CHAPTER

BAYESIAN MIXTURE MODELS 20

CHAPTER CONTENTS

In this chapter, within the Bayesian inference framework explained in Chapter 17, practical Bayesian inference algorithms for mixture models are presented. First, algorithms based on variational approximation and Gibbs sampling for Gaussian mixture models are introduced. Then a variational Bayesian algorithm for topic models is explained.

20.1 GAUSSIAN MIXTURE MODELS

In this section, Bayesian inference algorithms for *Gaussian mixture models* (see Section 15.1) are introduced.

20.1.1 BAYESIAN FORMULATION

Let us consider the mixture of m Gaussian models:

$$q(\boldsymbol{x}|\mathcal{W},\mathcal{M},\mathcal{S}) = \sum_{\ell=1}^{m} w_\ell N(\boldsymbol{x}|\boldsymbol{\mu}_\ell, \boldsymbol{S}_\ell^{-1}),$$

where $\mathcal{W} = (w_1,\ldots,w_m)$, $\mathcal{M} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m)$, $\mathcal{S} = (\boldsymbol{S}_1,\ldots,\boldsymbol{S}_m)$, and $N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{S}^{-1})$ denotes the Gaussian density with expectation $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{S}^{-1}$ (or precision matrix $\boldsymbol{S}$):

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{S}^{-1}) = \frac{\sqrt{\det(\boldsymbol{S})}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{20.1}$$

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00031-5

For training samples $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ drawn independently from the true density $p(\boldsymbol{x})$, the likelihood $p(\mathcal{D}|\mathcal{W}, \mathcal{M}, \mathcal{S})$ is given by

$$p(\mathcal{D}|\mathcal{W}, \mathcal{M}, \mathcal{S}) = \prod_{i=1}^{n} q(\boldsymbol{x}_i|\mathcal{W}, \mathcal{M}, \mathcal{S}).$$

For mixing weights $\mathcal{W}$, the *symmetric Dirichlet distribution* (see Section 6.3) is considered as the prior probability:

$$p(\mathcal{W}; \alpha_0) = \mathrm{Dir}(\boldsymbol{W}; \alpha_0) \propto \prod_{\ell=1}^{m} w_\ell^{\alpha_0 - 1},$$

where $\mathrm{Dir}(\boldsymbol{W}; \alpha)$ denotes the symmetric Dirichlet density with concentration parameter α. Note that the Dirichlet distribution is *conjugate* (see Section 17.2) for the discrete distribution given by Eq. (20.3).

For Gaussian expectations $\mathcal{M}$ and Gaussian precision matrices $\mathcal{S}$, the product of the normal distribution and the Wishart distribution (see Section 6.4), called the *normal-Wishart distribution*, is considered as the prior probability:

$$\begin{aligned}
&p(\mathcal{M}, \mathcal{S}; \beta_0, \boldsymbol{W}_0, \nu_0) \\
&= \prod_{\ell=1}^{m} N(\boldsymbol{\mu}_\ell|\boldsymbol{0}, (\beta_0 \boldsymbol{S}_\ell)^{-1}) W(\boldsymbol{S}_\ell; \boldsymbol{W}_0, \nu_0) \\
&\propto \prod_{\ell=1}^{m} \det(\boldsymbol{S}_\ell)^{\frac{\nu_0 - d}{2} - 1} \exp\left(-\frac{\beta_0}{2} \boldsymbol{\mu}_\ell^\top \boldsymbol{S}_\ell \boldsymbol{\mu}_\ell - \frac{1}{2} \mathrm{tr}\left(\boldsymbol{W}_0^{-1} \boldsymbol{S}_\ell\right)\right),
\end{aligned}$$

where $W(\boldsymbol{S}; \boldsymbol{W}, \nu)$ denotes the Wishart density with ν degrees of freedom:

$$W(\boldsymbol{S}; \boldsymbol{W}, \nu) = \frac{\det(\boldsymbol{S})^{\frac{\nu - d - 1}{2}} \exp\left(-\frac{1}{2} \mathrm{tr}\left(\boldsymbol{W}^{-1} \boldsymbol{S}\right)\right)}{\det(2\boldsymbol{W})^{\frac{\nu}{2}} \Gamma_d\left(\frac{\nu}{2}\right)}.$$

Here, d denotes the dimensionality of input $\boldsymbol{x}$ and $\Gamma_d(\cdot)$ denotes the d-dimensional *gamma function* defined by Eq. (6.2):

$$\Gamma_d\left(\frac{\nu}{2}\right) = \int_{\mathbb{S}_d^+} \det(\boldsymbol{S})^{\frac{\nu - d - 1}{2}} \exp\left(-\mathrm{tr}(\boldsymbol{S})\right) \mathrm{d}\boldsymbol{S},$$

where $\mathbb{S}_d^+$ denotes the set of all $d \times d$ positive symmetric matrices. Note that the above *normal-Wishart distribution* is conjugate for the multivariate normal distribution with unknown expectation and unknown precision matrix.

The above formulation is summarized in Fig. 20.1. By the Bayes' theorem, the posterior probability $p(\mathcal{W}, \mathcal{M}, \mathcal{S}|\mathcal{D})$ is given as

$$\begin{aligned}
&p(\mathcal{W}, \mathcal{M}, \mathcal{S}|\mathcal{D}; \alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0) \\
&\quad = \frac{p(\mathcal{D}|\mathcal{W}, \mathcal{M}, \mathcal{S}) p(\mathcal{W}; \alpha_0) p(\mathcal{M}, \mathcal{S}; \beta_0, \boldsymbol{W}_0, \nu_0)}{p(\mathcal{D}; \alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0)},
\end{aligned} \tag{20.2}$$

which is not computationally tractable. Below, practical approximate inference methods are introduced.

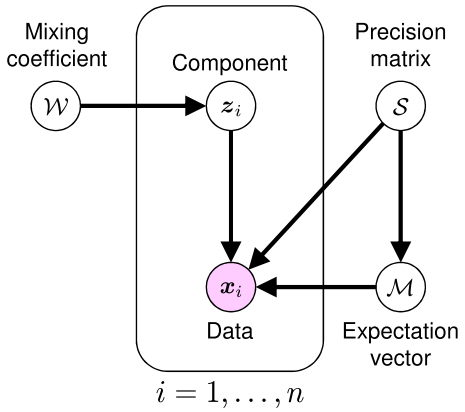


FIGURE 20.1

Variational Bayesian formulation of Gaussian mixture model.

To derive approximate inference methods, let us consider *latent variables*,

$$\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\},$$

for training samples $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where each $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,m})^\top$ is an m-dimensional vector that indicates mixture component selection. Specifically, if the kth component is selected, the kth element of $\boldsymbol{z}_i$ is 1 and all other elements are 0. The probability of observing $\mathcal{Z}$ given mixing weights $\mathcal{W} = (w_1, \ldots, w_m)$ can be expressed as

$$p(\mathcal{Z}|\mathcal{W}) = \prod_{i=1}^{n} \prod_{\ell=1}^{m} w_\ell^{z_{i,\ell}}. \tag{20.3}$$

20.1.2 VARIATIONAL INFERENCE

Here, the *variational technique* introduced in Section 18.2 is employed to approximately compute the posterior probability $p(\mathcal{W}, \mathcal{M}, \mathcal{S}|\mathcal{D})$ given by Eq. (20.2).

The marginal likelihood can be expressed as

$$\begin{aligned}
&\mathrm{ML}(\alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0) \\
&= p(\mathcal{D}; \alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0) \\
&= \iiiint p(\mathcal{D}, \mathcal{Z}, \mathcal{W}, \mathcal{M}, \mathcal{S}; \alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0) \mathrm{d}\mathcal{Z} \mathrm{d}\mathcal{W} \mathrm{d}\mathcal{M} \mathrm{d}\mathcal{S} \\
&= \iiiint p(\mathcal{D}|\mathcal{Z}, \mathcal{M}, \mathcal{S}) p(\mathcal{Z}|\mathcal{W}) p(\mathcal{W}; \alpha_0) \\
&\qquad \times p(\mathcal{M}, \mathcal{S}; \beta_0, \boldsymbol{W}_0, \nu_0) \mathrm{d}\mathcal{Z} \mathrm{d}\mathcal{W} \mathrm{d}\mathcal{M} \mathrm{d}\mathcal{S}.
\end{aligned}$$

For the above marginal likelihood, let us consider the following trial distributions:

$$q(\mathcal{Z})r(\mathcal{W},\mathcal{M},\mathcal{S}).$$

Then, from Eq. (18.4), the VB-E step is given by

$$\begin{aligned} q(\mathcal{Z}) &\propto \exp\left(\iiint r(\mathcal{W},\mathcal{M},\mathcal{S})\log p(\mathcal{D},\mathcal{Z}|\mathcal{W},\mathcal{M},\mathcal{S})\mathrm{d}\mathcal{W}\mathrm{d}\mathcal{M}\mathrm{d}\mathcal{S}\right) \\ &= \exp\left(\iiint r(\mathcal{W},\mathcal{M},\mathcal{S})\log p(\mathcal{D}|\mathcal{Z},\mathcal{M},\mathcal{S})\mathrm{d}\mathcal{W}\mathrm{d}\mathcal{M}\mathrm{d}\mathcal{S}\right. \\ &\quad \left. + \iiint r(\mathcal{W},\mathcal{M},\mathcal{S})\log p(\mathcal{Z}|\mathcal{W})\mathrm{d}\mathcal{W}\mathrm{d}\mathcal{M}\mathrm{d}\mathcal{S}\right). \end{aligned} \tag{20.4}$$

Similarly, from Eq. (18.5), the VB-M step is given by

$$\begin{aligned} r(\mathcal{W},\mathcal{M},\mathcal{S}) &\propto p(\mathcal{W},\mathcal{M},\mathcal{S};\alpha_0,\beta_0,\boldsymbol{W}_0,\nu_0) \\ &\quad \times \exp\left(\int q(\mathcal{Z})\log p(\mathcal{D},\mathcal{Z}|\mathcal{W},\mathcal{M},\mathcal{S})\mathrm{d}\mathcal{Z}\right) \\ &= p(\mathcal{W};\alpha_0)\exp\left(\int q(\mathcal{Z})\log p(\mathcal{Z}|\mathcal{W})\mathrm{d}\mathcal{Z}\right) \\ &\quad \times p(\mathcal{M},\mathcal{S};\beta_0,\boldsymbol{W}_0,\nu_0)\exp\left(\int q(\mathcal{Z})\log p(\mathcal{D}|\mathcal{Z},\mathcal{M},\mathcal{S})\mathrm{d}\mathcal{Z}\right). \end{aligned} \tag{20.5}$$

Combining Eq. (20.4) and Eq. (20.5) yields the VBEM algorithm described in Fig. 20.2 (see Section 10.2.1 of reference [15] for details).

As mentioned in Section 18.2.1, $r(\mathcal{W},\mathcal{M},\mathcal{S})$ obtained by the VBEM algorithm may be a good approximation to the posterior probability $p(\mathcal{W},\mathcal{M},\mathcal{S}|\mathcal{D})$:

$$r(\mathcal{W},\mathcal{M},\mathcal{S}) = \mathrm{Dir}(\mathcal{W}|\widehat{\boldsymbol{\alpha}})\prod_{\ell=1}^{m} N(\boldsymbol{\mu}_\ell|\widehat{\boldsymbol{h}}_\ell,(\widehat{\beta}_\ell\boldsymbol{S}_\ell)^{-1})W(\boldsymbol{S}_\ell|\widehat{\boldsymbol{W}}_\ell,\widehat{\nu}_\ell).$$

The expectations of the marginals of $r(\mathcal{W},\mathcal{M},\mathcal{S})$ can be obtained as

$$\widehat{w}_\ell = \frac{\widehat{\alpha}_\ell}{\sum_{\ell'=1}^{m}\widehat{\alpha}_{\ell'}}, \quad \widehat{\boldsymbol{\mu}}_\ell = \widehat{\boldsymbol{h}}_\ell, \quad \text{and} \quad \widehat{\boldsymbol{S}}_\ell = \widehat{\nu}_\ell\widehat{\boldsymbol{W}}_\ell,$$

which may be used as the most plausible parameter values. Then the following density estimator is obtained:

$$\widehat{p}(\boldsymbol{x}) = \sum_{\ell=1}^{m}\widehat{w}_\ell N(\boldsymbol{x}|\widehat{\boldsymbol{\mu}}_\ell,\widehat{\boldsymbol{S}}_\ell^{-1}).$$

A MATLAB code for computing this VBEM-based density estimator is given in Fig. 20.3, and its behavior is illustrated in Fig. 20.4. Here, the mixture model of five Gaussian components is fitted to the mixture of two Gaussian distributions. As shown in Fig. 20.4, two out of five Gaussian components fit the true two Gaussian distributions well, and the remaining three Gaussian components are almost eliminated—the learned mixing coefficients are given as

1. Initialize parameters $\{\widehat{\alpha}_\ell, \widehat{\boldsymbol{h}}_\ell, \widehat{\beta}_\ell, \widehat{\boldsymbol{W}}_\ell, \widehat{\nu}_\ell\}_{\ell=1}^m$.
2. VB-E step: Compute the distribution of $\mathcal{Z} = \{z_1, \ldots, z_n\}$ from current solution $\{\widehat{\alpha}_\ell, \widehat{\boldsymbol{h}}_\ell, \widehat{\beta}_\ell, \widehat{\boldsymbol{W}}_\ell, \widehat{\nu}_\ell\}_{\ell=1}^m$:

$$q(\mathcal{Z}) = \prod_{i=1}^{n} \prod_{\ell=1}^{m} \widehat{\eta}_{i,\ell}^{z_{i,\ell}}, \quad \text{where} \quad \widehat{\eta}_{i,\ell} = \frac{\widehat{\rho}_{i,\ell}}{\sum_{\ell'=1}^m \widehat{\rho}_{i,\ell'}}.$$

$\{\widehat{\eta}_{i,\ell}\}_{i=1,\,\ell=1}^{n\quad m}$ are the responsibilities computed as

$$\widehat{\rho}_{i,\ell} = \exp\left(\psi(\widehat{\alpha}_\ell) - \psi\left(\sum_{\ell'=1}^{m} \widehat{\alpha}_{\ell'}\right) + \frac{1}{2}\sum_{j=1}^{d} \psi\left(\frac{\widehat{\nu}_\ell + 1 - j}{2}\right) + \frac{1}{2}\log\det(\widehat{\boldsymbol{W}}_\ell) - \frac{d}{2\widehat{\beta}_\ell} - \frac{\widehat{\nu}_\ell}{2}(\boldsymbol{x}_i - \widehat{\boldsymbol{h}}_\ell)^\top \widehat{\boldsymbol{W}}_\ell (\boldsymbol{x}_i - \widehat{\boldsymbol{h}}_\ell)\right),$$

where $\psi(\alpha)$ denotes the *digamma function* defined as the log-derivative of the gamma function $\Gamma(\alpha)$:

$$\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\log\Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}.$$

3. VB-M step: Compute the joint distribution of $\mathcal{W} = (w_1, \ldots, w_m)$, $\mathcal{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m)$, and $\mathcal{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m)$ from current responsibilities $\{\widehat{\eta}_{i,\ell}\}_{i=1,\,\ell=1}^{n\quad m}$:

$$r(\mathcal{W}, \mathcal{M}, \mathcal{S}) = \mathrm{Dir}(\mathcal{W}|\widehat{\boldsymbol{\alpha}}) \prod_{\ell=1}^{m} N(\boldsymbol{\mu}_\ell|\widehat{\boldsymbol{h}}_\ell, (\widehat{\beta}_\ell \boldsymbol{S}_\ell)^{-1}) W(\boldsymbol{S}_\ell|\widehat{\boldsymbol{W}}_\ell, \widehat{\nu}_\ell),$$

where

$$\widehat{\gamma}_\ell = \sum_{i=1}^{n} \widehat{\eta}_{i,\ell}, \qquad \widehat{\boldsymbol{c}}_\ell = \frac{1}{\widehat{\gamma}_\ell}\sum_{i=1}^{n} \widehat{\eta}_{i,\ell}\boldsymbol{x}_i, \qquad \widehat{\boldsymbol{h}}_\ell = \frac{\widehat{\gamma}_\ell}{\widehat{\beta}_\ell}\widehat{\boldsymbol{c}}_\ell,$$

$$\widehat{\alpha}_\ell = \alpha_0 + \widehat{\gamma}_\ell, \qquad \widehat{\beta}_\ell = \beta_0 + \widehat{\gamma}_\ell, \qquad \widehat{\nu}_\ell = \nu_0 + \widehat{\gamma}_\ell, \quad \text{and}$$

$$\widehat{\boldsymbol{W}}_\ell = \left(\boldsymbol{W}_0^{-1} + \sum_{i=1}^{n} \widehat{\eta}_{i,\ell}(\boldsymbol{x}_i - \widehat{\boldsymbol{c}}_\ell)(\boldsymbol{x}_i - \widehat{\boldsymbol{c}}_\ell)^\top + \frac{\beta_0\widehat{\gamma}_\ell}{\beta_0 + \widehat{\gamma}_\ell}\widehat{\boldsymbol{c}}_\ell\widehat{\boldsymbol{c}}_\ell^\top\right)^{-1}.$$

4. Iterate 2–3 until convergence.

FIGURE 20.2

VBEM algorithm for Gaussian mixture model. $(\alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0)$ are hyperparameters.

```
x=[2*randn(1,100)-5 randn(1,50); randn(1,100) randn(1,50)+3];
[d,n]=size(x); m=5; e=rand(n,m); W=zeros(d,d,m); b0=1;
for o=1:10000
  e=e./repmat(sum(e,2),[1 m]);
  g=sum(e); a=1+g; b=b0+g; nu=3+g; w=a/sum(a);
  xe=x*e; c=xe./repmat(g,[d 1]); h=xe./repmat(b,[d 1]);
  for k=1:m
    t1=x-repmat(c(:,k),[1 n]); t2=x-repmat(h(:,k),[1 n]);
    W(:,:,k)=inv(eye(d)+(t1.*repmat(e(:,k)',[d 1]))*t1' ...
                 +c(:,k)*c(:,k)'*b0*g(k)/(b0+g(k)));
    t3=sum(psi((nu(k)+1-[1:d])/2))+log(det(W(:,:,k)));
    e(:,k)=exp(t3/2+psi(a(k))-psi(sum(a))-d/2/b(k) ...
               -sum(t2.*(W(:,:,k)*t2))*nu(k)/2);
  end
  if o>1 && norm(w-w0)+norm(h-h0)+norm(W(:)-W0(:))<0.001
    break
  end
  w0=w; h0=h; W0=W;
end

figure(1); clf; hold on
plot(x(1,:),x(2,:),'ro'); v=linspace(0,2*pi,100);
for k=1:m
  [V,D]=eig(nu(k)*W(:,:,k));
  X=3*w(k)*V'*[cos(v)/D(1,1); sin(v)/D(2,2)];
  plot(h(1,k)+X(1,:),h(2,k)+X(2,:),'b-')
end
```

FIGURE 20.3

MATLAB code of VBEM algorithm for Gaussian mixture model.

$$(\widehat{w}_1, \widehat{w}_2, \widehat{w}_3, \widehat{w}_4, \widehat{w}_5) = (0.01, 0.33, 0.01, 0.01, 0.65).$$

As explained in Section 18.2.1, the negative of the variational free energy,

$$F(q,r) = \iiiint q(\mathcal{Z}) r(\mathcal{W}, \mathcal{M}, \mathcal{S}) \times \log \frac{q(\mathcal{Z}) r(\mathcal{W}, \mathcal{M}, \mathcal{S})}{p(\mathcal{D}, \mathcal{Z}, \mathcal{W}, \mathcal{M}, \mathcal{S}; \alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0)} \mathrm{d}\mathcal{Z}\mathrm{d}\mathcal{W}\mathrm{d}\mathcal{M}\mathrm{d}\mathcal{S},$$

gives a lower bound of the log marginal likelihood, $\mathrm{ML}(\alpha_0, \beta_0, \boldsymbol{W}_0, \nu_0)$. Based on the dependency in Fig. 20.1, the variational free energy can be expressed as

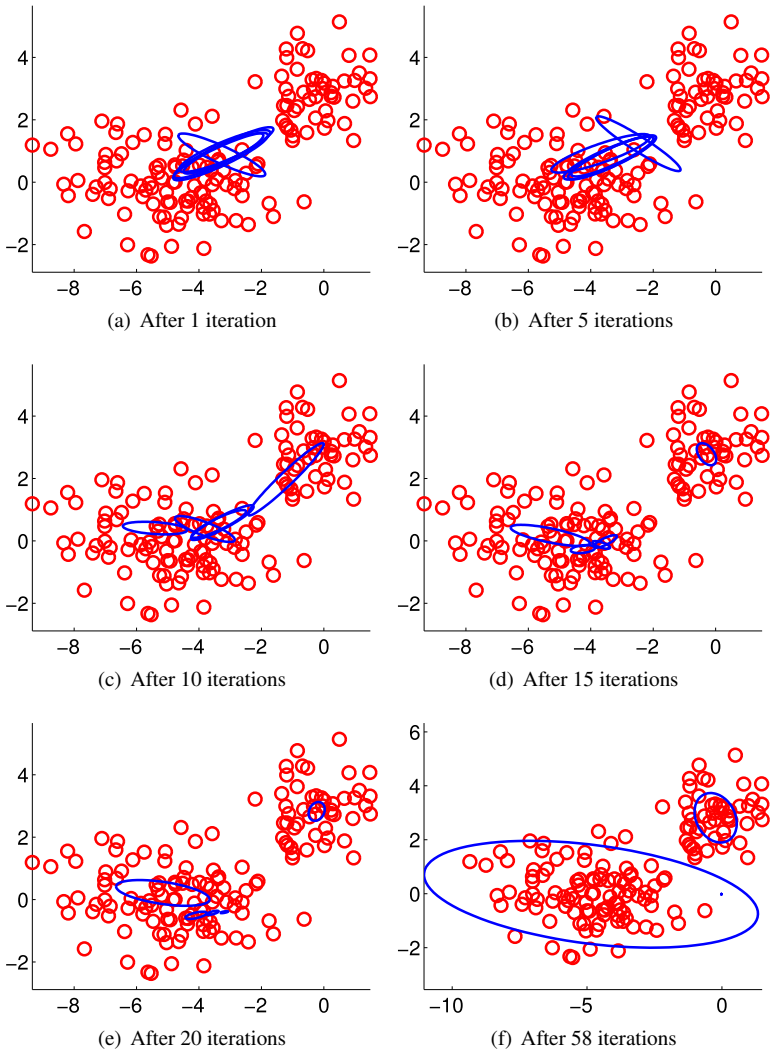


FIGURE 20.4

Example of VBEM algorithm for Gaussian mixture model. The size of ellipses is proportional to the mixing weights $\{w_\ell\}_{\ell=1}^m$. A mixture model of five Gaussian components is used here, but three components have mixing coefficient close to zero and thus they are almost eliminated.

$$F(q,r) = \iiiint q(\mathcal{Z})r(\mathcal{W},\mathcal{M},\mathcal{S})\Big(\log q(\mathcal{Z}) + \log r(\mathcal{W},\mathcal{M},\mathcal{S}) - \log p(\mathcal{D}|\mathcal{Z},\mathcal{M},\mathcal{S}) - \log p(\mathcal{Z}|\mathcal{W}) - \log p(\mathcal{W};\alpha_0) - \log p(\mathcal{M},\mathcal{S};\beta_0,\boldsymbol{W}_0,\nu_0)\Big)\mathrm{d}\mathcal{Z}\mathrm{d}\mathcal{W}\mathrm{d}\mathcal{M}\mathrm{d}\mathcal{S},$$

which allows us to choose the hyperparameters and the number of mixing components by the empirical Bayes method (see Section 17.4). See Section 10.2.2 of reference [15] for more details of computing the variational free energy for Gaussian mixture models.

20.1.3 GIBBS SAMPLING

Next, the *Gibbs sampling technique* introduced in Section 19.3.3 is employed to numerically approximate the posterior probability $p(\mathcal{W},\mathcal{M},\mathcal{S}|\mathcal{D})$ given by Eq. (20.2).

Collapsed Gibbs sampling for assignment $\boldsymbol{z}_i$ can be performed as follows [74]:

$$\begin{aligned} p(z_{i,\ell}=1|\widetilde{\mathcal{Z}},\mathcal{D};\alpha_0,\beta_0,\boldsymbol{W}_0,\nu_0) \\ \propto (\widehat{\alpha}_\ell-1)\times t\left(\boldsymbol{x}_i\middle|\widehat{\boldsymbol{c}}_\ell,\frac{\widehat{\beta}_\ell+1}{\widehat{\beta}_\ell(\widehat{\nu}_\ell-d+1)}\widehat{\boldsymbol{W}}_\ell^{-1},\widehat{\nu}_\ell-d+1\right), \end{aligned} \tag{20.6}$$

where $\widetilde{\mathcal{Z}}=\{\boldsymbol{z}_{i'}\}_{i'\neq i}$, $\widehat{\alpha}_\ell=\alpha_0+n_\ell$, $\widehat{\beta}_\ell=\beta_0+n_\ell$, $\widehat{\nu}_\ell=\nu_0+n_\ell$,

$$\widehat{\boldsymbol{W}}_\ell^{-1}=\left(\boldsymbol{W}_0+\sum_{i:z_{i,\ell}=1}\boldsymbol{x}_i\boldsymbol{x}_i^\top-\frac{n_\ell^2}{\widehat{\beta}_\ell}\widehat{\boldsymbol{c}}_\ell\widehat{\boldsymbol{c}}_\ell^\top\right)^{-1},$$

$n_\ell=\sum_{i=1}^n z_{i,\ell}$, $\widehat{\boldsymbol{c}}_\ell=\frac{1}{n_\ell}\sum_{i:z_{i,\ell}=1}^n\boldsymbol{x}_i$, and $t(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ denotes the probability density function of the *multivariate Student's t-distribution*:

$$t(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma(\frac{\nu}{2})(\nu\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}}\left(1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{\nu+d}{2}}.$$

After estimating assignment $\mathcal{Z}$, parameters $\mathcal{W},\mathcal{M},\mathcal{S}$ for each Gaussian component may be estimated separately as

$$\widehat{w}_\ell=\frac{\widehat{\alpha}_\ell}{\sum_{\ell'=1}^m\widehat{\alpha}_{\ell'}},\quad \widehat{\boldsymbol{\mu}}_\ell=\frac{n_\ell}{\widehat{\beta}_\ell}\widehat{\boldsymbol{c}}_\ell,\quad\text{and}\quad \widehat{\boldsymbol{S}}_\ell=\widehat{\nu}_\ell\widehat{\boldsymbol{W}}_\ell.$$

A MATLAB code of collapsed Gibbs sampling for Gaussian mixture models is given in Fig. 20.5, and its behavior is illustrated in Fig. 20.6. A mixture model of five Gaussian components is used here, but only two components remain and no samples belong to the remaining three components after Gibbs sampling.

The *Dirichlet process*, denoted by $\mathrm{DP}(\alpha_0,p_0)$, is a probability distribution of probability distributions, and a Dirichlet distribution is generated from a Dirichlet process. p_0 is called the *base distribution* which corresponds to the mean of the Dirichlet distribution, while α_0 is called the *concentration parameter* which corresponds to the *precision* (i.e., the inverse variance) of the Dirichlet distribution. The *Dirichlet process mixture* is a mixture model using the Dirichlet process as a prior probability and is equivalent to considering an infinite-dimensional Dirichlet distribution. Indeed, taking the limit $m\to\infty$ for $\alpha_0=\alpha_0'/m$ in Eq. (20.6) gives a *collapsed Gibbs sampling* procedure for the Dirichlet process mixture model.

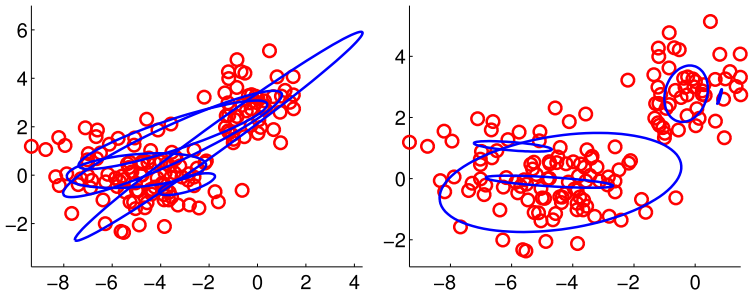
```
x=[2*randn(1,100)-5 randn(1,50); randn(1,100) randn(1,50)+3];
[d,n]=size(x); m=5; z=mod(randperm(n),m)+1;
a0=1; b0=1; n0=1; W0=eye(d);
for o=1:100
  for i=1:n
    g=(1:n~=i); X=x(:,g); Z=z(g);
    for k=1:m
      p(k)=0; e=(Z==k); t=sum(e);
      if t~=0
        u=n0+t-d+1; b=b0+t; c=sum(X(:,e),2); xi=x(:,i)-c/t;
        W=inv((b+1)/b/u*(W0+X(:,e)*X(:,e)'-c*c'/b));
        p(k)=(a0+t-1)*gamma((u+d)/2)/gamma(u/2)*u^(-d/2) ...
             *sqrt(det(W))*(1+xi'*W*xi/u)^(-(u+d)/2);
    end, end
    z(i)=find(cumsum(p/sum(p))>rand,1);
end, end

figure(1); clf; hold on
plot(x(1,:),x(2,:),'ro'); v=linspace(0,2*pi,100);
for k=1:m
  e=(z==k); t=sum(e); nu(k)=n0+t; u=nu(k)-d+1; b=b0+t;
  c=sum(x(:,e),2); w(k)=a0+t; h(:,k)=c/b; W(:,:,k)=zeros(d);
  if t~=0
    W(:,:,k)=inv((W0+x(:,e)*x(:,e)'-c*c'/b));
  end
end
w=w./sum(w);
for k=1:m
  [V,D]=eig(nu(k)*W(:,:,k));
  X=3*w(k)*V'*[cos(v)/D(1,1); sin(v)/D(2,2)];
  plot(h(1,k)+X(1,:),h(2,k)+X(2,:),'b-')
end
```

FIGURE 20.5

MATLAB code of collapsed Gibbs sampling for Gaussian mixture model.

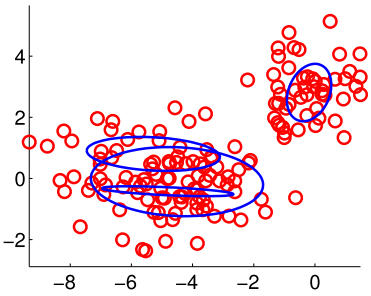
20.2 LATENT DIRICHLET ALLOCATION (LDA)

One of the most successful applications of the Bayesian generative approach is *topic modeling* called LDA [16]. In this section, an implementation of LDA based on *Gibbs sampling* is explained [51].

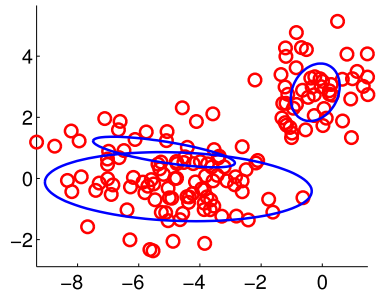


(a) After 1 iteration

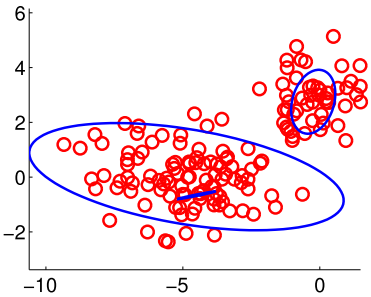
(b) After 5 iterations



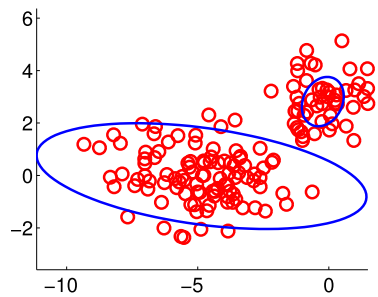
(c) After 10 iterations



(d) After 20 iterations



(e) After 30 iterations



(f) After 100 iterations

FIGURE 20.6

Example of collapsed Gibbs sampling for Gaussian mixture model. A mixture model of five Gaussian components is used here, but only two components remain and no samples belong to the remaining three components.

20.2.1 TOPIC MODELS

A topic model is a generative model for a set of D documents,

$$\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\},$$

where the dth document $\boldsymbol{w}_d = (w_{d,1}, \ldots, w_{d,n_d})$ is a sequence of n_d words in a corpus with W unique words. More specifically, each word $w_{n,d} \in \{1, \ldots, W\}$ is supposed to be given by the mixture of T topics as

$$p(w_{n,d}|\boldsymbol{\theta}^{(d)}, \boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(T)}) = \sum_{t_{n,d}=1}^{T} p(t_{n,d}|\boldsymbol{\theta}^{(d)}) p(w_{n,d}|t_{n,d}, \boldsymbol{\phi}^{(t_{n,d})}),$$

where $t_{n,d} \in \{1, \ldots, T\}$ is a latent variable that indicates the topic of $w_{n,d}$. For topic $t = 1, \ldots, T$, the probability,

$$p(t|\boldsymbol{\theta}^{(d)}) = \theta_t^{(d)} \geq 0,$$

denotes the discrete probability of topic t with parameter $\boldsymbol{\theta}^{(d)} = (\theta_1^{(d)}, \ldots, \theta_T^{(d)})^\top$ such that $\sum_{t=1}^{T} \theta_t^{(d)} = 1$. For $w = 1, \ldots, W$, the probability,

$$p(w|t, \boldsymbol{\phi}^{(t)}) = \phi_w^{(t)} \geq 0,$$

denotes the discrete probability of observing word w under topic t with parameter $\boldsymbol{\phi}^{(t)} = (\phi_1^{(t)}, \ldots, \phi_W^{(t)})^\top$ such that $\sum_{w=1}^{W} \phi_w^{(t)} = 1$. This generative model means that, for each word w in document d, topic t is chosen following the discrete probability with parameter $\boldsymbol{\theta}^{(d)}$ and then word w is chosen following the discrete probability with parameter $\boldsymbol{\phi}^{(t)}$.

The goal of topic modeling is to specify the multinomial parameters $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$ from data $\mathcal{W}$.

20.2.2 BAYESIAN FORMULATION

Here, instead of directly learning the parameters $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$, let us consider the posterior probability of the set $\mathcal{T}$ of all topics for all documents $\mathcal{W}$:

$$\mathcal{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_D\},$$

where $\boldsymbol{t}_d = (t_{d,1}, \ldots, t_{d,n_d})$.

For the multinomial parameters $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$, the *symmetric Dirichlet distributions* (see Section 6.3) are considered as the prior probabilities:

$$p(\boldsymbol{\theta}^{(d)}; \alpha) = \mathrm{Dir}(\boldsymbol{\theta}^{(d)}; \alpha) \quad \text{and} \quad p(\boldsymbol{\phi}^{(t)}; \beta) = \mathrm{Dir}(\boldsymbol{\phi}^{(t)}; \beta),$$

where $\mathrm{Dir}(\cdot\,; \alpha)$ denotes the symmetric Dirichlet density with concentration parameter α. Note that the Dirichlet distributions are *conjugate* (see Section 17.2) for discrete distributions.

The posterior probability of topics $\mathcal{T}$ is given by

$$p(\mathcal{T}|\mathcal{W}; \alpha, \beta) = \frac{p(\mathcal{W}, \mathcal{T}; \alpha, \beta)}{\sum_{\mathcal{T}} p(\mathcal{W}, \mathcal{T}; \alpha, \beta)}, \tag{20.7}$$

where the marginal likelihood $p(\mathcal{W}, \mathcal{T}; \alpha, \beta)$ can be expressed *analytically* as

$$\begin{aligned} & p(\mathcal{W}, \mathcal{T}; \alpha, \beta) \\ &= \prod_{d=1}^{D} \prod_{n=1}^{n_d} \int \theta_{t_{n,d}}^{(d)} \mathrm{Dir}(\boldsymbol{\theta}^{(d)}; \alpha) \mathrm{d}\boldsymbol{\theta}^{(d)} \int \phi_{w_{n,d}}^{(t_{n,d})} \mathrm{Dir}(\boldsymbol{\phi}^{(t_{n,d})}; \beta) \mathrm{d}\boldsymbol{\phi}^{(t_{n,d})} \\ &= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \right)^D \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(n_t^{(d)} + \alpha)}{\Gamma(\sum_{t=1}^{T} n_t^{(d)} + T\alpha)} \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \right)^T \prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(n_t^{(w)} + \beta)}{\Gamma(\sum_{w=1}^{W} n_t^{(w)} + W\beta)}, \end{aligned}$$

where $n_t^{(d)}$ denotes the number of times a word from document d is assigned to topic t, $n_t^{(w)}$ denotes the number of times word w is assigned to topic t, and $\Gamma(\cdot)$ denotes the *gamma function* (see Section 4.3).

20.2.3 GIBBS SAMPLING

Let us use *Gibbs sampling* (see Section 19.3.3) to approximate the posterior probability $p(\mathcal{T}|\mathcal{W}; \alpha, \beta)$.

Let $\widetilde{\mathcal{W}}$ and $\widetilde{\mathcal{T}}$ be defined in the same way as $\mathcal{W}$ and $\mathcal{T}$, but $w_{n,d}$ and $t_{n,d}$ are excluded. Then the conditional probability needed for Gibbs sampling is given by

$$\begin{aligned} p(t_{n,d}|\widetilde{\mathcal{T}}, \mathcal{W}) &= p(t_{n,d}|\widetilde{\mathcal{T}}, w_{n,d}, \widetilde{\mathcal{W}}) \\ &\propto p(t_{n,d}|\widetilde{\mathcal{T}}) p(w_{n,d}|t_{n,d}, \widetilde{\mathcal{T}}, \widetilde{\mathcal{W}}). \end{aligned} \tag{20.8}$$

The first term in Eq. (20.8) can be expressed as

$$p(t_{n,d}|\widetilde{\mathcal{T}}) = \int p(t_{n,d}|\boldsymbol{\theta}^{(d)}) p(\boldsymbol{\theta}^{(d)}|\widetilde{\mathcal{T}}) \mathrm{d}\boldsymbol{\theta}^{(d)}, \tag{20.9}$$

where Bayes' theorem yields

$$p(\boldsymbol{\theta}^{(d)}|\widetilde{\mathcal{T}}) \propto p(\widetilde{\mathcal{T}}|\boldsymbol{\theta}^{(d)}) p(\boldsymbol{\theta}^{(d)}; \alpha).$$

Here, the prior probability $p(\boldsymbol{\theta}^{(d)}; \alpha)$ is chosen to be symmetric Dirichlet distribution with concentration parameter α, which is conjugate for the discrete distribution $p(\widetilde{\mathcal{T}}|\boldsymbol{\theta}^{(d)})$. Then the posterior probability $p(\boldsymbol{\theta}^{(d)}|\widetilde{\mathcal{T}})$ is also the symmetric Dirichlet distribution with concentration parameter $\widetilde{n}_{t_{n,d}}^{(d)} + \alpha$, where $\widetilde{n}_{t_{n,d}}^{(d)}$ is defined in the same way as $n_{t_{n,d}}^{(d)}$, but $w_{n,d}$ and $t_{n,d}$ are excluded. Then Eq. (20.9) can be computed analytically as

$$p(t_{n,d}|\widetilde{\mathcal{T}}) = \int \theta_{t_{n,d}}^{(d)} \mathrm{Dir}(\boldsymbol{\theta}^{(d)}; \widetilde{n}_{t_{n,d}}^{(d)} + \alpha) \mathrm{d}\boldsymbol{\theta}^{(d)} = \frac{\widetilde{n}_{t_{n,d}}^{(d)} + \alpha}{\sum_{t=1}^{T} \widetilde{n}_t^{(d)} + T\alpha}.$$

Similarly, the second term in Eq. (20.8) can be expressed as

$$p(w_{n,d}|t_{n,d}, \widetilde{\mathcal{T}}, \widetilde{\mathcal{W}}) = \int p(w_{n,d}|t_{n,d}, \boldsymbol{\phi}^{(t_{n,d})}) p(\boldsymbol{\phi}^{(t_{n,d})}|\widetilde{\mathcal{T}}, \widetilde{\mathcal{W}}) \mathrm{d}\boldsymbol{\phi}^{(t_{n,d})}, \tag{20.10}$$

where Bayes' theorem yields

$$p(\boldsymbol{\phi}^{(t_{n,d})}|\widetilde{\mathcal{T}},\widetilde{\mathcal{W}}) \propto p(\widetilde{\mathcal{W}}|\boldsymbol{\phi}^{(t_{n,d})},\widetilde{\mathcal{T}})p(\boldsymbol{\phi}^{(t_{n,d})};\beta).$$

Here, the prior probability $p(\boldsymbol{\phi}^{(t_{n,d})};\beta)$ is chosen to be the symmetric Dirichlet distribution with concentration parameter β, which is conjugate for the discrete distribution $p(\widetilde{\mathcal{W}}|\boldsymbol{\phi}^{(t_{n,d})},\widetilde{\mathcal{T}})$. Then the posterior probability $p(\boldsymbol{\phi}^{(t_{n,d})}|\widetilde{\mathcal{T}},\widetilde{\mathcal{W}})$ is also the symmetric Dirichlet distribution with concentration parameter $\widetilde{n}_{t_{n,d}}^{(w_{n,d})} + \beta$, where $\widetilde{n}_{t_{n,d}}^{(w_{n,d})}$ is defined in the same way as $n_{t_{n,d}}^{(w_{n,d})}$, but $w_{n,d}$ and $t_{n,d}$ are excluded. Then Eq. (20.10) can be computed analytically as

$$\begin{aligned} p(w_{n,d}|t_{n,d},\widetilde{\mathcal{T}},\widetilde{\mathcal{W}}) &= \int \phi_{w_{n,d}}^{(t_{n,d})} \mathrm{Dir}(\boldsymbol{\phi}^{(t_{n,d})};\widetilde{n}_{t_{n,d}}^{(w_{n,d})} + \beta)\mathrm{d}\boldsymbol{\phi}^{(t_{n,d})} \\ &= \frac{\widetilde{n}_{t_{n,d}}^{(w_{n,d})} + \beta}{\sum_{w=1}^{W} \widetilde{n}_{t_{n,d}}^{(w)} + W\beta}. \end{aligned}$$

Summarizing the above equations, Eq. (20.8) can be expressed as

$$p(t_{n,d}|\widetilde{\mathcal{T}},\mathcal{W}) \propto \frac{\widetilde{n}_{t_{n,d}}^{(d)} + \alpha}{\sum_{t=1}^{T} \widetilde{n}_{t}^{(d)} + T\alpha} \times \frac{\widetilde{n}_{t_{n,d}}^{(w_{n,d})} + \beta}{\sum_{w=1}^{W} \widetilde{n}_{t_{n,d}}^{(w)} + W\beta},$$

which shows that Gibbs sampling can be efficiently performed just by counting the number of words.

Finally, with estimated topics $\widehat{\mathcal{T}}$, the solutions $\widehat{\theta}_t^{(d)}$ and $\widehat{\phi}_t^{(w)}$ for any $w \in \{1,\dots,W\}$, $t \in \{1,\dots,T\}$, and $d \in \{1,\dots,D\}$ can be computed as

$$\widehat{\theta}_t^{(d)} = \frac{\widehat{n}_{t_{n,d}}^{(d)} + \alpha}{\sum_{t=1}^{T} \widehat{n}_{t}^{(d)} + T\alpha} \quad \text{and} \quad \widehat{\phi}_t^{(w)} = \frac{\widehat{n}_{t_{n,d}}^{(w_{n,d})} + \beta}{\sum_{w=1}^{W} \widehat{n}_{t_{n,d}}^{(w)} + W\beta}.$$