \mathbf{U}, \mathbf{D})$; and the *hidden variables* are represented by all the $\mathbf{y}_i$ and $\mathbf{x}_{ij}$. Note that $\mathbf{z}_{ij}$ is not also hidden, because if $\mathbf{m}_{ij}$, $\mathbf{y}_i$ and $\mathbf{x}_{ij}$ are given, then $\mathbf{z}_{ij}$ is determined.

1.1 Data

We are given N observations of the form $\mathbf{m}_{ij}$, for K speakers, so that $i = 1 \cdots K$. There are n_i observations per speaker, so that $j = 1 \cdots n_i$. We denote the matrix of all the observations for speaker i as $\mathbf{M}_i = \begin{bmatrix}\mathbf{m}_{i1} \cdots \mathbf{m}_{in_i}\end{bmatrix}$.

The *zero-order statistic* for speaker i is n_i and the global zero-order statsistic is $N = \sum_{i=1}^{K} n_i$. The *first-order statistic* for speaker i is:

$$\mathbf{f}_i = \sum_{j=1}^{n_i} \mathbf{m}_{ij} \tag{5}$$

and the global *second-order statistic* is:

$$\mathbf{S} = \sum_{ij} \mathbf{m}_{ij}\mathbf{m}'_{ij}. \tag{6}$$

1.2 Prior

The joint prior for the hidden variables for a speaker i is:

$$p(\mathbf{y}_i, \mathbf{X}_i) = p(\mathbf{y}_i)p(\mathbf{X}_i) \propto \exp\left(-\frac{1}{2}\mathbf{y}_i'\mathbf{y}_i - \frac{1}{2}\operatorname{tr}(\mathbf{X}_i'\mathbf{X}_i)\right), \tag{7}$$

where $\mathbf{X}_i = \left[\mathbf{x}_{i1} \cdots \mathbf{x}_{in_i}\right]$.

1.3 Likelihood

The complete-data log-likelihood, for speaker i is:

$$p(\mathbf{M}_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\lambda}) = \prod_{j=1}^{n_i} \mathcal{N}(\mathbf{m}_{ij}|\mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_{ij}, \mathbf{D}^{-1}) \tag{8}$$

$$\propto \exp\left(-\frac{1}{2}\sum_{j=1}^{n_i}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i - \mathbf{U}\mathbf{x}_{ij})'\mathbf{D}(\mathbf{m}_{ij} - \mathbf{V}\mathbf{y}_i - \mathbf{U}\mathbf{x}_{ij})\right) \tag{9}$$

$$\propto \exp\sum_{j=1}^{n_i}\Big(-\frac{1}{2}\mathbf{m}_{ij}'\mathbf{D}\mathbf{m}_{ij} + \mathbf{m}_{ij}'\mathbf{D}\mathbf{V}\mathbf{y}_i + \mathbf{m}_{ij}'\mathbf{D}\mathbf{U}\mathbf{x}_{ij} - \frac{1}{2}\mathbf{y}_i'\mathbf{V}'\mathbf{D}\mathbf{V}\mathbf{y}_i - \mathbf{y}_i'\mathbf{V}'\mathbf{D}\mathbf{U}\mathbf{x}_{ij} - \frac{1}{2}\mathbf{x}_{ij}'\mathbf{U}'\mathbf{D}\mathbf{U}\mathbf{x}_{ij}\Big) \tag{10}$$

1.4 Joint

$$p(\mathbf{M}_i, \mathbf{y}_i, \mathbf{X}_i|\boldsymbol{\lambda}) \tag{11}$$

$$\propto \exp\left(-\frac{1}{2}\mathbf{y}_i'\mathbf{L}_i\mathbf{y}_i + \sum_{j=1}^{n_i} -\frac{1}{2}\mathbf{m}_{ij}'\mathbf{D}\mathbf{m}_{ij} + \mathbf{m}_{ij}'\mathbf{D}\mathbf{V}\mathbf{y}_i + \mathbf{m}_{ij}'\mathbf{D}\mathbf{U}\mathbf{x}_{ij} - \mathbf{x}_{ij}'\mathbf{J}\mathbf{y}_i - \frac{1}{2}\mathbf{x}_{ij}'\mathbf{K}\mathbf{x}_{ij}\right) \tag{12}$$

where

$$\mathbf{J} = \mathbf{U}'\mathbf{D}\mathbf{V} \tag{13}$$

$$\mathbf{K} = \mathbf{U}'\mathbf{D}\mathbf{U} + \mathbf{I} \tag{14}$$

$$\mathbf{L}_i = n_i\mathbf{V}'\mathbf{D}\mathbf{V} + \mathbf{I} \tag{15}$$

1.5 Posterior

We assemble the joint posterior from two factors:

$$p(\mathbf{y}_i, \mathbf{X}_i | \mathbf{M}_i \boldsymbol{\lambda}) = p(\mathbf{X}_i | \mathbf{y}_i, \mathbf{M}_i, \boldsymbol{\lambda}) p(\mathbf{y}_i | \mathbf{M}_i \boldsymbol{\lambda}), \tag{16}$$

which we find below:

1.5.1 Outer posterior

The conditional posterior for $\mathbf{X}_i$ is:

$$\begin{aligned} p(\mathbf{X}_i | \mathbf{M}_i, \mathbf{y}_i, \boldsymbol{\lambda}) &\propto p(\mathbf{X}_i, \mathbf{M}_i, \mathbf{y}_i | \boldsymbol{\lambda}) && (17) \\ &\propto \exp\left(\sum_{j=1}^{n_i} \mathbf{x}'_{ij} (\mathbf{U}'\mathbf{D}\mathbf{m}_{ij} - \mathbf{J}\mathbf{y}_i) - \frac{1}{2} \mathbf{x}'_{ij} \mathbf{K} \mathbf{x}_{ij} \right) && (18) \\ &\propto \exp\left(\sum_{j=1}^{n_i} \mathbf{x}'_{ij} (\mathbf{K}\tilde{\mathbf{x}}_{ij} - \mathbf{J}\mathbf{y}_i) - \frac{1}{2} \mathbf{x}'_{ij} \mathbf{K} \mathbf{x}_{ij} \right) && (19) \\ &\propto \exp\left(\sum_{j=1}^{n_i} \mathbf{x}'_{ij} \mathbf{K}\hat{\mathbf{x}}_{ij} - \frac{1}{2} \mathbf{x}'_{ij} \mathbf{K} \mathbf{x}_{ij} \right) && (20) \\ &\propto \prod_j \mathcal{N}(\mathbf{x}_{ij} | \hat{\mathbf{x}}_{ij}, \mathbf{K}^{-1}), && (21) \end{aligned}$$

where

$$\begin{aligned} \mathbf{K}\tilde{\mathbf{x}}_{ij} &= \mathbf{U}'\mathbf{D}\mathbf{m}_{ij}, & \tilde{\mathbf{x}}_{ij} &= \mathbf{K}^{-1}\mathbf{U}'\mathbf{D}\mathbf{m}_{ij}, && (22) \\ \mathbf{K}\hat{\mathbf{x}}_{ij} &= \mathbf{K}\tilde{\mathbf{x}}_{ij} - \mathbf{J}\mathbf{y}_i, & \hat{\mathbf{x}}_{ij} &= \tilde{\mathbf{x}}_{ij} - \mathbf{K}^{-1}\mathbf{J}\mathbf{y}_i. && (23) \end{aligned}$$

1.5.2 Inner posterior

$$\begin{aligned} p(\mathbf{y}_i | \mathbf{M}_i, \boldsymbol{\lambda}) \propto p(\mathbf{y}_i, \mathbf{M}_i | \boldsymbol{\lambda}) &= \left. \frac{p(\mathbf{y}_i, \mathbf{X}_i, \mathbf{M}_i | \boldsymbol{\lambda})}{p(\mathbf{X}_i | \mathbf{y}_i, \mathbf{M}_i, \boldsymbol{\lambda})} \right|_{\mathbf{X}_i = \mathbf{0}} && (24) \\ &\propto \frac{\exp\left(-\frac{1}{2} \mathbf{y}'_i \mathbf{L}_i \mathbf{y}_i + \sum_j \mathbf{m}'_{ij} \mathbf{D}\mathbf{V}\mathbf{y}_i \right)}{\exp\left(-\frac{1}{2} \sum_j \hat{\mathbf{x}}'_{ij} \mathbf{K} \hat{\mathbf{x}}_{ij} \right)} && (25) \end{aligned}$$

Now expand:

$$\begin{aligned} \frac{1}{2} \sum_j \hat{\mathbf{x}}'_{ij} \mathbf{K} \hat{\mathbf{x}}_{ij} &= \frac{1}{2} \sum_j (\tilde{\mathbf{x}}'_{ij} - \mathbf{y}'_i \mathbf{J}' \mathbf{K}^{-1})(\mathbf{K}\tilde{\mathbf{x}}_{ij} - \mathbf{J}\mathbf{y}_i) && (26) \\ &= +\frac{n_i}{2} \mathbf{y}'_i \mathbf{J}' \mathbf{K}^{-1} \mathbf{J} \mathbf{y}_i - \mathbf{y}'_i \mathbf{J}' \tilde{\mathbf{x}}_i + \text{const.} && (27) \end{aligned}$$

where

$$\tilde{\mathbf{x}}_i = \sum_j \tilde{\mathbf{x}}_{ij}. \tag{28}$$

Now use this in (25):

$$p(\mathbf{y}_i|\mathbf{M}_i, \boldsymbol{\lambda}) \propto \exp\Big(\mathbf{y}_i'\mathbf{P}_i\hat{\mathbf{y}}_i - \frac{1}{2}\mathbf{y}_i'\mathbf{P}_i\mathbf{y}_i\Big) \tag{29}$$

$$\propto \mathcal{N}(\mathbf{y}_i|\hat{\mathbf{y}}_i, \mathbf{P}_i^{-1}), \tag{30}$$

where

$$\mathbf{P}_i = n_i(\mathbf{V}'\mathbf{D}\mathbf{V} - \mathbf{J}'\mathbf{K}^{-1}\mathbf{J}) + \mathbf{I}, \tag{31}$$

$$\mathbf{P}_i\hat{\mathbf{y}}_i = \mathbf{V}'\mathbf{D}\mathbf{f}_i - \mathbf{J}'\tilde{\mathbf{x}}_i \tag{32}$$

1.6 Marginal (EM Objective)

$$p(\mathbf{M}_i|\boldsymbol{\lambda}) = \frac{p(\mathbf{M}_i|\mathbf{X}_i, \mathbf{y}_i, \boldsymbol{\lambda})p(\mathbf{X}_i)p(\mathbf{y}_i)}{p(\mathbf{X}_i|\mathbf{y}_i, \mathbf{M}_i, \boldsymbol{\lambda})p(\mathbf{y}_i|\mathbf{M}_i, \boldsymbol{\lambda})}\Bigg|_{\mathbf{y}_i=\mathbf{0},\mathbf{X}_i=\mathbf{0}} \tag{33}$$

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 $\boldsymbol{\lambda}$ of the model. We devote subsections to the E-step, the M-step and the (minim-divergence) MD-step.

2.1 EM auxiliary

$$\tilde{\mathcal{Q}} = \Big\langle \sum_i \log p(\mathbf{M}_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\lambda}) + \text{const} \Big\rangle \tag{34}$$

$$= \Big\langle \sum_{ij} \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}(\mathbf{m}_{ij} - \mathbf{W}\mathbf{z}_{ij})'\mathbf{D}(\mathbf{m}_{ij} - \mathbf{W}\mathbf{z}_{ij}) \Big\rangle \tag{35}$$

$$= \Big\langle \sum_{ij} \frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathbf{m}_{ij}'\mathbf{D}\mathbf{m}_{ij} - \frac{1}{2}\mathbf{z}_{ij}'\mathbf{W}'\mathbf{D}\mathbf{W}\mathbf{z}_{ij} + \mathbf{m}_{ij}'\mathbf{D}\mathbf{W}\mathbf{z}_{ij} \Big\rangle \tag{36}$$

$$= \frac{N}{2}\log|\mathbf{D}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}\mathbf{D}) - \frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{W}'\mathbf{D}\mathbf{W}) + \operatorname{tr}(\mathbf{T}\mathbf{D}\mathbf{W}) \tag{37}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix}, \qquad \mathbf{z}_{ij} = \begin{bmatrix} \mathbf{x}_{ij} \\ \mathbf{y}_i \end{bmatrix}, \tag{38}$$

$$\mathbf{S} = \sum_{ij} \mathbf{m}_{ij}\mathbf{m}'_{ij} \qquad \mathbf{R} = \sum_{ij} \left\langle \mathbf{z}_{ij}\mathbf{z}'_{ij} \right\rangle, \tag{39}$$

$$\mathbf{T} = \sum_{ij} \langle \mathbf{z}_{ij} \rangle \, \mathbf{m}'_{ij}, \qquad N = \sum_{i} n_i. \tag{40}$$

2.2 M-step

Differentiating w.r.t $\mathbf{W}$ and setting to zero gives (independently of $\mathbf{D}$):

$$\mathbf{W}' = \mathbf{R}^{-1}\mathbf{T}. \tag{41}$$

Differentiating w.r.t. $\mathbf{D}$, setting to zero and solving gives:

$$\mathbf{D}^{-1} = \frac{1}{N}(\mathbf{S} + \mathbf{WRW}' - 2\mathbf{WT}) \tag{42}$$

$$= \frac{1}{N}(\mathbf{S} - \mathbf{WT}), \tag{43}$$

where we used (41) for simplification. We can zero the off-diagonals, to make $\mathbf{D}$ diagonal[1]. If we want to further constrain $\mathbf{D}$, to be isotropic, so that $\mathbf{D} = d\mathbf{I}$, then we find:

$$\frac{1}{d} = \frac{1}{ND} \operatorname{tr}(\mathbf{S} + \mathbf{WRW}' - 2\mathbf{WT}) \tag{44}$$

$$= \frac{1}{ND} \operatorname{tr}(\mathbf{S} - \mathbf{WT}), \tag{45}$$

where D is the dimensionality.

2.3 Expectations

To complete the M-step, we need to express $\mathbf{T}$ and $\mathbf{R}$ in terms of the posteriors that we found in section 1.5:

$$\mathbf{T} = \sum_{ij} \langle \mathbf{z}_{ij} \rangle \, \mathbf{m}'_{ij} = \sum_{ij} \left\langle \begin{bmatrix} \mathbf{x}_{ij} \\ \mathbf{y}_i \end{bmatrix} \right\rangle \mathbf{m}'_{ij} = \begin{bmatrix} \mathbf{T}_{\mathbf{x}} \\ \mathbf{T}_{\mathbf{y}} \end{bmatrix}, \tag{46}$$

[1]See Tom Minka's Matrix Calculus tutorial.

and where

$$\mathbf{T_y} = \sum_{ij} \hat{\mathbf{y}}_i \mathbf{m}'_{ij} = \sum_i \hat{\mathbf{y}}_i \mathbf{f}'_i, \tag{47}$$

$$\mathbf{T_x} = \sum_{ij} \langle \hat{\mathbf{x}}_{ij}(\mathbf{y}) \rangle \mathbf{m}'_{ij} = \sum_{ij} \mathbf{K}^{-1}(\mathbf{U}'\mathbf{D}\mathbf{m}_{ij} - \mathbf{J}\hat{\mathbf{y}}_i)\mathbf{m}'_{ij} \tag{48}$$

$$= \mathbf{K}^{-1}\mathbf{U}'\mathbf{D} \sum_{ij} \mathbf{m}_{ij}\mathbf{m}'_{ij} - \mathbf{K}^{-1}\mathbf{J} \sum_i \hat{\mathbf{y}}_i \mathbf{f}'_i \tag{49}$$

$$= \mathbf{K}^{-1}(\mathbf{U}'\mathbf{D}\mathbf{S} - \mathbf{J}\mathbf{T_y}). \tag{50}$$

Finally:

$$\mathbf{R} = \sum_{ij} \left\langle \begin{bmatrix} \mathbf{x}_{ij} \\ \mathbf{y}_i \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{ij} & \mathbf{y}'_i \end{bmatrix} \right\rangle = \begin{bmatrix} \mathbf{R_{xx}} & \mathbf{R_{xy}} \\ \mathbf{R}'_{\mathbf{xy}} & \mathbf{R_{yy}} \end{bmatrix}, \tag{51}$$

where

$$\mathbf{R_{yy}} = \sum_{ij} \langle \mathbf{y}_i \mathbf{y}'_i \rangle = \sum_i n_i (\mathbf{P}_i^{-1} + \hat{\mathbf{y}}_i \hat{\mathbf{y}}'_i), \tag{52}$$

$$\mathbf{R_{xy}} = \sum_{ij} \langle \mathbf{x}_{ij} \mathbf{y}'_i \rangle = \sum_{ij} \left\langle \mathbf{K}^{-1}(\mathbf{U}'\mathbf{D}\mathbf{m}_{ij} - \mathbf{J}\mathbf{y}_i)\mathbf{y}'_i \right\rangle \tag{53}$$

$$= \mathbf{K}^{-1} \left(\mathbf{U}'\mathbf{D}\mathbf{T}'_{\mathbf{y}} - \mathbf{J}\mathbf{R_{yy}} \right), \tag{54}$$

$$\mathbf{R_{xx}} = \sum_{ij} \left\langle \mathbf{x}_{ij} \mathbf{x}'_{ij} \right\rangle = N\mathbf{K}^{-1} + \sum_{ij} \langle \hat{\mathbf{x}}_{ij}(\mathbf{y}) \hat{\mathbf{x}}_{ij}(\mathbf{y})' \rangle \tag{55}$$

where

$$\sum_{ij} \langle \hat{\mathbf{x}}_{ij}(\mathbf{y}) \hat{\mathbf{x}}_{ij}(\mathbf{y})' \rangle \tag{56}$$

$$= \sum_{ij} \left\langle \mathbf{K}^{-1}(\mathbf{U}'\mathbf{D}\mathbf{m}_{ij} - \mathbf{J}\mathbf{y}_i)(\mathbf{m}'_{ij}\mathbf{D}\mathbf{U} - \mathbf{y}'_i\mathbf{J}')\mathbf{K}^{-1} \right\rangle \tag{57}$$

$$= \mathbf{K}^{-1} \sum_{ij} \left\langle \mathbf{U}'\mathbf{D}\mathbf{m}_{ij}\mathbf{m}'_{ij}\mathbf{D}\mathbf{U} - \mathbf{U}'\mathbf{D}\mathbf{m}_{ij}\mathbf{y}'_i\mathbf{J}' - \mathbf{J}\mathbf{y}_i\mathbf{m}'_{ij}\mathbf{D}\mathbf{U} + \mathbf{J}\mathbf{y}_i\mathbf{y}'_i\mathbf{J}' \right\rangle \mathbf{K}^{-1} \tag{58}$$

$$= \mathbf{K}^{-1}(\mathbf{U}'\mathbf{D}\mathbf{S}\mathbf{D}\mathbf{U} - \mathbf{U}'\mathbf{D}\mathbf{T}'_{\mathbf{y}}\mathbf{J}' - \mathbf{J}\mathbf{T_y}\mathbf{D}\mathbf{U} + \mathbf{J}\mathbf{R_{yy}}\mathbf{J}')\mathbf{K}^{-1} \tag{59}$$

2.4 MD-step

Here we temporarily allow a more general prior for the hidden variables:

$$p(\mathbf{y}_i) = \mathcal{N}(\mathbf{y}_i | \mathbf{0}, \mathcal{Y}), \tag{60}$$

$$p(\mathbf{x}_{ij} | \mathbf{y}_i) = \mathcal{N}(\mathbf{x} | \mathbf{G}\mathbf{y}_i, \mathcal{X}) \tag{61}$$

and then maximize the following complementary auxiliary w.r.t. to the new prior parameters:

$$\breve{\mathcal{Q}} = \left\langle \sum_i \log \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathcal{Y}) + \sum_j \log \mathcal{N}(\mathbf{x}_{ij}|\mathbf{G}\mathbf{y}, \mathcal{X}) \right\rangle \tag{62}$$

$$= \sum_i \langle \log \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathcal{Y}) \rangle + \sum_i \sum_j \langle \log \mathcal{N}(\mathbf{x}_{ij}|\mathbf{G}\mathbf{y}_i, \mathcal{X}) \rangle \tag{63}$$

This maximization gives:

$$\mathcal{Y} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{P}^{-1} + \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i', \tag{64}$$

$$\mathbf{G}' = \mathbf{R}_{\mathbf{yy}}^{-1} \mathbf{R}_{\mathbf{xy}}', \tag{65}$$

$$\mathcal{X} = \frac{1}{N} (\mathbf{R}_{\mathbf{xx}} - \mathbf{G}\mathbf{R}_{\mathbf{xy}}'). \tag{66}$$

These non-standard priors can now be transformed back to standard form, by absorbing their effects into **U** and **V**:

$$\mathbf{U} \to \mathbf{U}\,\mathrm{chol}(\mathcal{X})', \tag{67}$$

$$\mathbf{V} \to \mathbf{V}\,\mathrm{chol}(\mathcal{Y})' + \mathbf{U}\mathbf{G}, \tag{68}$$

where **U** on the RHS of (68) is the new value and where $\mathrm{chol}(\mathcal{X})\,\mathrm{chol}(\mathcal{X})' = \mathcal{X}$ denotes Cholesky decomposition.