VARIATIONAL LEARNING OF CLUSTERS OF UNDERCOMPLETE NONSYMMETRIC INDEPENDENT COMPONENTS

Kwokleung Chan, Te-Won Lee and Terrence Sejnowski

The Salk Institute, Computational Neurobiology Laboratory,
10010 N. Torrey Pines Road,
La Jolla,, CA 92037, USA
{*kwchan,tewon,terry*}*@salk.edu*

ABSTRACT

We apply a variational method to automatically determine the number of mixtures of independent components in high-dimensional datasets, in which the sources may be non-symmetrically distributed. The data is modeled by clusters where each cluster is described as a linear mixing of independent factors. Because of the variational bayesian treatment, this method can yield an accurate density model for the observed data without overfitting problems. This allows us also to identify the dimensionality of the data for each cluster. The new method is applied to a difficult real-world medical dataset and is successful in diagnosing glaucoma.

1. INTRODUCTION

In pattern classification, the performance of a method is often determined by how well it can model the underlying statistical distribution of the data. Independent component analysis (ICA) is an example for modeling non-Gaussian structure, e.g., platykurtic or leptokurtic probability density functions. In many applications of ICA, the form of the source distribution (or equivalently the "non-linearity") is fixed and usually symmetric. Real data sets often contain both super and sub-gaussian sources. These sources may be skewed and therefore non-symmetric and they may appear in clusters. Furthermore, the dimensionality within each cluster could be different. In unsupervised classification, one is interested in obtaining a close fit to the observed data distribution without running into overfitting problems.

Clusters of data can be described by an ICA mixture model [1]. Instead of assuming fixed source distributions within the cluster we can use mixture of Gaussians [2][3] to model non-symmetric sources. We use ensemble learning [4] (a.k.a. variational method [5]) to tackle the problem of finding number of clusters and number of sources in each cluster, when modeling high dimensional data.

In this paper, we extend the mixture model of [5] and ICA model of [4], and propose a mixture of undercomplete non-symmetric ICA solution to describe the underlying distribution of small but high dimension dataset.

2. THEORY AND METHOD

Observations $\mathbf{X} = \{\mathbf{x}_t \in \mathcal{R}^N\}, t = [1, \cdots, T]$, are assumed to be generated from one of C clusters with diagonal gaussian noise $\mathbf{\Psi}^c$ and cluster mean ν^c.

$$P(\mathbf{x}_t|\mathbf{A}^c, \boldsymbol{\nu}^c, \mathbf{\Psi}^c) = \sum_c^C P(c_t = c|\rho_c) \int \mathcal{N}(\mathbf{x}_t|\mathbf{A}^c\mathbf{s}_t^c + \boldsymbol{\nu}^c, \mathbf{\Psi}^c) P(\mathbf{s}_t^c)\, d\mathbf{s}_t^c \tag{1}$$

Inside each cluster, observation $\mathbf{x}_t$ is a linear combination of M independent sources $\mathbf{s}_t^c$. To allow for non-symmetric sources, the density of each is modeled by a mixture of K gaussians

$$P(s_{mt}^c) = \sum_k^K \pi_{mk}^c \mathcal{N}(s_{mt}^c|\phi_{mk}^c, \beta_{mk}^c) \tag{2}$$

We also assume a zero mean gaussian density for $\mathbf{A}_n^c$,

$$P(A_{nm}^c) \quad = \quad \mathcal{N}(A_{nm}^c|0, \alpha_m^c) \tag{3}$$

Instead of the likelihood of the data $P(\mathbf{X}|\theta)$ (θ denote the collection of the parameters), we aim to maximize the evidence on the data $P(\mathbf{X})$. Introducing posterior probability $Q(\theta)$ and using the Jensen's inequality, log of the evidence is lower bounded by

$$\log P(\mathbf{X}) \quad \geq \quad \int Q(\theta) \sum_t \log P(\mathbf{x}_t|\theta)\, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \tag{4}$$

repeatly introduce $Q(c_t)$, $Q(\mathbf{s}_t^c)$, we arrive at

$$\log P(\mathbf{X}) \geq \int Q(\theta) \sum_t \sum_{c_t} Q(c_t) \log \frac{P(c_t|\rho_{c_t})}{Q(c_t)}\, d\theta + \int Q(\theta) \sum_t \sum_{c_t} Q(c_t) \int Q(\mathbf{s}_t^c) \Big[\log P(\mathbf{x}_t|\mathbf{s}_t^c, \theta) + \log \frac{P(\mathbf{s}_t^c|\theta)}{Q(\mathbf{s}_t^c)} \Big] d\mathbf{s}_t^c d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \qquad (5)$$

finally, $\log P(\mathbf{s}_t^c|\theta)$ is replaced as

$$\log P(\mathbf{s}_t^c|\theta) = \sum_m \log P(s_{mt}^c|\theta) \geq \sum_{m\,k} Q(k_{mt}^c) \Big[\log \mathcal{N}(s_{mt}^c|\theta_{mk}^c) + \log \frac{\pi_{mk}^c}{Q(k_{mt}^c)} \Big] \qquad (6)$$

to complete the expansion. Notice that $Q(k_{mt}^c)$ is a short form for $Q(k_{mt}^c = k)$.

Learning is accomplished by functional maximization of the lower bound of $\log P(\mathbf{X})$ over $Q(\theta)$, $Q(\mathbf{s}_t^c)$, $Q(c_t)$ and $Q(k_{mt}^c)$. We need a separable posterior $Q(\theta)$

$$Q(\theta) \;=\; Q(\boldsymbol{\rho}) \prod_c \Big[\prod_n Q(\nu_n^c) Q(\Psi_n^c) Q(\mathbf{A}_n^c) \prod_m Q(\alpha_m^c) Q(\boldsymbol{\pi}_m^c) \prod_{m\,k} Q(\phi_{mk}^c) Q(\beta_{mk}^c) \Big] \qquad (7)$$

in order to obtain analytical solutions. Learning rule for $Q(\theta)$ is in the Appendix.

Translational and scale degeneracy present in the model as described by equation 1, 2 and 3. After each update of $Q(\pi_{mk}^c)$, $Q(\phi_{mk}^c)$ and $Q(\beta_{mk}^c)$, we rescale $P(\mathbf{s}_{mt}^c)$ to be zero mean and unit variances. Distribution of $Q(\mathbf{A}^c)$, $Q(\alpha_m^c)$ and $Q(\boldsymbol{\nu}^c)$ etc. are adjusted accordingly. This removes the above two degeneracy and speed up convergence.

Local maxima of $\log P(\mathbf{X})$ exist since each cluster is itself a mixture of (correlated) gaussians. For example, in some solutions, two clusters may be regarded as one containing one bimodal sub-gaussian source. This adversely affects the effectiveness of identifying other sources. We employ an index very similar to the Fisher's discriminant

$$J = |\phi_1 - \phi_2| \times \sqrt{\pi_1 \beta_1 + \pi_2 \beta_2} \qquad (8)$$

J is computed for each pair of adjacent gaussians in $P(s_{mt}^c)$ and $J > 5$ seems to be a good criteria to split the cluster.

By virtue of Central Limit Theorem, linear mixing of arbitrary sources of finite variances would result in a near-gaussian density. As a result, at the early stage of learning when $\mathbf{A}^c$ is randomly initialized, $P(s_{mt}^c)$ sometimes would be driven to have only one single gaussian, especially when the sources to be learned contain some near gaussian components. This is discouraged by reinitializing the sources when all but one of the gaussians die, while keeping the $\mathbf{A}^c$ unchanged.

To compare different models $\mathcal{H}$ resulting from different initial conditions, we compute their corresponding upper bounds $\mathcal{E}(\mathbf{X}|\mathcal{H})$ (equations 5, 6) on the evidence $P(\mathbf{X})$.

$$\mathcal{E} = \sum_t \log Z_t + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \qquad (9)$$

Where Z_t is defined in equation 23. From $\mathcal{E}(\mathbf{X}|\mathcal{H})$ we can select the model $\mathcal{H}$ with highest evidence.

3. EXPERIMENTS

3.1. Synthetic Data

In this simulation experiment, we mix sources of various skewness and kurtosis, namely laplacian, uniform, gamma, beta, generalized gaussian ($\propto \exp(-|x|^q)$), and rectified generalized gaussian, to form 5 clusters in a 2 dimensional space. Number of points in each cluster range from 200 to 400. 0.1% noise is added to the data. The model is initialized with 8 to 10 clusters. Most of the time a 5 clusters solution is obtained. Subplots a) to e) of figure 1 show the densities of the 10 sources recovered. We can see that 3 gaussians are adequate for most of the sources densities. The mixture of gaussians (MOG) fit the source histograms well. Discrepancy from the true distribution arises from randomness in samples generation. Bottom row of the figure draws the initial and final configuration. In the middle right of figure 1, we plot the evolution of the evidence over iterations. Dips correspond to splitting of clusters, and large jumps correspond to vanishing of some clusters. The average signal to noise ratio (SNR) for the mixed sources was 9 dB and the SNR for the recovered sources was on average 38 dB.

3.2. Dimensions Reduction

In this experiment, we embedded 3 clusters containing 2,3 and 4 sources respectively in a four dimensional space. Each cluster has 250 data points and 1% of noise. Correct number of clusters and intrinsic dimension of each are obtained in all trials of runs. The original and learned mixing matrix $\mathbf{A}$'s from one run are shown in Table 1. Besides columns of $\mathbf{A}$, the corresponding rows of s_{mt}^c display negligible values for those 'killed' components. Signal to noise ratios (SNR) for the mixed and recovered sources of each cluster are listed in Table 2. The average SNR for mixed and recovered sources are 5 dB and 22 dB respectively.

3.3. Medical Data Set Analysis: Glaucoma

To evaluate the unsupervised classification ability of the derived learning algorithm on high dimensional data, we apply

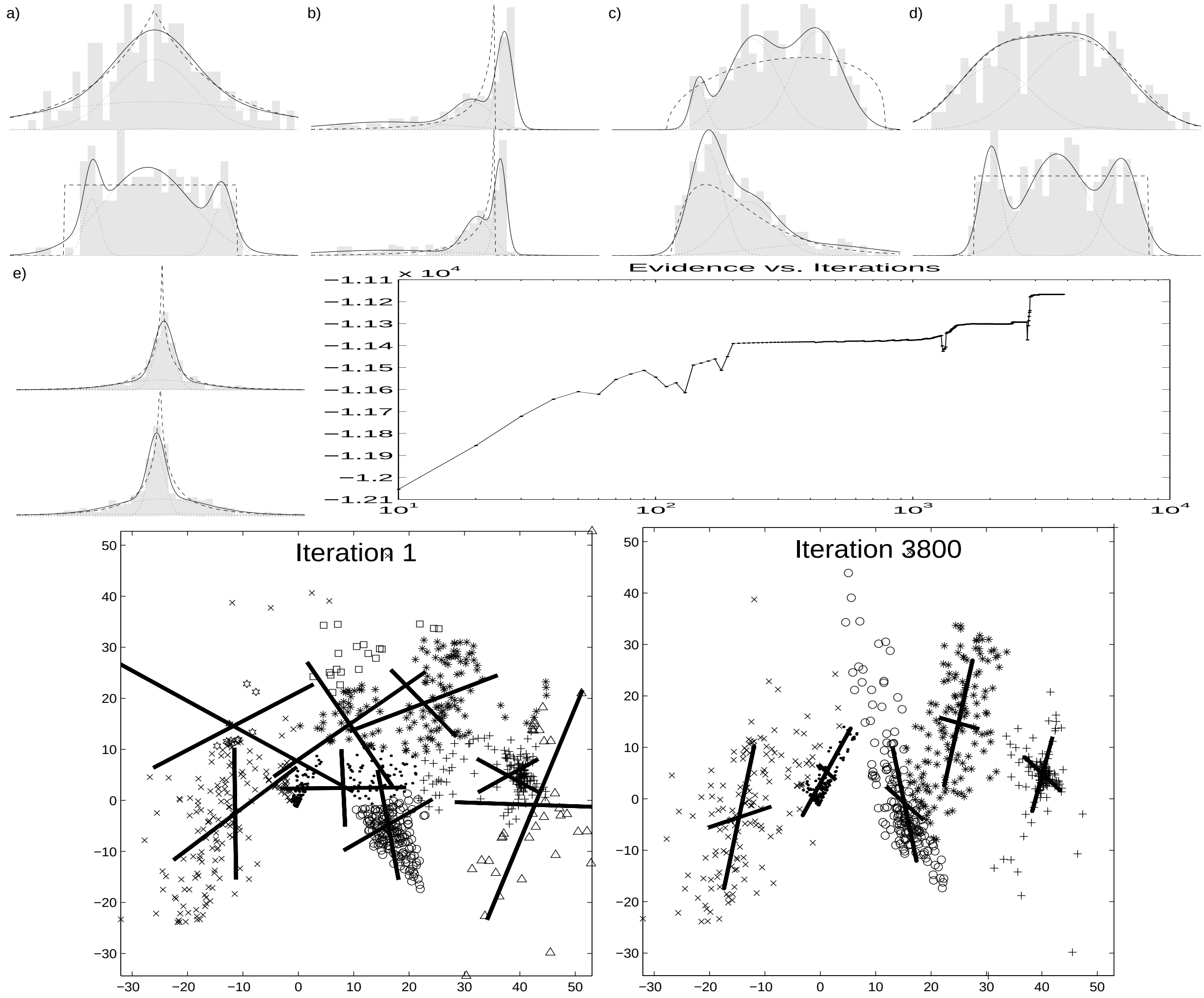

Fig. 1. Application of nonsymmetric undercomplete ICA to synthetic data. a) to e), histograms: recovered sources distribution; dashed lines: original probability densities; solid line: mixture of gaussians modeled probability densities; dotted lines: individual gaussian contribution. Middle row right: Evolution of evidence as function of number of iterations. Bottom row left: after 1 iteration; right: final solution.

it to a glaucoma data set. Glaucoma is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field [6]. Visual field sensitivity test is hence commonly used in clinical setting to evaluate glaucoma. The data vector is composed of the 52 visual sensitivities (measured in dB) over the visual field and the patient's age. Our dataset consists of 189 normal fields and 156 glaucomatous fields, as defined by the presence of glaucomatous optic neuropathy (GON). We started with 1 cluster and look for 20 or less sources. The most stable solution consists of 2 clusters after some split and deletion. Figure 2 shows the strength of the sources in the 2 clusters, and the density distribution of the leading sources. It suggests 12 dimensions in cluster 1 and 6 dimensions in cluster 2. When matching the 2 clusters to the unseen label GON (cluster 1=glaucoma, 2=normal), we get a true positive rate (sensitivity) of 106/156=68% and a true negative rate (specificity) of 187/189=99%. The traditionally used index GHT (glaucoma hemifield test) on the same data yields a sensitivity of 67% and a specificity of 100%. Specificity of $> 95\%$ is desired in the glaucoma community. The large difference between sensitivity and specificity occurs because the glaucoma class contains large number of 'normal looking' examples, while the normal class data is relatively pure. On the right of figure 2 are the grey scale plots of values of column $\mathbf{A}_1$ for the 2 clusters, mapped onto the retina. It is interesting to see that the first principal source for the glaucoma cluster indicates a contrast in visual sensi-

Table 1. Original and learned mixing matrix $\mathbf{A}$'s for the 3 clusters in experiment 2

CLUSTER	ORIGINAL A	LEARNED A
1	$\begin{pmatrix} -3.0 & 2.0 & 0.1 & 0.0 \\ 2.0 & 2.0 & -3.0 & 3.0 \\ 0.0 & 3.0 & 1.0 & 2.0 \\ 1.0 & 1.0 & 0.5 & 0.0 \end{pmatrix}$	$\begin{pmatrix} 3.04 & 0.11 & 0.23 & -1.99 \\ -1.95 & 2.56 & 3.28 & -1.80 \\ -0.15 & -1.26 & 2.36 & -2.63 \\ -0.96 & -0.72 & 0.14 & -0.91 \end{pmatrix}$
2	$\begin{pmatrix} 2.0 & 2.0 & 3.0 \\ 1.0 & 3.0 & -1.0 \\ -3.0 & 0.0 & 2.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 2.51 & -0.00 & 2.72 & -1.93 \\ 2.88 & 0.00 & -1.30 & -0.76 \\ 0.20 & 0.00 & 1.96 & 2.85 \\ 1.18 & -0.00 & 0.88 & -0.95 \end{pmatrix}$
3	$\begin{pmatrix} -3.0 & 2.0 \\ 2.0 & -3.0 \\ 1.0 & 3.0 \\ 2.0 & -4.0 \end{pmatrix}$	$\begin{pmatrix} 3.15 & 1.80 & 0.00 & 0.00 \\ -2.25 & -2.85 & -0.00 & -0.00 \\ -0.71 & 3.04 & 0.00 & 0.00 \\ 1.62 & -4.13 & -0.00 & -0.00 \end{pmatrix}$

tivity between the upper and lower fields, while the normal group shows a relatively uniform visual sensitivity.

4. DISCUSSION

In this paper, we have derived the learning rules for variational learning of mixture of undercomplete non-symmetric ICA solution. Modeling independent source densities by mixture of gaussians is not new. Here we extend the algorithm to the multi-clusters case and study its use as unsupervised classification. This is a combination of Ghahramani's variational learning of mixture of factor analysers [5], Miskin and Lapalainen's ensemble learning of ICA [4][3] and Attais' Independent Factor Analysis [2]. The proposed model has been successfully applied on the glaucoma data set to identify hidden sources and perform unsupervised classification. The discovered fields of regions for glaucoma and non-glaucoma are supported by physiological evidence since they are most commonly used by physicians to determine the disease.

Correctly identifying the number of sources in signal mixtures have always been an important issue. In particular, different number of components may be identified for each cluster. A common conventional way to obtain undercomplete ICA solution is to perform complete ICA on PCA reduced data. Although some efficient methods (e.g. [7]) have been proposed for performing undercomplete ICA skipping PCA, there were no general guidelines on how many sources to look for. This paper employs the automatic dimensions reduction property of bayesian method to identify the number of sources in undercomplete noisy ICA. The use of arbitrary source densities allows us a flexible linear model for data densities fitting. And the variational bayesian treatment prevents us from over-learning.

Table 2. Signal to noise ratio (SNR) of mixed and recovered sources in experiment 2

CLUSTER	MIXTURE (dB)				RECOVERED (dB)			
1	6	5	1	2	21	18	24	19
2	5	8	3		31	21	19	
3	5	11			25	24		

5. REFERENCES

[1] T-W. Lee, M. S. Lewicki, and T. J. Sejnowski, "ICA mixture models for unsupervised classification with non-Gaussian sources and automatic context switching in blind signal separation," *IEEE Transactions on Pattern Recognition and Machine Learning*, vol. 22, no. 10, pp. 1–12, Oct 2000.

[2] Hagai Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[3] H. Lappalainen, "Ensemble learning for independent component analysis," in *International Workshop on ICA and Blind Signal Separation*, Aussois, Jan. 11-15 1999, pp. 7–12.

[4] James Miskin, "Ensemble learning for independent component analysis," Ph.D. thesis, Department of Physics, University of Cambridge, UK, Jun. 2000.

[5] Zoubin Ghahramani and Matthew J. Beal, "Variational inference for bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems 12*, S. Solla, Todd K. Leen, and K.-R. Muller, Eds. 2000, pp. 449–455, MIT Press.

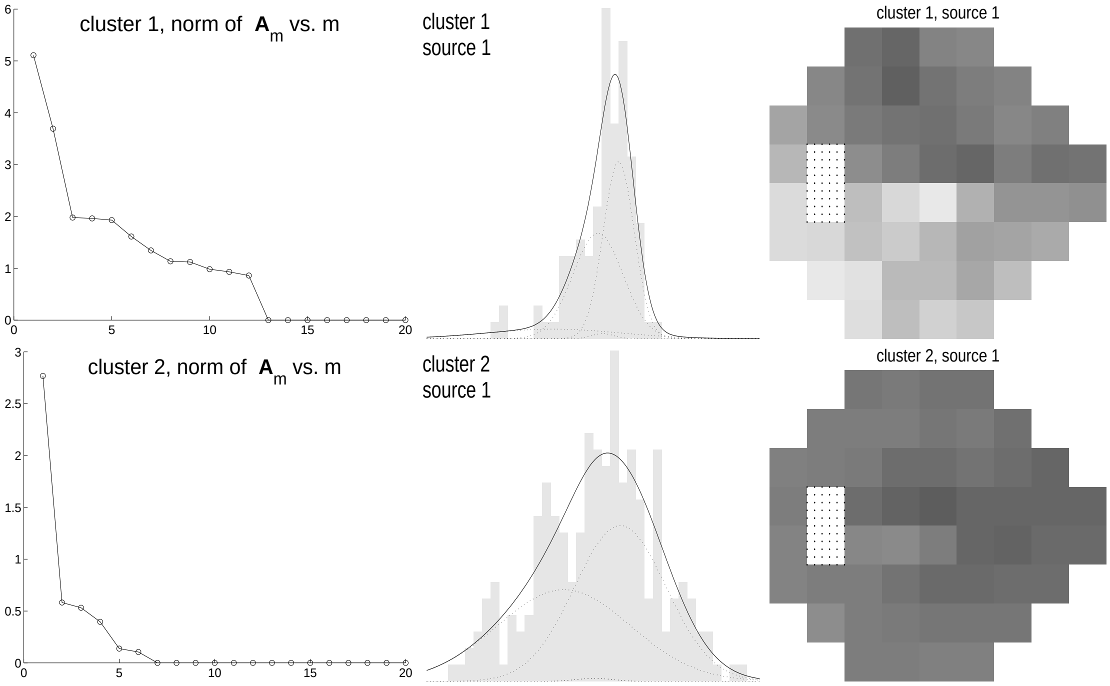

Fig. 2. Illustration for the 2 clusters solution on the glaucoma dataset. Left: standard deviation ($\propto |\mathbf{A}_m|$) of the sources. It shows an intrinsic dimensions of 12 for cluster 1 and 6 for cluster 2. Middle: Density distributions for the first source of each cluster. Right: grey scale visual fields map of $\mathbf{A}_1$.

[6] Kwokleung Chan, Michael Goldbaum, Pamela A. Sample, Te-Won Lee, and Terrence J. Sejnowski, "Comparison of machine learning and traditional classifier in glaucoma diagnosis," in *8th Joint Symposium on Neural Computation*. Institute for Neural Computation, UCSD, May 19, 2001, vol. 11, to appear.

[7] S.-I. Amari, "Natural gradient learning for over- and undercomplete bases in ICA," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, Nov 1999.

A. APPENDIX

Besides the mixture density (equation 1), sources $\mathbf{s}_t^c$ (equation 2) and the mixing matrix $\mathbf{A}^c$ (equation 3) , we employ the following priors on the parameters and hyper-parameters.

$$\begin{aligned}
P(\boldsymbol{\pi}_m^c) &= \mathcal{D}\left(\pi_{m1}^c, \cdots, \pi_{mK}^c | d_o(\pi_{m1}^c), \cdots, d_o(\pi_{mK}^c)\right) \\
P(\phi_{mk}^c) &= \mathcal{N}(\phi_{mk}^c | \mu_o(\phi_{mk}^c), \Lambda_o(\phi_{mk}^c)) \\
P(\beta_{mk}^c) &= \mathcal{G}(\beta_{mk}^c | a_o(\beta_{mk}^c), b_o(\beta_{mk}^c)) \qquad (10)
\end{aligned}$$

$$P(\alpha_m^c) = \mathcal{G}(\alpha_m^c | a_o(\alpha_m^c), b_o(\alpha_m^c)) \qquad (11)$$

$$\begin{aligned}
P(\nu_n^c) &= \mathcal{N}(\nu_n^c | \mu_o(\nu_n^c), \Lambda_o(\nu_n^c)) \\
P(\Psi_n^c) &= \mathcal{G}(\Psi_n^c | a_o(\Psi_n^c), b_o(\Psi_n^c)) \\
p(c|\rho_c) &= \rho_c \\
P(\boldsymbol{\rho}) &= \mathcal{D}(\rho_1, \cdots, \rho_C | d_o(\rho_1), \cdots, d_o(\rho_C)) \,(12)
\end{aligned}$$

Where $\mathcal{N}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{D}(\cdot)$ are the Normal, Gamma and Dirichlet distribution respectively. And we use the following values for the hyper-parameter in the priors. $\mu_o(\nu_n^c) = 0, \Lambda_o(\nu_n^c) = 0.001, d_o(\rho_c) = d_o(\pi_{mk}^c) = 0.001$. $\mu_o(\phi_{mk}^c) = 0, \Lambda_o(\phi_{mk}^c) = 1, a_o(\beta_{mk}^c) = 1.2, b_o(\beta_{mk}^c) = 0.1, a_o(\alpha_m^c) = b_o(\alpha_m^c) = 0.001, a_o(\Psi_n^c) = b_o(\Psi_n^c) = 0.001$.

Using the separable posterior $Q(\theta)$ (equation 7) together with the posterior on the hidden variables $Q(c_t)$, $Q(\mathbf{s}_t^c)$ and $Q(k_{mt}^c)$, we perform functional maximization on the evidence (equation 5 & 6) to obtain the following recursive learning rules. Because of the choice of conjugate prior, free-form optimization results in the same form of $Q(\cdot)$ as $P(\cdot)$, but of different hyper-parameters. The only exception is $Q(\mathbf{s}_t^c)$.

$$\begin{aligned}
Q(\mathbf{s}_t^c) &= \mathcal{N}(\mathbf{s}_t^c | \boldsymbol{\mu}(\mathbf{s}_t^c), \boldsymbol{\Lambda}(\mathbf{s}_t^c)) \\
\boldsymbol{\Lambda}(\mathbf{s}_t^c) &= \langle \mathbf{A}^{c\top} \boldsymbol{\Psi}^c \mathbf{A}^c \rangle \\
&\quad + \mathrm{diag}(\textstyle\sum_k Q(k_{mt}^c) \langle \beta_{mk}^c \rangle) \\
[(\boldsymbol{\Lambda}(\mathbf{s}_t^c)) \boldsymbol{\mu}(\mathbf{s}_t^c)]_m &= \left[\langle \mathbf{A}^{c\top} \boldsymbol{\Psi}^c (\mathbf{x}_t - \boldsymbol{\nu}^c) \rangle\right]_m \\
&\quad + \textstyle\sum_k Q(k_{mt}^c) \langle \beta_{mk}^c \phi_{mk}^c \rangle \,(13)
\end{aligned}$$

$$\begin{aligned}\Lambda(\phi^c_{mk}) &= \Lambda_o(\phi^c_{mk}) + \textstyle\sum_t Q(c_t)Q(k^c_{mt})\langle\beta^c_{mk}\rangle \\ \mu(\phi^c_{mk}) &= \frac{\sum_t Q(c_t)Q(k^c_{mt})\langle\beta^c_{mk}s^c_{mt}\rangle}{\Lambda(\phi^c_{mk})}\end{aligned} \tag{14}$$

$$\begin{aligned}a(\beta^c_{mk}) &= a_o(\beta^c_{mk}) + \frac{1}{2}\sum_t Q(c_t)Q(k^c_{mt}) \\ b(\beta^c_{mk}) &= b_o(\beta^c_{mk}) + \\ &\quad \frac{1}{2}\sum_t Q(c_t)Q(k^c_{mt})\langle(s_{mt} - \phi_{mk})^2\rangle\end{aligned} \tag{15}$$

$$d(\pi^c_{mk}) = d_o(\pi^c_{mk}) + \textstyle\sum_t Q(c_t)Q(k^c_{mt}) \tag{16}$$

$$\begin{aligned}\boldsymbol{\Lambda}(\mathbf{A}^c_n) &= \mathrm{diag}(\langle\alpha^c_1\rangle, \cdots, \langle\alpha^c_m\rangle) + \textstyle\sum_t Q(c_t)\langle\Psi^c_n\rangle\langle\mathbf{s}^c_t\mathbf{s}^{c\top}_t\rangle \\ \boldsymbol{\mu}(\mathbf{A}^c_n) &= [\langle\Psi^c_n\rangle\textstyle\sum_t Q(c_t)\langle(x_{nt} - \nu_n)\mathbf{s}^{c\top}_t\rangle](\boldsymbol{\Lambda}(\mathbf{A}^c_n))^{-1}\end{aligned} \tag{17}$$

$$\begin{aligned}a(\alpha^c_m) &= a_o(\alpha^c_m) + \frac{N}{2} \\ b(\alpha^c_m) &= b_o(\alpha^c_m) + \frac{1}{2}\sum_n\langle A^2_{nm}\rangle\end{aligned} \tag{18}$$

$$\begin{aligned}\Lambda(\nu^c_n) &= \Lambda_o(\nu^c_n) + \textstyle\sum_t Q(c_t)\langle\Psi^c_n\rangle \\ \mu(\nu^c_n) &= \left[\textstyle\sum_t Q(c_t)\langle(x_{nt} - \mathbf{A}^c_n\mathbf{s}^c_t)\Psi_n\rangle\right]/\Lambda(\nu^c_n)\end{aligned} \tag{19}$$

$$\begin{aligned}a(\Psi^c_n) &= a_o(\Psi^c_n) + \frac{1}{2}\sum_t Q(c_t) \\ b(\Psi^c_n) &= b_o(\Psi^c_n) + \frac{1}{2}\sum_n Q(c_t)\langle(x_{nt} - \mathbf{A}^c_n\mathbf{s}^c_t - \nu^c_n)^2\rangle\end{aligned} \tag{20}$$

$$d(\rho_c) = d_o(\rho_c) + \textstyle\sum_t Q(c_t) \tag{21}$$

$\langle\cdot\rangle$ denote the expectation of over the posterior distributions $Q(\cdot)$. Hidden variables distributions $Q(c_t)$ and $Q(k^c_{mt})$ are given by

$$\begin{aligned}\log Q(k^c_{mt}) = \langle\log\pi^c_{mk}\rangle + \langle\log\sqrt{\frac{\beta^c_{mk}}{2\pi}}\rangle \\ -\frac{1}{2}\langle\beta^c_{mk}(s^c_{mt} - \mu^c_{mk})^2\rangle - \log z^c_{mt}\end{aligned} \tag{22}$$

$$\begin{aligned}\log Q(c_t) = \langle\log\rho_c\rangle + \langle\log P(\mathbf{x}_t|\mathbf{s}^c_t, \mathbf{A}^c, \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c)\rangle \\ -\langle\log Q(\mathbf{s}^c_t)\rangle + \textstyle\sum_m\log z^c_{mt} - \log Z_t\end{aligned} \tag{23}$$

where z^c_{mt} and Z_t are the normalization constants.