4 Maximum a-posteriori approximation

Maximum a-posteriori (MAP) approximation is a well-known and widely used approximation for Bayesian inference. The approximation covers all variables including model parameters Θ, latent variables Z, and classification categories C (word sequence W in the automatic speech recognition case). For example, the Viterbi algorithm ($\arg\max_Z p(Z|\mathbf{O})$) in the continuous density hidden Markov model (CDHMM), as discussed in Section 3.3.2, corresponds to the MAP approximation of latent variables, while the forward–backward algorithm, as discussed in Section 3.3.1, corresponds to an exact inference of these variables. As another example, the MAP decision rule ($\arg\max_C p(C|\mathbf{O})$) in Eq. (3.2) also corresponds to the MAP approximation of inferring the posterior distribution of classification categories. Since the final goal of automatic speech recognition is to output the word sequence, the MAP approximation of the word sequence matches the final goal.[1] Thus, the MAP approximation can be applied to all probabilistic variables in speech and language processing as an essential technique.

This chapter starts to discuss the MAP approximation of Bayesian inference in detail, but limits the discussion only to model parameters Θ in Section 4.1. In the MAP approximation for model parameters, the prior distributions work as a regularization of these parameters, which makes the estimation of the parameters more robust than that of the maximum likelihood (ML) approach. Another interesting property of the MAP approximation for model parameters is that we can easily involve the inference of latent variables by extending the EM algorithm from ML to MAP estimation. Section 4.2 describes the general EM algorithm with the MAP approximation by following the ML-based EM algorithm, as discussed in Section 3.4. Based on the general MAP–EM algorithm, Section 4.3 provides MAP–EM solutions for CDHMM parameters, and introduces the well-known applications based on speaker adaptation. Section 4.5 describes the parameter smoothing method in discriminative training of the CDHMM, which actually corresponds to the MAP solution for discriminative parameter estimation. Section 4.6 focuses on the MAP estimation of GMM parameters, which is a subset of the MAP estimation of CDHMM parameters. It is used to construct speaker GMMs that are used

[1] However, if we consider some other spoken language processing applications given automatic speech recognition inputs (e.g., dialog, machine translation, and information retrieval), we need to consider how to provide $p(W|\mathbf{O})$ rather than $\hat{W} = \arg\max_W p(W|\mathbf{O})$ to avoid propagating any speech recognition errors to the post-processing applications.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

for speaker verification. Section 4.7 provides an MAP solution of n-gram parameters that leads to one instance of interpolation smoothing, as discussed in Section 3.6.2. Finally, Section 4.8 deals with the adaptive MAP estimation of latent topic model parameters.

4.1 MAP criterion for model parameters

This section begins with a general discussion of the MAP approximation for model parameters Θ. For simplicity, we first review the posterior distribution of model parameters given observations $\mathbf{O}$ without latent variables Z. Instead of estimating posterior distributions, the MAP estimation focuses on the following parameter estimation:

$$\Theta^{\text{MAP}} = \arg\max_{\Theta} p(\Theta|\mathbf{O}). \tag{4.1}$$

This corresponds to estimating the model parameter Θ^{MAP} given training data $\mathbf{O}$. By using the product and sum rules, as discussed in Section 2.1.1, we can rewrite the above equation as follows:

$$\begin{aligned}\Theta^{\text{MAP}} &= \arg\max_{\Theta} p(\Theta|\mathbf{O}) \\ &= \arg\max_{\Theta} \frac{p(\mathbf{O}|\Theta)p(\Theta)}{\int p(\mathbf{O}|\Theta)p(\Theta)d\Theta} \\ &= \arg\max_{\Theta} \underbrace{p(\mathbf{O}|\Theta)}_{\text{likelihood}} \times \underbrace{p(\Theta)}_{\text{prior}}.\end{aligned} \tag{4.2}$$

Since $p(\mathbf{O}) = \int p(\mathbf{O}|\Theta)p(\Theta)d\Theta$ does not depend on Θ, we can avoid computing this integral directly.[2] Furthermore, if we use an exponential family distribution for a likelihood function and a conjugate distribution for a prior distribution, as discussed in Section 2.1.3, the MAP estimate is represented as the mode of the corresponding conjugate posterior distribution, analytically. This is an advantage of using conjugate distributions.[3]

Equation (4.2) also tells us that the posterior distribution is composed of the likelihood function and the prior distribution, thus the estimation is based on the maximum likelihood function with the additional contribution of the prior distribution. That is, the prior distribution acts to regularize model parameters in the ML estimation, as we discussed in Section 2.3.1 as the best-known Bayesian advantage over ML. For example, let us consider the likelihood function $p(O|\Theta) = \prod_t \mathcal{N}(o_t|\mu, 1)$ as a one-dimensional Gaussian distribution with mean μ and precision as 1, and the prior distribution $p(\Theta)$ as a

[2] This term is called the evidence, which is neglected in the MAP approximation. However, the importance of the evidence is discussed in Chapter 5.

[3] In other words, it is not simple to obtain the mode of the posterior distribution, if we do not use the conjugate distribution, since we cannot obtain the posterior distribution analytically. For example, if we use the Laplace distribution for the prior distribution, the mode of posterior distributions cannot be obtained analytically, and we need some numerical computation.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

one-dimensional Gaussian distribution of the mean vector μ^0 and the scale parameter r. Then, the MAP estimation can be represented as follows:

$$\begin{aligned}
&\arg\max_{\mu} \log p(O|\Theta) + \log p(\Theta) \\
&\qquad\qquad = \arg\max_{\mu} \log\left(\prod_t \mathcal{N}(o_t|\mu, 1)\right) + \log \mathcal{N}(\mu|\mu^0, r^{-1}) \\
&\qquad\qquad = \arg\max_{\mu} \sum_t (o_t - \mu)^2 + r(\mu - \mu^0)^2,
\end{aligned} \tag{4.3}$$

where from the second to the third lines, we use the definition of a Gaussian distribution (Appendix C.5) as follows:

$$\mathcal{N}(x|\mu, r^{-1}) \triangleq (2\pi)^{-\frac{1}{2}}(r)^{\frac{1}{2}} \exp\left(-\frac{r(x-\mu)^2}{2}\right). \tag{4.4}$$

Thus, the optimization problem of the MAP solution corresponds to solving the minimum mean square error (MMSE) estimation with the l^2 regularization term around μ^0. The scale parameter r behaves as a tuning parameter to balance the MMSE estimation and the regularization term. These parameters are called regularization parameters, which can be hyperparameters of the prior distribution. Equation (4.3) can be analytically solved by using the conjugate distribution rule, as discussed in Section 2.1.4, or by using the following derivative method:

$$\begin{aligned}
\frac{\partial}{\partial\mu} \sum_{t=1}^{T} (o_t - \mu)^2 + r(\mu - \mu^0)^2 &= -2\sum_{t=1}^{T}(o_t - \mu) + 2r(\mu - \mu^0) \\
&= -2\left(\sum_{t=1}^{T} o_t + r\mu^0\right) + 2(T + r)\mu = 0.
\end{aligned} \tag{4.5}$$

We obtain the MAP estimate of μ analytically as:

$$\begin{aligned}
\mu^{\mathrm{MAP}} &= \frac{\sum_t^T o_t + r\mu^0}{T + r} \\
&= \frac{\mu^{\mathrm{ML}} + \frac{r}{T}\mu^0}{1 + \frac{r}{T}}.
\end{aligned} \tag{4.6}$$

Thus, the regularization term sets a constraint for the ML estimate $\mu^{\mathrm{ML}} = \frac{\sum_t^T o_t}{T}$ with a regularization constant r.

Similarly, if we use a Laplace distribution as a prior distribution, the prior distribution works as an l^1 regularization term. The Laplace distribution is defined as follows (Appendix C.10):

$$\mathrm{Lap}(x|\mu, \beta) \triangleq \frac{1}{2\beta} \exp\left(-\frac{|x - \mu|}{\beta}\right). \tag{4.7}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Therefore, using $\mathrm{Lap}(\mu|\mu^0, \beta)$ instead of $\mathcal{N}(\mu|\mu^0, r^{-1})$, Eq. (4.3) is rewritten as follows:

$$\begin{aligned}\arg\max_{\mu} \log p(O|\Theta) + \log p(\Theta) \\ &= \arg\max_{\mu} \log\left(\prod_t \mathcal{N}(o_t|\mu, 1)\right) + \log \mathrm{Lap}(\mu|\mu^0, \beta) \\ &= \arg\max_{\mu} \sum_t (o_t - \mu)^2 + \frac{1}{\beta}|\mu - \mu^0|. \end{aligned} \tag{4.8}$$

Thus, the prior distribution effect in the MAP parameter estimation is often regarded as a regularization of parameters. Consequently, Eq. (4.3) can incorporate the prior knowledge of parameters via hyperparameters $\boldsymbol{\mu}^0$, and the MAP approximation retains the Bayesian advantage of use of prior knowledge, as discussed in Section 2.3.1. Note that Eq. (4.8) is not differentiable with respect to μ, and it does not have a well-defined conjugate distribution. Therefore, the MAP estimation with the Laplace prior (l^1 regularization) is often undertaken by a numerical method.

Now, we introduce a useful mathematical operation for the MAP approximation of model parameters. To compute the expected values of the posterior distribution with respect to model parameters Θ, the MAP approximation can use the following posterior distribution represented by a Dirac delta function:

$$p(\Theta|\mathbf{O}) = \delta(\Theta - \Theta^{\mathrm{MAP}}), \tag{4.9}$$

where the Dirac delta function has the following property:

$$\int f(\mathbf{a})\delta(\mathbf{a} - \mathbf{a}^*) = f(\mathbf{a}^*). \tag{4.10}$$

This posterior distribution intuitively corresponds to having a location parameter with the MAP estimate Θ^{MAP} and very small (0) variance. If the model parameters are represented by discrete variables, we can use the Kronecker delta function. Therefore, once we obtain the MAP estimation of model parameters Θ^{MAP}, we can compute the expected value of function $f(\Theta)$ based on Eq. (4.9) as follows:

$$\begin{aligned}\mathbb{E}_{(\Theta)}[f(\Theta)|\mathbf{O}] &= \int f(\Theta)p(\Theta)d\Theta = \int f(\Theta)\delta(\Theta - \Theta^{\mathrm{MAP}})d\Theta \\ &= f(\Theta^{\mathrm{MAP}}). \end{aligned} \tag{4.11}$$

Here we use Eq. (4.10) to calculate the integral. Since this is equivalent to just plugging in the MAP estimates to the $f(\Theta)$, this procedure is called plug-in MAP (Lee & Huo 2000). For example, if we use the likelihood function of unseen data $\mathbf{O}'$ for $f(\Theta)$, Eq. (4.11) is rewritten as follows:

$$\mathbb{E}_{(\Theta)}[p(\mathbf{O}'|\Theta)|\mathbf{O}] = p(\mathbf{O}'|\Theta^{\mathrm{MAP}}). \tag{4.12}$$

That is, the likelihood function of unseen data can be obtained by simply replacing Θ with the MAP estimate Θ^{MAP}. This can be used as a likelihood function to compute likelihood values in prediction and classification steps. The Dirac delta-function-based

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

posterior representation is very useful, since the representation connects the analytical relationship between the MAP-based *point* estimation and the Bayesian *distribution* estimation.

Thus, the MAP approximation does not need to solve the marginalization explicitly in the training and prediction/classification steps. Although the approximation lacks the Bayesian advantages of the model selection and marginalization, as discussed in Section 2.3, it still has the most effective Bayesian advantage over ML, namely *use of prior knowledge*. Equation (4.3) also shows that the effect of the prior distribution in the MAP estimation works as a regularization. Therefore, the MAP approximation is widely used in practical Bayesian applications. In addition, the MAP approximation is simply extended to deal with latent variables based on the EM algorithm, which is a key technique in training statistical models in speech and language processing, as we discussed in Chapter 3. The next section discusses the MAP version of the EM algorithm.

4.2 MAP extension of EM algorithm

As we discussed in Section 3.4, most statistical models used in speech and language processing have to deal with latent variables Z, e.g., HMM states and mixture components of the CDHMM in acoustic modeling, and latent topics in language modeling. The maximum likelihood approach has an efficient solution based on the EM algorithm, which optimizes the auxiliary function $Q(\Theta'|\Theta)$ instead of a (log) likelihood function. This section describes the EM extension of MAP parameter estimation in general.

4.2.1 Auxiliary function

Following the discussion in Section 3.4, we prove that the EM steps ultimately lead to the local optimum value Θ^{MAP} in terms of the MAP criterion. First, since the logarithmic function is a monotonic function, the MAP criterion in Eq. (4.2) is represented as follows:

$$\begin{aligned}\Theta^{\text{MAP}} &= \arg\max_{\Theta} p(\mathbf{O}|\Theta)p(\Theta)\\ &= \arg\max_{\Theta} \log\left(p(\mathbf{O}|\Theta)p(\Theta)\right). \end{aligned} \tag{4.13}$$

By introducing latent variable Z, the above equation can be written as

$$\Theta^{\text{MAP}} = \arg\max_{\Theta} \log\left(\sum_{Z} p(\mathbf{O}, Z|\Theta)p(\Theta)\right). \tag{4.14}$$

As discussed in the ML–EM algorithm, the summation over latent variable $\sum_Z$ is computationally very difficult since the latent variable in speech and language processing is represented as a possible sequence, and the number of these variables is exponential. Therefore, we need to avoid having to compute $\sum_Z$ directly.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Similarly to the ML–EM algorithm in Section 3.4, in M-step, we maximize the MAP version of the auxiliary function with respect to the parameters Θ, and estimate new parameters by

$$\Theta^{\mathrm{MAP}} = \arg\max_{\Theta'} Q^{\mathrm{MAP}}(\Theta'|\Theta). \tag{4.15}$$

The updated parameters Θ' are then treated as the current parameters for the next iteration of EM steps. The $Q^{\mathrm{MAP}}(\Theta'|\Theta)$ is defined as the expectation of the joint distribution $p(\mathbf{O}, Z, \Theta')$ with respect to $p(Z|\mathbf{O}, \Theta)$ as follows:

$$\begin{aligned} Q^{\mathrm{MAP}}(\Theta'|\Theta) &\triangleq \mathbb{E}_{(Z)}[\log p(\mathbf{O}, Z, \Theta')|\mathbf{O}, \Theta] \\ &= \mathbb{E}_{(Z)}[\log p(\mathbf{O}, Z|\Theta')p(\Theta')|\mathbf{O}, \Theta] \\ &= \underbrace{\mathbb{E}_{(Z)}[\log p(\mathbf{O}, Z|\Theta')|\mathbf{O}, \Theta]}_{Q^{\mathrm{ML}}(\Theta'|\Theta)} + \log p(\Theta'), \end{aligned} \tag{4.16}$$

where $\log p(\Theta')$ does not depend on Z, and can be separated from the expectation operation. Compared with the ML auxiliary function $Q^{\mathrm{ML}}(\Theta'|\Theta)$ (Eq. (3.78)), we have an additional term $\log p(\Theta')$, which comes from a prior distribution of model parameters.

Now we explain how optimization of the auxiliary function Q^{MAP} leads to the local optimization of $p(\mathbf{O}|\Theta)p(\Theta)$ or $p(\Theta|\mathbf{O})$. For the explanation, we define the logarithmic function of the joint distribution $p(\mathbf{O}, \Theta') = p(\mathbf{O}|\Theta')p(\Theta')$ as follows:

$$L^{\mathrm{MAP}}(\Theta') \triangleq \log\left(p(\mathbf{O}|\Theta)p(\Theta)\right). \tag{4.17}$$

This is similar to $L(\Theta')$ in Eq. (3.83), but $L^{\mathrm{MAP}}(\Theta')$ has an additional factor from $p(\Theta)$. Now, we represent $p(\mathbf{O}|\Theta)$ in the above equation with the distributions of latent variable Z based on the product rule of probabilistic variables, as follows:

$$p(\mathbf{O}|\Theta') = \frac{p(\mathbf{O}, Z|\Theta')}{p(Z|\mathbf{O}, \Theta')}. \tag{4.18}$$

Therefore, by substituting Eq. (4.18) into Eq. (4.17), we obtain

$$L^{\mathrm{MAP}}(\Theta') = \log p(\mathbf{O}, Z|\Theta') - \log p(Z|\mathbf{O}, \Theta') + \log p(\Theta'). \tag{4.19}$$

Now we perform the expectation operation with respect to $p(Z|\mathbf{O}, \Theta)$ for both sides of the equation, and obtain the following relationship:

$$\begin{aligned} L^{\mathrm{MAP}}(\Theta') &= \mathbb{E}_{(Z)}[\log p(\mathbf{O}, Z|\Theta')|\mathbf{O}, \Theta] - \mathbb{E}_{(Z)}[\log p(Z|\mathbf{O}, \Theta')|\mathbf{O}, \Theta] + \log p(\Theta') \\ &= \underbrace{\mathbb{E}_{(Z)}[\log p(\mathbf{O}, Z|\Theta')|\mathbf{O}, \Theta] + \log p(\Theta')}_{Q^{\mathrm{MAP}}(\Theta'|\Theta)} \underbrace{- \mathbb{E}_{(Z)}[\log p(Z|\mathbf{O}, \Theta')|\mathbf{O}, \Theta]}_{H(\Theta'|\Theta)}, \end{aligned} \tag{4.20}$$

where $L^{\mathrm{MAP}}(\Theta')$ and $\log p(\Theta')$ are not changed since these do not depend on Z. Note that the third term of Eq. (4.20) is represented as $H(\Theta'|\Theta)$, which is exactly the same definition as Eq. (3.85). Thus, we derive a similar equation to that of the ML auxiliary function Eq. (3.86):

$$Q^{\mathrm{MAP}}(\Theta'|\Theta) = L(\Theta')^{\mathrm{MAP}} + H(\Theta'|\Theta). \tag{4.21}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Since $H(\Theta'|\Theta)$ is the same as in Eq. (3.86) and has a bound based on the Jensen's inequality, we can apply the same discussion to $Q^{\mathrm{MAP}}(\Theta'|\Theta)$ to show that $Q^{\mathrm{MAP}}(\Theta'|\Theta)$ is the auxiliary function of the MAP criterion.

Thus, we prove that the E-step performing the expectation and M-step maximizing the auxiliary function with respect to the model parameters Θ, always increase the joint likelihood value as:

$$\Theta^{\mathrm{MAP}} = \arg\max_{\Theta'} Q^{\mathrm{MAP}}(\Theta'|\Theta) \Rightarrow p(\mathbf{O}, \Theta^{\mathrm{MAP}}) \geq p(\mathbf{O}, \Theta). \tag{4.22}$$

This leads to a local optimization of the joint likelihood function $p(\mathbf{O}, \Theta)$, which corresponds to the MAP criterion in Eq. (4.13). We call this procedure the MAP–EM algorithm.

4.2.2 A recipe

Based on the previous discussions, we summarize in the text box a general procedure to obtain the MAP estimation of model parameters. The following section describes the concrete form of MAP–EM solutions for CDHMMs similarly to Section 3.4.

1. Set a likelihood function for a statistical model (generative model) with model parameters.
2. Set appropriate prior distributions for model parameters (possibly conjugate distributions to obtain analytical results based on the conjugate distribution discussion in Section 2.1.4).
3. Solve the parameter estimation by maximizing the MAP objective function:
   i. Solve posterior distributions for model parameters when we can use conjugate priors;
   ii. Solve the parameter estimation by getting the modes of posterior distributions.

4.3 Continuous density hidden Markov model

This section describes the MAP estimation of HMM parameters based on the MAP–EM algorithm (Lee *et al.* 1991, Gauvain & Lee 1994). Following the general procedure for MAP estimation (as set out in Section 4.2.2), we first review a likelihood function of the CDHMM, as discussed in Section 3.2.3. Then we provide a concrete form of the prior distribution $p(\Theta)$ for full and diagonal covariance cases. Then, according to the derivation of the ML–EM algorithm in Section 3.4, we also derive the concrete form solutions of the MAP–EM algorithm of the CDHMM.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

4.3.1 Likelihood function

We first provide the complete data likelihood function with speech feature sequence $\mathbf{O} = \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \cdots, T\}$, HMM state sequence $S = \{s_t \in \{1, \cdots, J\} | t = 1, \cdots, T\}$, and GMM component sequence $V = \{v_t \in \{1, \cdots, K\} | t = 1, \cdots, T\}$, which is introduced in Eq. (3.43) as follows:

$$p(\mathbf{O}, S, V|\Theta) = \pi_{s_1} \omega_{s_1 v_1} p(\mathbf{o}_1 | \Theta_{s_1 v_1}) \prod_{t=2}^{T} a_{s_{t-1} s_t} \omega_{s_t v_t} p(\mathbf{o}_t | \Theta_{s_t v_t}). \tag{4.23}$$

Recall that a set of HMM parameters Θ holds:

- Initial state probability π_j;
- State transition probability a_{ij};
- Mixture weight ω_{jk};
- Gaussian mean vector $\boldsymbol{\mu}_{jk}$;
- Gaussian covariance matrix $\boldsymbol{\Sigma}_{jk}$.

The next section provides prior distribution $p(\Theta)$.

4.3.2 Conjugate priors (full covariance case)

The prior distribution is considered to be the following joint distribution form:

$$p(\Theta) = p\left(\{\pi_j\}_{j=1}^{J}), \{\{a_{ij}\}_{i=1}^{J}\}_{j=1}^{J}, \{\{\omega_{jk}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\}_{j=1}^{J}\}_{k=1}^{K}\right). \tag{4.24}$$

However, since it is difficult to handle this joint distribution, we usually factorize it by assuming conditional independence for each HMM state and mixture component. Then, the prior distribution is rewritten as follows:

$$\begin{aligned} p(\Theta) &= p(\boldsymbol{\pi}) p(A) p(\boldsymbol{\omega}) p(\boldsymbol{\mu}, \mathbf{R}) \\ &= p(\{\pi_j\}_{j=1}^{J}) \left(\prod_{i=1}^{J} p(\{a_{ij}\}_{j=1}^{J}) \right) \left(\prod_{j=1}^{J} p(\{\omega_{jk}\}_{k=1}^{K}) \right) \left(\prod_{j=1}^{J} \prod_{k=1}^{K} p(\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \right), \end{aligned} \tag{4.25}$$

where we also assume that π_j, a_{ij}, ω_{jk}, and $\{\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\}$ are independent of each other, although we keep the dependency of $\boldsymbol{\mu}_{jk}$ and $\boldsymbol{\Sigma}_{jk}$.

Now we provide the concrete forms of the above prior distributions for the HMM parameters based on the conjugate distribution discussion in Section 2.1.4. We first focus on the prior distributions of the initial state probability π_j, state transition probability a_{ij}, and Gaussian mixture weight ω_{jk}. Note that these probabilistic variables have the same constraint that $\pi_j \geq 0, \sum_{j=1}^{J} \pi_j = 1$, $a_{ij} \geq 0, \sum_{j=1}^{J} a_{ij} = 1$, and $\omega_{jk} \geq 0, \sum_{k=1}^{K} \omega_{jk} = 1$. In addition, these are represented by a multinomial distribution in the complete data likelihood function. Therefore, according to Table 2.1, these are represented by a Dirichlet distribution with hyperparameters as follows:

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$
\begin{aligned}
p(\{\pi_j\}_{j=1}^J) &= \mathrm{Dir}(\{\pi_j\}_{j=1}^J|\{\phi_j^\pi\}_{j=1}^J), \\
p(\{a_{ij}\}_{j=1}^J) &= \mathrm{Dir}(\{a_{ij}\}_{j=1}^J|\{\phi_{ij}^a\}_{j=1}^J), \\
p(\{\omega_{jk}\}_{k=1}^K) &= \mathrm{Dir}(\{\omega_{jk}\}_{k=1}^K|\{\phi_{jk}^\omega\}_{k=1}^K), \qquad (4.26)
\end{aligned}
$$

where $\phi_j^\pi \geq 0$, $\phi_{ij}^a \geq 0$, and $\phi_{jk}^\omega \geq 0$.

Next we consider the prior distribution of Gaussian mean vector $\boldsymbol{\mu}_{jk}$ and Gaussian precision matrix $\boldsymbol{\Sigma}_{jk}$. For simplicity, we focus on the precision matrix $\mathbf{R}$, which is the inverse matrix of the covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$
\mathbf{R}_{jk} \triangleq (\boldsymbol{\Sigma}_{jk})^{-1}. \qquad (4.27)
$$

According to Table 2.1, the joint prior distribution of Gaussian mean vector $\boldsymbol{\mu}_{jk}$ and Gaussian precision matrix $\mathbf{R}_{jk}$ can be written as follows:

$$
\begin{aligned}
p(\boldsymbol{\mu}_{jk}, \mathbf{R}_{jk}) &= p(\boldsymbol{\mu}_{jk}|\mathbf{R}_{jk})p(\mathbf{R}_{jk}) \\
&= \mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\mu}_{jk}^0, (\phi_{jk}^{\boldsymbol{\mu}}\mathbf{R}_{jk})^{-1})\mathcal{W}(\mathbf{R}_{jk}|\mathbf{R}_{jk}^0, \phi_{jk}^{\mathbf{R}}), \qquad (4.28)
\end{aligned}
$$

where $\mathcal{W}(\cdot)$ is a Wishart distribution, which is defined in Appendix C.14. Note that the prior distribution $p(\boldsymbol{\mu}_{jk}|\mathbf{R}_{jk})$ of the mean vector depends on covariance matrix $\mathbf{R}_{jk}$, and cannot be factorized independently. Instead, these parameters are represented by the joint prior distribution $p(\boldsymbol{\mu}_{jk}, \mathbf{R}_{jk})$ with the Gaussian–Wishart distribution (Appendix C.15) or the product form in Eq. (4.28), as we discussed in Section 2.1.4. We can also provide the prior distribution for the original covariance matrix $\boldsymbol{\Sigma}$ by using the inverse-Wishart distribution instead of the Wishart distribution in Eq. (4.28). Both representations yield the same result in the MAP estimation of HMM parameters.

Consequently, the conjugate prior distribution of a CDHMM is represented by the following factorization form with each parameter:

$$
\begin{aligned}
p(\Theta) = {} & \mathrm{Dir}(\{\pi_j\}_{j=1}^J|\{\phi_j^\pi\}_{j=1}^J) \\
& \times \left(\prod_{i=1}^J \mathrm{Dir}(\{a_{ij}\}_{j=1}^J|\{\phi_{ij}^a\}_{j=1}^J)\right)\left(\prod_{j=1}^J \mathrm{Dir}(\{\omega_{jk}\}_{k=1}^K|\{\phi_{jk}^\omega\}_{k=1}^K)\right) \\
& \times \left(\prod_{j=1}^J \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\mu}_{jk}^0, (\phi_{jk}^{\boldsymbol{\mu}}\mathbf{R}_{jk})^{-1})\mathcal{W}(\mathbf{R}_{jk}|\mathbf{R}_{jk}^0, \phi_{jk}^{\mathbf{R}})\right). \qquad (4.29)
\end{aligned}
$$

Note that the prior distribution of a CDHMM is represented by three types of distributions, i.e., Dirichlet, Gaussian, and Wishart distributions. The prior distribution has five scalar hyperparameters $\phi^\pi, \phi^a, \phi^\omega, \phi^{\boldsymbol{\mu}}, \phi^{\mathbf{R}}$, one vector hyperparameter $\boldsymbol{\mu}^0$, and one matrix hyperparameter $\mathbf{R}^0$. A set of these hyperparameters is written as Ψ in this chapter, i.e.,

$$
\Psi \triangleq \{\phi_j^\pi, \phi_{ij}^a, \phi_{jk}^\omega, \phi_{jk}^{\boldsymbol{\mu}}, \phi_{jk}^{\mathbf{R}}, \boldsymbol{\mu}_{jk}^0, \mathbf{R}_{jk}^0 | i = 1, \cdots, J, j = 1, \cdots, J, k = 1, \cdots, K\}. \qquad (4.30)
$$

In the following sections, we sometimes represent the prior distribution as $p(\Theta|\Psi)$ instead of $p(\Theta)$ to deal with the hyperparameter dependency on the prior distributions explicitly.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

4.3.3 Conjugate priors (diagonal covariance case)

In practical use, the Gaussians in a CDHMM are often represented by a diagonal covariance matrix, as we discussed in Section 3.2.3. To deal with a conjugate distribution of a diagonal covariance matrix, we need to provide a specific distribution rather than the Wishart distribution since the off-diagonal elements are always zero, and it is not suitable to represent these random variables as the Wishart distribution. Instead, we use the gamma distribution for each diagonal component. We first define the $d - d$ element of the precision matrix as r_d:

$$r_d \triangleq [\mathbf{R}]_{dd}. \tag{4.31}$$

Then the joint prior distribution of Eq. (4.28) is factorized by a feature dimension, and it is replaced with the gamma distribution as follows:

$$\begin{aligned} p(\boldsymbol{\mu}_{jk}, \mathbf{R}_{jk}) &= \prod_{d=1}^{D} p(\mu_{jkd}|r_{jkd})p(r_{jkd}) \\ &= \prod_{d=1}^{D} \mathcal{N}(\mu_{jkd}|\mu^0_{jkd}, (\phi^{\boldsymbol{\mu}}_{jk} r_{jkd})^{-1}) \operatorname{Gam}(r_{jkd}|r^0_{jkd}, \phi^{\mathbf{R}}_{jk}), \end{aligned} \tag{4.32}$$

where a set of hyperparameters Ψ is represented as

$$\Psi \triangleq \{\phi^{\pi}_j, \phi^a_{ij}, \phi^{\omega}_{jk}, \phi^{\boldsymbol{\mu}}_{jk}, \phi^{\mathbf{R}}_{jk}, \boldsymbol{\mu}^0_{jk}, \mathbf{r}^0_{jk}|i = 1, \cdots, J, j = 1, \cdots, J, k = 1, \cdots, K\}, \tag{4.33}$$

where

$$\mathbf{r}^0 \triangleq [r^0_1, \cdots, r^0_D]^{\intercal}. \tag{4.34}$$

Similarly to the full covariance case, Eq. (4.32) can also be represented by a Gaussian-gamma distribution (Appendix C.13).

The dependency of the hyperparameter ϕ is not unique and can be arranged by considering applications due to the flexible parameterization of an exponential family distribution, as discussed in Section 2.1.3. For example, $\phi^{\boldsymbol{\mu}}_{jk}$ and $\phi^{\mathbf{R}}_{jk}$ can be changed depending on a dimension d (i.e., $\phi^{\boldsymbol{\mu}}_{jkd}$, $\phi^{\mathbf{R}}_{jkd}$). These make the model more precise, but need more effort to set $\phi^{\boldsymbol{\mu}}_{jkd}$, $\phi^{\mathbf{R}}_{jkd}$ for all dimensions manually or automatically. Actually, these values are often shared among all js and ks (i.e., $\phi^{\boldsymbol{\mu}}_{jk} \to \phi^{\boldsymbol{\mu}}$ etc.).

4.3.4 Expectation step

Once we set the prior distributions and likelihood function, by following the recipe in Section 4.2.2, we can perform the MAP–EM algorithm to estimate the model parameter Θ. This section considers the concrete form of the MAP expectation step. This procedure is very similar to Section 3.4 except for the additional consideration of the prior distribution $p(\Theta)$. The auxiliary function used in the MAP–EM algorithm is represented as

$$Q^{\mathrm{MAP}}(\Theta'|\Theta) = Q^{\mathrm{ML}}(\Theta'|\Theta) + \log p(\Theta'). \tag{4.35}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

According to Section 3.4, the auxiliary function $Q^{\mathrm{ML}}(\Theta'|\Theta)$ is factorized by a sum of four individual auxiliary functions as:

$$\begin{aligned}Q^{\mathrm{ML}}(\Theta'|\Theta) &= \sum_{S}\sum_{V} p(S, V|\mathbf{O}, \Theta)\Bigg[\log \pi'_{s_1} + \log \omega'_{s_1 v_1} + \log p(\mathbf{o}_1|\boldsymbol{\mu}'_{s_1 v_1}, \boldsymbol{\Sigma}'_{s_1 v_1}) \\ &\quad + \sum_{t=2}^{T}\Big(\log a'_{s_{t-1}s_t} + \log \omega'_{s_t v_t} + \log p(\mathbf{o}_t|\boldsymbol{\mu}'_{s_t v_t}, \boldsymbol{\Sigma}'_{s_t v_t})\Big)\Bigg] \\ &= Q^{\mathrm{ML}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) + Q^{\mathrm{ML}}(A'|A) + Q^{\mathrm{ML}}(\boldsymbol{\omega}'|\boldsymbol{\omega}) + Q^{\mathrm{ML}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}).\end{aligned} \tag{4.36}$$

Similarly, from Eq. (4.29), the prior distribution of all model parameters $p(\Theta)$ can be decomposed into the four individual prior distributions as

$$\begin{aligned}\log p(\Theta) &= \log \underbrace{\Big(\mathrm{Dir}(\{\pi_j\}_{j=1}^{J}|\{\phi_j^{\pi}\}_{j=1}^{J})\Big)}_{\triangleq p(\boldsymbol{\pi})} + \log \underbrace{\prod_{i=1}^{J}\Big(\mathrm{Dir}(\{a_{ij}\}_{j=1}^{J}|\{\phi_{ij}^{a}\}_{j=1}^{J})\Big)}_{\triangleq p(A)} \\ &\quad + \log \underbrace{\prod_{i=1}^{J}\Big(\mathrm{Dir}(\{\omega_{jk}\}_{k=1}^{K}|\{\phi_{jk}^{\omega}\}_{k=1}^{K})\Big)}_{\triangleq p(\boldsymbol{\omega})} \\ &\quad + \log \underbrace{\prod_{j=1}^{J}\prod_{k=1}^{K}\Big(\mathcal{N}(\boldsymbol{\mu}_{jk}|\boldsymbol{\mu}_{jk}^{0}, (\phi_{jk}^{\boldsymbol{\mu}}\mathbf{R}_{jk})^{-1})\mathcal{W}(\mathbf{R}_{jk}|\mathbf{R}_{jk}^{0}, \phi_{jk}^{\mathbf{R}})\Big)}_{\triangleq p(\boldsymbol{\mu}, \mathbf{R})}.\end{aligned} \tag{4.37}$$

Therefore, by using the factorization forms of Eqs. (4.36) and (4.37), similarly to the ML case, Eq. (4.36) is also represented as a sum of four individual auxiliary functions defined as follows:

$$\begin{aligned}Q^{\mathrm{MAP}}(\Theta'|\Theta) = Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) &+ Q^{\mathrm{MAP}}(A'|A) \\ &+ Q^{\mathrm{MAP}}(\boldsymbol{\omega}'|\boldsymbol{\omega}) + Q^{\mathrm{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}),\end{aligned} \tag{4.38}$$

where

$$\begin{aligned}Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) &= Q^{\mathrm{ML}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}) \\ &= \sum_{j=1}^{J}\xi_1(j)\log \pi'_j + \log\Big(\mathrm{Dir}(\{\pi_j\}_{j=1}^{J}|\{\phi_j^{\pi}\}_{j=1}^{J})\Big),\end{aligned} \tag{4.39}$$

$$\begin{aligned}Q^{\mathrm{MAP}}(A'|A) &= Q^{\mathrm{ML}}(A'|A) + \log p(A) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{J}\sum_{j=1}^{J}\xi_t(i,j)\log a'_{ij} + \sum_{i=1}^{J}\log\Big(\mathrm{Dir}(\{a'_{ij}\}_{j=1}^{J}|\{\phi_{ij}^{a}\}_{j=1}^{J})\Big),\end{aligned} \tag{4.40}$$

$$\begin{aligned} Q^{\mathrm{MAP}}(\boldsymbol{\omega}'|\boldsymbol{\omega}) &= Q^{\mathrm{ML}}(\boldsymbol{\omega}'|\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K} \gamma_t(j,k) \log \omega'_{j,k} + \sum_{j=1}^{J} \log \left(\mathrm{Dir}(\{\omega'_{jk}\}_{k=1}^{K} | \{\phi^{\omega}_{jk}\}_{k=1}^{K}) \right), \end{aligned} \tag{4.41}$$

$$\begin{aligned} Q^{\mathrm{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}) &= Q^{\mathrm{ML}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}) + \log p(\boldsymbol{\mu}, \mathbf{R}) \\ &\propto \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\gamma_t(j,k)}{2} \left[\log |\mathbf{R}'_{jk}| - (\mathbf{o}_t - \boldsymbol{\mu}'_{jk})^{\mathsf{T}} \mathbf{R}'_{jk} (\mathbf{o}_t - \boldsymbol{\mu}'_{jk}) \right] \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} \log \left(\mathcal{N}(\boldsymbol{\mu}'_{jk} | \boldsymbol{\mu}^0_{jk}, (\phi^{\boldsymbol{\mu}}_{jk} \mathbf{R}'_{jk})^{-1}) \mathcal{W}(\mathbf{R}'_{jk} | \mathbf{R}^0_{jk}, \phi^{\mathbf{R}}_{jk}) \right), \end{aligned} \tag{4.42}$$

where $\xi_1(j)$, $\xi_t(i,j)$, and $\gamma_t(j,k)$ are the posterior probabilities, which are introduced in Eqs. (3.99), (3.118), (3.119) as follows:

$$\begin{aligned} \xi_1(j) &\triangleq p(s_1 = j | \mathbf{O}, \Theta) \\ \xi_t(i,j) &\triangleq p(s_t = i, s_{t+1} = j | \mathbf{O}, \Theta) \\ \gamma_t(j,k) &\triangleq p(s_t = j, v_t = k | \mathbf{O}, \Theta). \end{aligned} \tag{4.43}$$

Note that Θ is estimated by using the MAP estimation Θ^{MAP} for Θ instead of the ML estimation, which is discussed in Section 4.3.6.

We can also obtain the auxiliary function of diagonal covariance Gaussians instead of Eq. (4.42) by using a Gaussian-gamma distribution as a prior distribution, as discussed in Eq. (4.32) as follows:

$$\begin{aligned} & Q^{\mathrm{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}) \\ & \propto \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{d=1}^{D} \frac{\gamma_t(j,k)}{2} \left[\log r'_{jk} - (o_{td} - \mu'_{jkd})^2 r'_{jkd} \right] \\ & \quad + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{d=1}^{D} \log \left(\mathcal{N}(\mu'_{jkd} | \mu^0_{jkd}, (\phi^{\boldsymbol{\mu}}_{jk} r'_{jkd})^{-1}) \mathrm{Gam}_2 \left(r'_{jkd} \middle| \phi^{\mathbf{R}}_{jk}, r_{jkd} \right) \right). \end{aligned} \tag{4.44}$$

Here we use the gamma distribution $\mathrm{Gam}_2(y|\phi, r^0)$ described in Eq. (C.81) instead of the original gamma distribution defined in Eq. (C.74), which provides a good relationship with the Wishart distribution, i.e., if $\mathbf{R}$ is a scalar value (the number of dimension $D = 1$), the hyperparameters of the Wishart distribution become the same as ϕ^{r}_{jkd} and r^0_{jk},[4] as we discussed in Example 2.6. Note that the vector and matrix operations in Eq. (4.42) are represented as scalar operations with the summation over the dimension. This is a very good property for which to obtain the analytical solutions due to the simplicity of the scalar calculations. In addition, this representation avoids vector and matrix computations, which also makes implementation simple. This section provides

[4] We can set a hyperparameter ϕ which depends on each element d, i.e., $\phi^{\mathbf{R}}_{jk} \to \phi^{\mathbf{R}}_{jkd}$. Considering the compatibility with the Wishart distribution, this book uses $\phi^{\mathbf{R}}_{jk}$, which is independent of d.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

both full and diagonal covariance solutions, but the diagonal covariance solution is used for most of our applications.

4.3.5 Maximization step

The maximization step in the ML–EM algorithm obtains the maximum values of parameters by using derivative techniques, as discussed in Section 3.4.3. In this section, we provide other solutions for this problem which:

1. Calculate the posterior distributions;
2. Obtain the mode values of the posterior distributions, which are used as the MAP estimates.

In general, it is difficult to analytically obtain the posterior distributions. However, since we use the conjugate prior distributions of a CDHMM, as discussed in Section 2.1.3, we can easily obtain the posterior distributions for these problems.

Initial weight

We first focus on $Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi})$ in Eq. (4.39):

$$Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) = \sum_{j=1}^{J} \xi_1(j) \log \pi'_j + \log \left(\mathrm{Dir}(\{\pi'_j\}_{j=1}^J | \{\phi_j^\pi\}_{j=1}^J) \right). \tag{4.45}$$

Recall that the Dirichlet distribution (Appendix C.4) is represented as follows:

$$\mathrm{Dir}(\{\pi_j\}_{j=1}^J | \{\phi_j^\pi\}_{j=1}^J) = C_{\mathrm{Dir}}(\{\phi_j^\pi\}_{j=1}^J) \prod_{j=1}^{J} (\pi_j)^{\phi_j^\pi - 1}. \tag{4.46}$$

Then, by substituting Eq. (4.46) into Eq. (4.45), Eq. (4.45) is re-written as follows:

$$\begin{aligned} Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) &= \sum_{j=1}^{J} \xi_1(j) \log \pi'_j + (\phi_j^\pi - 1) \log \pi'_j + \log C_{\mathrm{Dir}}(\{\phi_j^\pi\}_{j=1}^J) \\ &= \sum_{j=1}^{J} (\xi_1(j) + \phi_j^\pi - 1) \log \pi'_j + \log C_{\mathrm{Dir}}(\{\phi_j^\pi\}_{j=1}^J) \\ &= \log \prod_{j=1}^{J} (\pi'_j)^{\xi_1(j) + \phi_j^\pi - 1} + \log C_{\mathrm{Dir}}(\{\phi_j^\pi\}_{j=1}^J). \end{aligned} \tag{4.47}$$

Comparing the result with Eq. (4.46), it is the same function form with different hyperparameters. Therefore, the auxiliary function $Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi})$ is represented by the following Dirichlet distribution:

$$\begin{aligned} Q^{\mathrm{MAP}}(\boldsymbol{\pi}'|\boldsymbol{\pi}) = \log \left(\mathrm{Dir}(\{\pi'_j\}_{j=1}^J | \{\hat{\phi}_j^\pi\}_{j=1}^J) \right) - \log C_{\mathrm{Dir}}(\{\hat{\phi}_j^\pi\}_{j=1}^J) \\ + \log C_{\mathrm{Dir}}(\{\phi_j^\pi\}_{j=1}^J) \end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$
\begin{aligned}
&= \log\left(\mathrm{Dir}(\{\pi'_j\}_{j=1}^J | \{\hat{\phi}^\pi_j\}_{j=1}^J)\right) + \log \frac{C_{\mathrm{Dir}}(\{\phi^\pi_j\}_{j=1}^J)}{C_{\mathrm{Dir}}(\{\hat{\phi}^\pi_j\}_{j=1}^J)} \\
&\propto \log\left(\mathrm{Dir}(\{\pi'_j\}_{j=1}^J | \{\hat{\phi}^\pi_j\}_{j=1}^J)\right), \qquad (4.48)
\end{aligned}
$$

where

$$\hat{\phi}^\pi_j \triangleq \phi^\pi_j + \xi_1(j). \qquad (4.49)$$

We finally omit the ratio of the normalization factor of the prior and posterior distributions, which do not depend on π'_j. Actually, this Dirichlet distribution corresponds to the posterior distribution of π with new hyperparameter $\hat{\phi}$. This result is similar to that of the conjugate prior and posterior distributions for multinomial likelihood function, as we discussed in Example 2.8.

Once we obtain the analytical form of the posterior distribution, the MAP estimate can be obtained as the mode of the Dirichlet distribution (Appendix C.4):

$$\pi_j^{\mathrm{MAP}} = \frac{\phi^\pi_j + \xi_1(j) - 1}{\sum_{j'=1}^J (\phi^\pi_{j'} + \xi_1(j') - 1)}. \qquad (4.50)$$

Thus, we obtain the MAP estimate of the initial weight π^{MAP}, which is proportional to the hyperparameter $\hat{\phi}^\pi$. We discuss the meaning of this solution in Section 4.3.7.

State transition

Similarly, the auxiliary function of state transition parameters A is obtained as follows:

$$
\begin{aligned}
Q^{\mathrm{MAP}}(A'|A) &= \log\left(\prod_{i=1}^J \mathrm{Dir}(\{a'_{ij}\}_{j=1}^J | \{\hat{\phi}^a_{ij}\}_{j=1}^J)\right) + \log\left(\prod_{i=1}^J \frac{C_{\mathrm{Dir}}(\{\phi^a_{ij}\}_{j=1}^J)}{C_{\mathrm{Dir}}(\{\hat{\phi}^a_{ij}\}_{j=1}^J)}\right) \\
&\propto \log\left(\prod_{i=1}^J \mathrm{Dir}(\{a'_{ij}\}_{j=1}^J | \{\hat{\phi}^a_{ij}\}_{j=1}^J)\right), \qquad (4.51)
\end{aligned}
$$

where

$$\hat{\phi}^a_{ij} \triangleq \phi^a_{ij} + \sum_{t=1}^{T-1} \xi_t(i,j). \qquad (4.52)$$

Therefore, the mode of the Dirichlet distribution is obtained as:

$$a_{ij}^{\mathrm{MAP}} = \frac{\phi^a_{ij} + \sum_{t=1}^{T-1} \xi_t(i,j) - 1}{\sum_{j'=1}^J (\phi^a_{ij'} + \sum_{t=1}^{T-1} \xi_t(i,j') - 1)}. \qquad (4.53)$$

The solution is similar to the initial weight in Eq. (4.50), and it is computed from the statistics of the accumulated posterior values of the state transition $\sum_{t=1}^{T-1} \xi_t(i,j)$ and prior parameter ϕ^a_{ij}.

Again, the result indicates that the auxiliary function of the state transition is represented by the same Dirichlet distribution as that used in the prior distribution with different hyperparameters. This result corresponds to the conjugate distribution analysis for multinomial distribution, as we discussed in Section 2.8. Therefore, although we need to handle the latent variables within a MAP–EM framework based on the iterative

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

calculation, the conjugate prior provides an analytic solution in the M-step, and it makes the estimation process efficient.

Mixture weight

Similar to the state transition, the auxiliary function of the mixture weight parameters $\boldsymbol{\omega}$ is as follows:

$$\begin{aligned} Q^{\text{MAP}}(\boldsymbol{\omega}'|\boldsymbol{\omega}) &= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\gamma_t(j,k)\log\omega'_{jk} + \log\left(\prod_{j=1}^{J}\text{Dir}(\{\omega'_{jk}\}_{k=1}^{K}|\{\phi^{\omega}_{jk}\}_{k=1}^{K})\right) \\ &= \log\left(\prod_{j=1}^{J}\text{Dir}(\{\omega'_{jk}\}_{k=1}^{K}|\{\hat{\phi}^{\omega}_{jk}\}_{k=1}^{K})\right) + \log\left(\prod_{j=1}^{J}\frac{C_{\text{Dir}}(\{\phi^{\omega}_{jk}\}_{k=1}^{K})}{C_{\text{Dir}}(\{\hat{\phi}^{\omega}_{jk}\}_{k=1}^{K})}\right) \\ &\propto \log\left(\prod_{j=1}^{J}\text{Dir}(\{\omega'_{jk}\}_{k=1}^{K}|\{\hat{\phi}^{\omega}_{jk}\}_{k=1}^{K})\right), \end{aligned} \tag{4.54}$$

where

$$\hat{\phi}^{\omega}_{jk} \triangleq \phi^{\omega}_{jk} + \sum_{t=1}^{T}\gamma_t(j,k). \tag{4.55}$$

Therefore, the mode of the Dirichlet distribution is obtained as:

$$\omega^{\text{MAP}}_{jk} = \frac{\phi^{\omega}_{jk} + \sum_{t=1}^{T}\gamma_t(j,k) - 1}{\sum_{k'=1}^{K}(\phi^{\omega}_{jk'} + \sum_{t=1}^{T}\gamma_t(j,k') - 1)}. \tag{4.56}$$

Again, it is computed from the statistics of the accumulated posterior values of the state occupancy $\sum_{t=1}^{T}\gamma_t(j,k)$ and prior parameter ϕ^{ω}_{jk}.

Mean vector and covariance matrix

Finally, we focus on the auxiliary function of the mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{R}$. Recall that the multivariate Gaussian distribution (Appendix C.6) is represented as follows:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{R}^{-1}) = C_{\mathcal{N}}(\mathbf{R}^{-1})\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{R}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{4.57}$$

and the Wishart distribution (Appendix C.14) is represented as follows:

$$\mathcal{W}(\mathbf{Y}|\mathbf{R}^0,\phi) = C_{\mathcal{W}}(\mathbf{R}^0,\phi)|\mathbf{Y}|^{\frac{\phi-D-1}{2}}\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{R}^0\mathbf{Y}\right]\right). \tag{4.58}$$

Then, $Q^{\text{MAP}}(\boldsymbol{\mu}',\mathbf{R}'|\boldsymbol{\mu},\mathbf{R})$ is represented by using Eqs. (4.42) with the normalization constant, (4.57), and (4.58), as follows:

$$\begin{aligned} Q^{\text{MAP}}(\boldsymbol{\mu}',\mathbf{R}'|\boldsymbol{\mu},\mathbf{R}) = &\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\gamma_t(j,k)}{2}\left(\log|\mathbf{R}'_{jk}| - (\mathbf{o}_t-\boldsymbol{\mu}'_{jk})^{\mathsf{T}}\mathbf{R}'_{jk}(\mathbf{o}_t-\boldsymbol{\mu}'_{jk})\right) \\ &+ \sum_{j=1}^{J}\sum_{k=1}^{K}\log\left(\mathcal{N}(\boldsymbol{\mu}'_{jk}|\boldsymbol{\mu}^0_{jk},(\phi^{\boldsymbol{\mu}}_{jk}\mathbf{R}'_{jk})^{-1})\mathcal{W}(\mathbf{R}'_{jk}|\mathbf{R}^0_{jk},\phi^{\mathbf{R}}_{jk})\right) \\ &- \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\gamma_t(j,k)D}{2}\log(2\pi) \end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\gamma_t(j,k)}{2}\left(\log|\mathbf{R}'_{jk}| - (\mathbf{o}_t - \boldsymbol{\mu}'_{jk})^{\mathsf{T}}\mathbf{R}'_{jk}(\mathbf{o}_t - \boldsymbol{\mu}'_{jk})\right) \\
&\quad + \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}\Bigg(\log\left|\mathbf{R}'_{jk}\right| - (\boldsymbol{\mu}'_{jk} - \boldsymbol{\mu}^0_{jk})^{\mathsf{T}}\phi^{\boldsymbol{\mu}}_{jk}\mathbf{R}'_{jk}(\boldsymbol{\mu}'_{jk} - \boldsymbol{\mu}^0_{jk}) \\
&\qquad\qquad + (\phi^{\mathbf{R}}_{jk} - D - 1)\log|\mathbf{R}'_{jk}| - \mathrm{tr}\left[\mathbf{R}^0_{jk}\mathbf{R}'_{jk}\right]\Bigg) \\
&\quad + \sum_{j=1}^{J}\sum_{k=1}^{K}\Bigg(-\sum_{t=1}^{T}\frac{\gamma_t(j,k)D}{2}\log(2\pi) - \frac{D}{2}\log(2\pi) + \frac{D}{2}\log\phi^{\boldsymbol{\mu}}_{jk} \\
&\qquad\qquad + \log C_{\mathcal{W}}(\mathbf{R}^0_{jk}, \phi^{\mathbf{R}}_{jk})\Bigg),
\end{aligned}
\tag{4.59}
$$

where the final line includes the terms that do not depend on $\boldsymbol{\mu}$ and $\mathbf{R}$. Then we rearrange Eq. (4.59) so that we can write it in a probabilistic form. First, we omit j, k, and $'$ in Eq. (4.59) for simplicity, and consider the following function:

$$
\begin{aligned}
g(\boldsymbol{\mu}, \mathbf{R}) &\triangleq \frac{1}{2}\Bigg(\sum_{t=1}^{T}\gamma_t\left(\log|\mathbf{R}| - (\mathbf{o}_t - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{R}(\mathbf{o}_t - \boldsymbol{\mu})\right) \\
&\qquad + \log|\mathbf{R}| - (\boldsymbol{\mu} - \boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\boldsymbol{\mu}}\mathbf{R}(\boldsymbol{\mu} - \boldsymbol{\mu}^0) \\
&\qquad + (\phi^{\mathbf{R}} - D - 1)\log|\mathbf{R}| - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\Bigg) \\
&= \frac{1}{2}\left(\sum_{t=1}^{T}\gamma_t\log|\mathbf{R}| + \log|\mathbf{R}| + (\phi^{\mathbf{R}} - D - 1)\log|\mathbf{R}| - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\right) \\
&\quad \underbrace{-\frac{1}{2}\left((\mathbf{o}_t - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{R}(\mathbf{o}_t - \boldsymbol{\mu}) - (\boldsymbol{\mu} - \boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\boldsymbol{\mu}}\mathbf{R}(\boldsymbol{\mu} - \boldsymbol{\mu}^0)\right)}_{\triangleq f(\boldsymbol{\mu},\mathbf{R})}.
\end{aligned}
\tag{4.60}
$$

Then, we focus on $f(\boldsymbol{\mu}, \mathbf{R})$ that has the terms in $g(\boldsymbol{\mu}, \mathbf{R})$ that depend on $\boldsymbol{\mu}$. $f(\boldsymbol{\mu}, \mathbf{R})$ is re-written as follows:

$$
\begin{aligned}
-2f(\boldsymbol{\mu}, \mathbf{R}) &= \boldsymbol{\mu}^{\mathsf{T}}\left(\left(\sum_t \gamma_t + \phi^{\boldsymbol{\mu}}\right)\mathbf{R}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathsf{T}}\left(\mathbf{R}\sum_t\gamma_t\mathbf{o}_t + \phi^{\boldsymbol{\mu}}\mathbf{R}\boldsymbol{\mu}^0\right) \\
&\quad + \sum_t \gamma_t\mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t + (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\boldsymbol{\mu}}\mathbf{R}\boldsymbol{\mu}^0.
\end{aligned}
\tag{4.61}
$$

Since this is a quadratic form of $\boldsymbol{\mu}$, it can be represented by a Gaussian distribution by arranging Eq. (4.61) into the complete square form. Although it is complicated to deal with the complete square form for vectors, we can use the following complete square rules found in Eqs. (B.16) and (B.17):

$$
\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - 2\mathbf{x}^{\mathsf{T}}\mathbf{b} + c = (\mathbf{x} - \mathbf{u})^{\mathsf{T}}\,\mathbf{A}\,(\mathbf{x} - \mathbf{u}) + v,
\tag{4.62}
$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

where

$$\begin{aligned} \mathbf{u} &\triangleq \mathbf{A}^{-1}\mathbf{b} \\ v &\triangleq c - \mathbf{b}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{b}. \end{aligned} \tag{4.63}$$

Therefore, by using this rule (e.g., $\mathbf{x} \to \boldsymbol{\mu}$, $\mathbf{A} \to \left(\sum_t \gamma_t + \phi^{\mu}\right)\mathbf{R}$, $\mathbf{b} \to \mathbf{R}\sum_t \gamma_t \mathbf{o}_t + \phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0$, and $c \to \sum_t \gamma_t \mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t + (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0$), Eq. (4.61) is rewritten with the complete square form as follows:

$$\begin{aligned} -2f(\boldsymbol{\mu},\mathbf{R}) &= \boldsymbol{\mu}^{\mathsf{T}}\left(\left(\sum_t \gamma_t + \phi^{\mu}\right)\mathbf{R}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathsf{T}}\left(\mathbf{R}\sum_t \gamma_t \mathbf{o}_t + \phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0\right) \\ &\quad + \sum_t \gamma_t \mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t + (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0 \\ &= (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^{\mathsf{T}}\left(\hat{\phi}^{\mu}\mathbf{R}\right)(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) + v(\mathbf{R}), \end{aligned} \tag{4.64}$$

where $\hat{\phi}^{\mu}$, $\hat{\boldsymbol{\mu}}$, and $v(\mathbf{R})$ are defined as follows:

$$\begin{aligned} \hat{\phi}^{\mu} &\triangleq \phi^{\mu} + \sum_t \gamma_t, \\ \hat{\boldsymbol{\mu}} &\triangleq \left(\left(\sum_t \gamma_t + \phi^{\mu}\right)\mathbf{R}\right)^{-1}\left(\mathbf{R}\sum_t \gamma_t \mathbf{o}_t + \phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0\right) \\ &= \frac{\phi^{\mu}\boldsymbol{\mu}^0 + \sum_t \gamma_t \mathbf{o}_t}{\phi^{\mu} + \sum_t \gamma_t}, \\ v(\mathbf{R}) &\triangleq \sum_t \gamma_t \mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t + (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0 - \left(\mathbf{R}\sum_t \gamma_t \mathbf{o}_t + \phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0\right)^{\mathsf{T}}\hat{\boldsymbol{\mu}} \\ &= \sum_t \gamma_t \mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t + (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\mu}\mathbf{R}\boldsymbol{\mu}^0 - \hat{\boldsymbol{\mu}}^{\mathsf{T}}\hat{\phi}^{\mu}\mathbf{R}\hat{\boldsymbol{\mu}}. \end{aligned} \tag{4.65}$$

Note that $\hat{\phi}^{\mu}$ and $\hat{\boldsymbol{\mu}}$ correspond to the mean and covariance hyperparameters of the conjugate Gaussian distribution of $\boldsymbol{\mu}$. Thus, $f(\boldsymbol{\mu},\mathbf{R})$ is rewritten with the definition of the Gaussian distribution in Eq. (4.57) as follows:

$$\begin{aligned} f(\boldsymbol{\mu},\mathbf{R}) &= -\frac{1}{2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^{\mathsf{T}}\left(\hat{\phi}^{\mu}\mathbf{R}\right)(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) - \frac{1}{2}v(\mathbf{R}) \\ &= \log\mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}},(\hat{\phi}^{\mu}\mathbf{R})^{-1}) - \log C_{\mathcal{N}}(\mathbf{R}^{-1}) - \frac{1}{2}v(\mathbf{R}) \\ &= \log\mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}},(\hat{\phi}^{\mu}\mathbf{R})^{-1}) + \frac{D}{2}\log(2\pi) - \frac{D}{2}\log\hat{\phi}^{\mu} - \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}v(\mathbf{R}), \end{aligned} \tag{4.66}$$

where $v(\mathbf{R})$ is used to obtain the analytic form of $\mathbf{R}$ with the rest of the $\mathbf{R}$-dependent terms in Eq. (4.60). That is, the auxiliary function (Eq. (4.60)) with Eq. (4.66) is represented as follows:

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$\begin{aligned} g(\boldsymbol{\mu}, \mathbf{R}) = \log \mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, (\hat{\phi}^{\boldsymbol{\mu}}\mathbf{R})^{-1}) + \frac{D}{2}\log(2\pi) - \frac{D}{2}\log\hat{\phi}^{\boldsymbol{\mu}} \\ \times \underbrace{\frac{1}{2}\left(-v(\mathbf{R}) + \left(\sum_t \gamma_t + \phi^{\mathbf{R}} - D - 1\right)\log|\mathbf{R}| - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\right)}_{\triangleq h(\mathbf{R})}. \end{aligned} \tag{4.67}$$

Now we focus on $h(\mathbf{R})$ that has the terms in the second line in Eq. (4.67). By using the definition of $v(\mathbf{R})$ in Eq. (4.65), we can rewrite $h(\mathbf{R})$ as follows:

$$\begin{aligned} h(\mathbf{R}) &= \frac{1}{2}\left(-v(\mathbf{R}) + \left(\sum_t \gamma_t + \phi^{\mathbf{R}} - D - 1\right)\log|\mathbf{R}| - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\right) \\ &= \frac{1}{2}\left(-\sum_t \gamma_t \mathbf{o}_t^{\mathsf{T}}\mathbf{R}\mathbf{o}_t - (\boldsymbol{\mu}^0)^{\mathsf{T}}\phi^{\boldsymbol{\mu}}\mathbf{R}\boldsymbol{\mu}^0 + \hat{\boldsymbol{\mu}}^{\mathsf{T}}\hat{\phi}^{\boldsymbol{\mu}}\mathbf{R}\hat{\boldsymbol{\mu}} - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\right) \\ &\quad + \frac{1}{2}\left(\sum_t \gamma_t + \phi^{\mathbf{R}} - D - 1\right)\log|\mathbf{R}|. \end{aligned} \tag{4.68}$$

Since Eq. (4.68) includes the trace operation, it is difficult to re-arrange this equation. Therefore, by using the trace rule of $a = \mathrm{tr}[a]$ (Eq. (B.1)), we represent all terms except for the $\log|\mathbf{R}|$ term in Eq. (4.68) as follows:

$$\begin{aligned} h(\mathbf{R}) = \frac{1}{2}\Bigg(&-\mathrm{tr}\left[\sum_t \gamma_t \mathbf{o}_t\mathbf{o}_t^{\mathsf{T}}\mathbf{R}\right] - \mathrm{tr}\left[\phi^{\boldsymbol{\mu}}\boldsymbol{\mu}^0(\boldsymbol{\mu}^0)^{\mathsf{T}}\mathbf{R}\right] \\ &+ \mathrm{tr}\left[\hat{\phi}^{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}}\mathbf{R}\right] - \mathrm{tr}\left[\mathbf{R}^0\mathbf{R}\right]\Bigg) \\ &+ \frac{1}{2}\left(\sum_t \gamma_t + \phi^{\mathbf{R}} - D - 1\right)\log|\mathbf{R}|. \end{aligned} \tag{4.69}$$

In addition, by using the trace rules of $\mathrm{tr}[\mathbf{ABC}] = \mathrm{tr}[\mathbf{BCA}]$ and $\mathrm{tr}[\mathbf{A}+\mathbf{B}] = \mathrm{tr}[\mathbf{A}] + \mathrm{tr}[\mathbf{B}]$ (Eqs. (B.2) and (B.3)), Eq. (4.69) is finally represented by comparing $h(\mathbf{R})$ with the definition of the Wishart distribution (Eq. (4.58)), as follows:

$$\begin{aligned} h(\mathbf{R}) &= -\frac{1}{2}\mathrm{tr}\Bigg[\underbrace{\left(\sum_t \gamma_t \mathbf{o}_t\mathbf{o}_t^{\mathsf{T}} + \phi^{\boldsymbol{\mu}}\boldsymbol{\mu}^0(\boldsymbol{\mu}^0)^{\mathsf{T}} - \hat{\phi}^{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^{\mathsf{T}} + \mathbf{R}^0\right)}_{\triangleq\hat{\mathbf{R}}}\mathbf{R}\Bigg] \\ &\quad + \frac{1}{2}\Bigg(\underbrace{\left(\sum_t \gamma_t + \phi^{\mathbf{R}}\right)}_{\triangleq\hat{\phi}^{\mathbf{R}}} - D - 1\Bigg)\log|\mathbf{R}| \\ &= \log\mathcal{W}(\mathbf{R}|\hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}) - \log C_{\mathcal{W}}(\hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}). \end{aligned} \tag{4.70}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Thus, Eq. (4.70) can be represented as the Wishart distribution with the following hyperparameters:

$$\begin{aligned}\hat{\phi}^{\mathbf{R}} &\triangleq \phi^{\mathbf{R}} + \sum_t \gamma_t, \\ \hat{\mathbf{R}} &\triangleq \sum_t \gamma_t \mathbf{o}_t \mathbf{o}_t^{\mathsf{T}} + \phi^{\boldsymbol{\mu}} \boldsymbol{\mu}^0 (\boldsymbol{\mu}^0)^{\mathsf{T}} - \hat{\phi}^{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^{\mathsf{T}} + \mathbf{R}^0.\end{aligned} \tag{4.71}$$

Note that these hyperparameters are almost equivalent to those of the conjugate distribution analysis discussed in Eq. (2.112). The only difference is that the statistics in Eq. (4.71) are computed from the expectation value of the posterior distribution of a latent variable γ_t.

Here, $g(\boldsymbol{\mu}, \mathbf{R})$ in the original Q function, Eq. (4.59), is represented by Eqs. (4.67) and (4.70) as follows:

$$\begin{aligned}g(\boldsymbol{\mu}, \mathbf{R}) = \log \mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, (\hat{\phi}^{\boldsymbol{\mu}}\mathbf{R})^{-1}) + \frac{D}{2}\log(2\pi) - \frac{D}{2}\log \hat{\phi}^{\boldsymbol{\mu}} \\ + \log \mathcal{W}(\mathbf{R}|\hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}) - \log C_{\mathcal{W}}(\hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}).\end{aligned} \tag{4.72}$$

Thus, we have found that $Q^{\text{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R})$ can be represented with the same distribution form as the prior distributions, which is represented by Gaussian and Wishart distributions as follows:

$$\begin{aligned}&Q^{\text{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} \log\left(\mathcal{N}(\boldsymbol{\mu}'_{jk}|\hat{\boldsymbol{\mu}}_{jk}, (\hat{\phi}^{\boldsymbol{\mu}}_{jk}\mathbf{R}'_{jk})^{-1})\mathcal{W}(\mathbf{R}'_{jk}|\hat{\mathbf{R}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk})\right) \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K}\left(-\sum_{t=1}^{T}\frac{\gamma_t(j,k)D}{2}\log(2\pi) + \frac{D}{2}\log\frac{\phi^{\boldsymbol{\mu}}_{jk}}{\hat{\phi}^{\boldsymbol{\mu}}_{jk}} + \log\frac{C_{\mathcal{W}}(\mathbf{R}^0_{jk}, \phi^{\mathbf{R}}_{jk})}{C_{\mathcal{W}}(\hat{\mathbf{R}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk})}\right) \\ &\propto \sum_{j=1}^{J}\sum_{k=1}^{K} \log\left(\mathcal{N}(\boldsymbol{\mu}'_{jk}|\hat{\boldsymbol{\mu}}_{jk}, (\hat{\phi}^{\boldsymbol{\mu}}_{jk}\mathbf{R}'_{jk})^{-1})\mathcal{W}(\mathbf{R}'_{jk}|\hat{\mathbf{R}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk})\right),\end{aligned} \tag{4.73}$$

where we recover the indexes j, k, and $'$. By using the definition of the Gaussian–Wishart distribution in Appendix C.15, $Q^{\text{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R})$ can also be represented as:

$$Q^{\text{MAP}}(\boldsymbol{\mu}', \mathbf{R}'|\boldsymbol{\mu}, \mathbf{R}) \propto \sum_{j=1}^{J}\sum_{k=1}^{K} \log\left(\mathcal{NW}(\boldsymbol{\mu}'_{jk}, \mathbf{R}'_{jk}|\hat{\boldsymbol{\mu}}_{jk}, \hat{\phi}^{\boldsymbol{\mu}}_{jk}, \hat{\mathbf{R}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk})\right). \tag{4.74}$$

By summarizing the result of Eqs. (4.65) and (4.71), the hyperparameters of these distributions are defined as:

$$\begin{aligned}\hat{\phi}^{\boldsymbol{\mu}}_{jk} &\triangleq \phi^{\boldsymbol{\mu}}_{jk} + \sum_t \gamma_t(j,k), \\ \hat{\boldsymbol{\mu}}_{jk} &\triangleq \frac{\phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^0_{jk} + \sum_t \gamma_t(j,k)\mathbf{o}_t}{\phi^{\boldsymbol{\mu}}_{jk} + \sum_t \gamma_t(j,k)},\end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$\begin{aligned}
\hat{\phi}^{\mathbf{R}}_{jk} &\triangleq \phi^{\mathbf{R}}_{jk} + \sum_t \gamma_t(j,k), \\
\hat{\mathbf{R}}_{jk} &\triangleq \sum_t \gamma_t(j,k)\mathbf{o}_t\mathbf{o}_t^\intercal + \phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^0_{jk}(\boldsymbol{\mu}^0_{jk})^\intercal - \hat{\phi}^{\boldsymbol{\mu}}_{jk}\hat{\boldsymbol{\mu}}_{jk}\hat{\boldsymbol{\mu}}^\intercal_{jk} + \mathbf{R}^0_{jk}.
\end{aligned} \tag{4.75}$$

The MAP estimates of these values are obtained by considering the modes of the Gaussian–Wishart distribution in Appendix C.15. The modes of $\boldsymbol{\mu}'_{jk}$ and $\boldsymbol{\Sigma}'_{jk}$ are represented as:

$$\begin{aligned}
\boldsymbol{\mu}^{\text{MAP}}_{jk} &= \hat{\boldsymbol{\mu}}_{jk}, \\
\boldsymbol{\Sigma}^{\text{MAP}}_{jk} &= \left(\mathbf{R}^{\text{MAP}}_{jk}\right)^{-1} = (\hat{\phi}^{\mathbf{R}}_{jk} - D - 1)^{-1}\hat{\mathbf{R}}_{jk}.
\end{aligned} \tag{4.76}$$

Note that $\boldsymbol{\Sigma}^{\text{MAP}}_{jk}$ cannot be obtained when $\hat{\phi}^{\mathbf{R}}_{jk} - D - 1 \leq 0$. Thus, we can analytically obtain the M-step solutions of CDHMM parameters (i.e., initial weight, state transition, mixture weight, mean vector and covariance matrix) in the MAP sense, thanks to the conjugate prior distributions. We summarize the solutions below. The hyperparameters of the posterior distributions are represented with Gaussian sufficient statistics and the prior hyperparameters as:

$$\begin{cases}
\hat{\phi}^{\pi}_j \triangleq \phi^{\pi}_j + \xi_1(j), \\
\hat{\phi}^{a}_{ij} \triangleq \phi^{a}_{ij} + \sum_{t=1}^{T-1} \xi_t(i,j), \\
\hat{\phi}^{\omega}_{jk} \triangleq \phi^{\omega}_{jk} + \sum_{t=1}^{T} \gamma_t(j,k), \\
\hat{\phi}^{\boldsymbol{\mu}}_{jk} \triangleq \phi^{\boldsymbol{\mu}}_{jk} + \sum_{t=1}^{T} \gamma_t(j,k), \\
\hat{\boldsymbol{\mu}}_{jk} \triangleq \dfrac{\phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^0_{jk} + \sum_{t=1}^T \gamma_t(j,k)\mathbf{o}_t}{\phi^{\boldsymbol{\mu}}_{jk} + \sum_{t=1}^T \gamma_t(j,k)}, \\
\hat{\phi}^{\mathbf{R}}_{jk} \triangleq \phi^{\mathbf{R}}_{jk} + \sum_{t=1}^{T} \gamma_t(j,k), \\
\hat{\mathbf{R}}_{jk} \triangleq \sum_{t=1}^{T} \gamma_t(j,k)\mathbf{o}_t\mathbf{o}_t^\intercal + \phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^0_{jk}(\boldsymbol{\mu}^0_{jk})^\intercal - \hat{\phi}^{\boldsymbol{\mu}}_{jk}\hat{\boldsymbol{\mu}}_{jk}\hat{\boldsymbol{\mu}}^\intercal_{jk} + \mathbf{R}^0_{jk}.
\end{cases} \tag{4.77}$$

A set of hyperparameters $\hat{\Psi}$ is defined as follows:

$$\hat{\Psi} \triangleq \{\hat{\phi}^{\pi}_j, \hat{\phi}^{a}_{ij}, \hat{\phi}^{\omega}_{jk}, \hat{\phi}^{\boldsymbol{\mu}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk}, \hat{\boldsymbol{\mu}}_{jk}, \hat{\mathbf{R}}_{jk} | i = 1, \cdots, J, j = 1, \cdots, J, k = 1, \cdots, K\}. \tag{4.78}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Then, the MAP solutions of HMM parameters are represented with the posterior hyperparameters as follows:

$$\begin{cases} \pi_j^{\mathrm{MAP}} = \dfrac{\hat{\phi}_j^{\pi} - 1}{\sum_{j'=1}^{J}(\hat{\phi}_{j'}^{\pi} - 1)}, \\ a_{ij}^{\mathrm{MAP}} = \dfrac{\hat{\phi}_{ij}^{a} - 1}{\sum_{j'=1}^{J}(\hat{\phi}_{ij'}^{a} - 1)}, \\ \omega_{jk}^{\mathrm{MAP}} = \dfrac{\hat{\phi}_{jk}^{\omega} - 1}{\sum_{k'=1}^{K}(\hat{\phi}_{jk'}^{\omega} - 1)}, \\ \boldsymbol{\mu}_{jk}^{\mathrm{MAP}} = \hat{\boldsymbol{\mu}}_{jk}, \\ \boldsymbol{\Sigma}_{jk}^{\mathrm{MAP}} = (\hat{\phi}_{jk}^{\mathbf{R}} - D - 1)^{-1}\hat{\mathbf{R}}_{jk}. \end{cases} \tag{4.79}$$

Compared with the ML M-step of CDHMM parameters in Eq. (3.151), these solutions are more complicated, and actually incur more computational cost than that of the ML solution. However, the computational cost of the M-step is much smaller than that of the E-step, and the additional computational cost of the MAP estimate can be disregarded in practical use.

Mean vector and diagonal covariance matrix

In practise, we often use the diagonal covariance matrix for a multivariate Gaussian distribution for HMMs, as we discussed in Section 3.2.3. This section provides the MAP solution for the diagonal covariance case (Gauvain & Lee 1991). Since the one-dimensional solution of the full-covariance Gaussian posterior distribution in the previous discussion corresponds to that of the diagonal covariance Gaussian posterior distribution of a diagonal element, we can obtain the hyperparameters of the posterior distribution of the CDHMM parameters by using $D \to 1$ for each diagonal component. We also summarize the solution of the MAP estimates of HMM parameters of the diagonal covariance Gaussian case below. The hyperparameters of the posterior distributions are represented with Gaussian sufficient statistics and the prior hyperparameters as:

$$\begin{cases} \hat{\phi}_j^{\pi} & \triangleq \phi_j^{\pi} + \xi_1(j), \\ \hat{\phi}_{ij}^{a} & \triangleq \phi_{ij}^{a} + \sum_{t=1}^{T-1} \xi_t(i,j), \\ \hat{\phi}_{jk}^{\omega} & \triangleq \phi_{jk}^{\omega} + \sum_{t=1}^{T} \gamma_t(j,k), \\ \hat{\phi}_{jk}^{\boldsymbol{\mu}} & \triangleq \phi_{jk}^{\boldsymbol{\mu}} + \sum_{t=1}^{T} \gamma_t(j,k), \\ \hat{\boldsymbol{\mu}}_{jk} & \triangleq \frac{\phi_{jk}^{\boldsymbol{\mu}}\boldsymbol{\mu}_{jk}^{0} + \sum_{t=1}^{T}\gamma_t(j,k)\mathbf{o}_t}{\phi_{jk}^{\boldsymbol{\mu}} + \sum_{t=1}^{T}\gamma_t(j,k)}, \\ \hat{\phi}_{jk}^{\mathbf{R}} & \triangleq \phi_{jk}^{\mathbf{R}} + \sum_{t=1}^{T} \gamma_t(j,k), \\ \hat{r}_{jkd} & \triangleq \sum_{t=1}^{T} \gamma_t(j,k) o_{td}^2 + \phi_{jk}^{\boldsymbol{\mu}}(\mu_{jkd}^{0})^2 - \hat{\phi}_{jk}^{\boldsymbol{\mu}}(\hat{\mu}_{jkd})^2 + r_{jkd}^{0}. \end{cases} \tag{4.80}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

In this case, a set of hyperparameters $\hat{\Psi}$ is defined as follows:

$$\hat{\Psi} \triangleq \{\hat{\phi}^{\pi}_{j}, \hat{\phi}^{a}_{ij}, \hat{\phi}^{\omega}_{jk}, \hat{\phi}^{\boldsymbol{\mu}}_{jk}, \hat{\phi}^{\mathbf{R}}_{jk}, \hat{\boldsymbol{\mu}}_{jk}, \hat{\mathbf{r}}_{jk} | i = 1, \cdots, J, j = 1, \cdots, J, k = 1, \cdots, K\}, \quad (4.81)$$

where

$$\hat{\mathbf{r}} \triangleq [\hat{r}_1, \cdots, \hat{r}_D]^{\mathsf{T}}. \quad (4.82)$$

Then, the MAP solutions of HMM parameters can be represented with the posterior hyperparameters as follows:

$$\begin{cases} \pi^{\text{MAP}}_{j} = \dfrac{\hat{\phi}^{\pi}_{j} - 1}{\sum_{j'=1}^{J}(\hat{\phi}^{\pi}_{j'} - 1)}, \\ a^{\text{MAP}}_{ij} = \dfrac{\hat{\phi}^{a}_{ij} - 1}{\sum_{j'=1}^{J}(\hat{\phi}^{a}_{ij'} - 1)}, \\ \omega^{\text{MAP}}_{jk} = \dfrac{\hat{\phi}^{\omega}_{jk} - 1}{\sum_{k'=1}^{K}(\hat{\phi}^{\omega}_{jk'} - 1)}, \\ \boldsymbol{\mu}^{\text{MAP}}_{jk} = \hat{\boldsymbol{\mu}}_{jk}, \\ \Sigma^{\text{MAP}}_{jkd} = \dfrac{\hat{r}_{jkd}}{\hat{\phi}^{\mathbf{R}}_{jk} - 2}. \end{cases} \quad (4.83)$$

Thus, we obtain the MAP estimates of HMM parameters in both diagonal and full covariance matrix cases.

4.3.6 Sufficient statistics

As we discussed in Eq. (4.43), the posterior probabilities of the state transition $\xi_t(i,j)$, and mixture occupation $\gamma_t(j,k)$ can be computed by plugging the MAP estimates Θ^{MAP} obtained by using Eqs. (4.79) or (4.83) into the variables in Sections 3.3 and 3.4.2. Note that the analytical results here are exactly the same as those in the ML–EM algorithm (except for the MAP estimates Θ^{MAP}): we list these for convenience.

First, the MAP forward variable $\alpha_t(j)$ is computed by using the following equation:

- Initialization

$$\begin{aligned} \alpha_1(j) &= p(\mathbf{o}_1, s_1 = j|\Theta^{\text{MAP}}) \\ &= \pi^{\text{MAP}}_{j} b^{\text{MAP}}_{j}(\mathbf{o}_1), \qquad 1 \le j \le J. \end{aligned} \quad (4.84)$$

- Induction

$$\begin{aligned} \alpha_t(j) &= p(\mathbf{o}_1, \cdots, \mathbf{o}_t, s_t = j|\Theta^{\text{MAP}}) \\ &= \left(\sum_{i=1}^{J} \alpha_{t-1}(i) a^{\text{MAP}}_{ij}\right) b^{\text{MAP}}_{j}(\mathbf{o}_t), \qquad \begin{matrix} 2 \le t \le T \\ 1 \le j \le J. \end{matrix} \end{aligned} \quad (4.85)$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

- Termination

$$p(\mathbf{O}|\Theta^{\text{MAP}}) = \sum_{j=1}^{J} \alpha_T(j). \tag{4.86}$$

Here, $b_j^{\text{MAP}}(\mathbf{o}_t)$ is a GMM emission probability distribution with the MAP estimate parameters defined as:

$$b_j^{\text{MAP}}(\mathbf{o}_t) \triangleq \sum_{k=1}^{K} \omega_{jk}^{\text{MAP}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk}^{\text{MAP}}, \boldsymbol{\Sigma}_{jk}^{\text{MAP}}). \tag{4.87}$$

The MAP backward variable $\beta_t(j)$ is computed by using the following equations:

- Initialization

$$\beta_T(j) = 1, \quad 1 \leq j \leq J. \tag{4.88}$$

- Induction

$$\begin{aligned}\beta_t(i) &= p(\mathbf{o}_{t+1}, \cdots, \mathbf{o}_T|s_t = i, \Theta^{\text{MAP}}) \\ &= \sum_{j=1}^{J} a_{ij}^{\text{MAP}} b_j^{\text{MAP}}(\mathbf{o}_{t+1})\beta_{t+1}(j), \\ & t = T-1, T-2, \cdots, 1, \quad 1 \leq i \leq J.\end{aligned} \tag{4.89}$$

- Termination

$$\begin{aligned}\beta_0 &\triangleq p(\mathbf{O}|\Theta^{\text{MAP}}) \\ &= \sum_{j=1}^{J} \pi_j^{\text{MAP}} b_j^{\text{MAP}}(\mathbf{o}_1)\beta_1(j).\end{aligned} \tag{4.90}$$

Therefore, based on the MAP forward and backward variables, we can compute the posterior probabilities as follows:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}^{\text{MAP}}\left(\sum_{k=1}^{K} \omega_{jk}^{\text{MAP}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk}^{\text{MAP}}, \boldsymbol{\Sigma}_{jk}^{\text{MAP}})\right)\beta_{t+1}(j)}{\sum_{i'=1}^{J}\sum_{j'=1}^{J} \alpha_t(i')a_{i'j'}^{\text{MAP}}\left(\sum_{k=1}^{K} \omega_{j'k}^{\text{MAP}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{j'k}^{\text{MAP}}, \boldsymbol{\Sigma}_{j'k}^{\text{MAP}})\right)\beta_{t+1}(j')}, \tag{4.91}$$

$$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j'=1}^{J} \alpha_t(j')\beta_t(j')} \cdot \frac{\omega_{jk}^{\text{MAP}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk}^{\text{MAP}}, \boldsymbol{\Sigma}_{jk}^{\text{MAP}})}{\sum_{k'=1}^{K} \omega_{jk'}^{\text{MAP}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk'}^{\text{MAP}}, \boldsymbol{\Sigma}_{jk'}^{\text{MAP}})}. \tag{4.92}$$

Once we have computed the posterior probabilities, we can compute the following sufficient statistics:

$$\begin{cases} \sum_{t=1}^{T-1} \xi_t(i,j) & \triangleq \xi_{ij}, \\ \sum_{t=1}^{T} \gamma_t(j,k) & \triangleq \gamma_{jk}, \\ \sum_{t=1}^{T} \gamma_t(j,k)\mathbf{o}_t & \triangleq \boldsymbol{\gamma}_{jk}^{(1)}, \\ \sum_{t=1}^{T} \gamma_t(j,k)\mathbf{o}_t\mathbf{o}_t^{\mathsf{T}} & \triangleq \boldsymbol{\Gamma}_{jk}^{(2)}, \\ \sum_{t=1}^{T} \gamma_t(j,k)o_{td}^2 & \triangleq \gamma_{jkd}^{(2)}. \end{cases} \tag{4.93}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

The superscripts $^{(1)}$ and $^{(2)}$ denote the 1st and 2nd order statistics.

This total EM algorithm for iteratively estimating the HMM parameters based on the MAP estimate is set out in Algorithm 8. Compared with the ML Baum–Welch algorithm, Algorithm 4, it requires the hyperparameters Ψ of the CDHMMs. Section 7.3 also introduces a variant of the Baum–Welch algorithm based on variational Bayes. Note again that the MAP E-step (computing forward variables, occupation probabilities, and accumulation) is exactly the same as that of the ML E-step, and retains the nice property of the parallelization and data scalability. In addition, since the E-step computation is dominant in the algorithm, the computational costs of the ML and MAP Baum–Welch algorithms are almost same.

Algorithm 8 MAP Baum–Welch algorithm

```
Require: $\Psi$ and $\Theta^{\mathrm{MAP}} \leftarrow \Theta^{\mathrm{init}}$
repeat
    Compute the forward variable $\alpha_t(j)$ from the forward algorithm
    Compute the backward variable $\beta_t(j)$ from the backward algorithm
    Compute the occupation probabilities $\gamma_1(j)$, $\gamma_t(j,k)$, and $\xi_t(i,j)$
    Accumulate the sufficient statistics $\xi(i,j)$, $\gamma(j,k)$, $\boldsymbol{\gamma}^{(1)}_{jk}$, and $\boldsymbol{\Gamma}^{(2)}_{jk}$ (or $\gamma^{(2)}_{jkd}$)
    Estimate the new hyperparameters $\hat{\Psi}$
    Estimate the new HMM parameters $(\Theta^{\mathrm{MAP}})'$
    Update the HMM parameters $\Theta^{\mathrm{MAP}} \leftarrow (\Theta^{\mathrm{MAP}})'$
until Convergence
```

4.3.7 Meaning of the MAP solution

This section discusses the meaning of the MAP solution obtained by Eqs. (4.79) and (4.83). We consider the two extreme case of the MAP solution, where the amount of data is small and large. That is, we consider the small data limit as $\xi_{ij}, \gamma_{jk} \to 0$. On the other hand, the large data limit corresponds to $\xi_{ij}, \gamma_{jk} \to \infty$.

- Mixture weight

  We first focus on the MAP estimate of the mixture weight ω, but the discussion can be applied to the state transition a.

  Large sample:

  The MAP estimate of the mixture weight is represented as follows:

$$\omega^{\mathrm{MAP}}_{jk} = \frac{\phi^{\omega}_{jk} + \gamma_{jk} - 1}{\sum_{k'=1}^{K}(\phi^{\omega}_{jk'} + \gamma_{jk'} - 1)} = \frac{\gamma_{jk}\left(\frac{\phi^{\omega}_{jk}-1}{\gamma_{jk}} + 1\right) - 1}{\sum_{k'=1}^{K}\gamma_{jk'}\left(\frac{\phi^{\omega}_{jk'}-1}{\gamma_{jk'}} + 1\right) - 1}. \tag{4.94}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Since $\frac{\phi^{\omega}_{jk}-1}{\gamma_{jk}} \to 0$ in the large sample case, the MAP estimate $\omega^{\mathrm{MAP}}_{jk}$ approaches the ML estimate $\omega^{\mathrm{ML}}_{jk}$ when γ_{jk} is sufficiently larger than $\phi^{\omega}_{jk} - 1$:

$$\omega^{\mathrm{MAP}}_{jk} \approx \frac{\gamma_{jk}}{\sum_{k'=1}^{K} \gamma_{jk'}} = \omega^{\mathrm{ML}}_{jk} \quad (\gamma_{jk} \gg \phi^{\omega}_{jk} - 1). \tag{4.95}$$

Small sample:

Similarly, the MAP estimate $\omega^{\mathrm{MAP}}_{jk}$ approaches the following value when γ_{jk} is sufficiently smaller than $\phi^{\omega}_{jk} - 1$:

$$\omega^{\mathrm{MAP}}_{jk} \approx \frac{\phi^{\omega}_{jk} - 1}{\sum_{k'=1}^{K}(\phi^{\omega}_{jk'} - 1)} \quad (\gamma_{jk} \ll \phi^{\omega}_{jk} - 1). \tag{4.96}$$

The weight is only computed from the prior hyperparameter ϕ^{ω}_{jk}. Thus, the mixture weight approaches the ML estimate when the amount of data is large, while it approaches the weight obtained only from the prior hyperparameters when the amount is small. Hyperparameter ϕ^{ω}_{jk} can be regarded as a scale.

- Mean

By using Eqs. (4.93) and (4.79), the MAP estimate of the mean vector can be rewritten as follows:

$$\begin{aligned}\boldsymbol{\mu}^{\mathrm{MAP}}_{jk} &= \frac{\phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^{0}_{jk} + \sum_t \gamma_t(j,k)\mathbf{o}_t}{\phi^{\boldsymbol{\mu}}_{jk} + \sum_t \gamma_t(j,k)} \\ &= \frac{\phi^{\boldsymbol{\mu}}_{jk}\boldsymbol{\mu}^{0}_{jk} + \boldsymbol{\gamma}^{(1)}_{jk}}{\phi^{\boldsymbol{\mu}}_{jk} + \gamma_{jk}} \\ &= \frac{\frac{\phi^{\boldsymbol{\mu}}_{jk}}{\gamma_{jk}}\boldsymbol{\mu}^{0}_{jk} + \boldsymbol{\mu}^{\mathrm{ML}}_{jk}}{\frac{\phi^{\boldsymbol{\mu}}_{jk}}{\gamma_{jk}} + 1}.\end{aligned} \tag{4.97}$$

This equation means that the MAP estimate $\boldsymbol{\mu}^{\mathrm{MAP}}_{jk}$ is linearly interpolated by the ML estimate $\boldsymbol{\mu}^{\mathrm{ML}}_{jk}$ and the hyperparameter $\boldsymbol{\mu}^{0}_{jk}$, as shown in Figure 4.1. $\frac{\phi^{\boldsymbol{\mu}}_{jk}}{\gamma_{jk}}$ is an interpolation ratio, and it has a specific meaning when the amount of data is sufficiently large ($\gamma_{jk} \gg \phi^{\omega}_{jk}$) or small ($\gamma_{jk} \ll \phi^{\omega}_{jk}$).

Large sample:

$$\boldsymbol{\mu}^{\mathrm{MAP}}_{jk} \approx \boldsymbol{\mu}^{\mathrm{ML}}_{jk}. \tag{4.98}$$

Similarly to the discussion of the mixture weight, the MAP estimate of the mean vector theoretically converges to the ML estimate.

Small sample:

$$\boldsymbol{\mu}^{\mathrm{MAP}}_{jk} \approx \boldsymbol{\mu}^{0}_{jk}. \tag{4.99}$$

This is a good property of the MAP estimate. Although the ML estimate with a small amount of data incorrectly estimates the mean vector, which degrades the

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

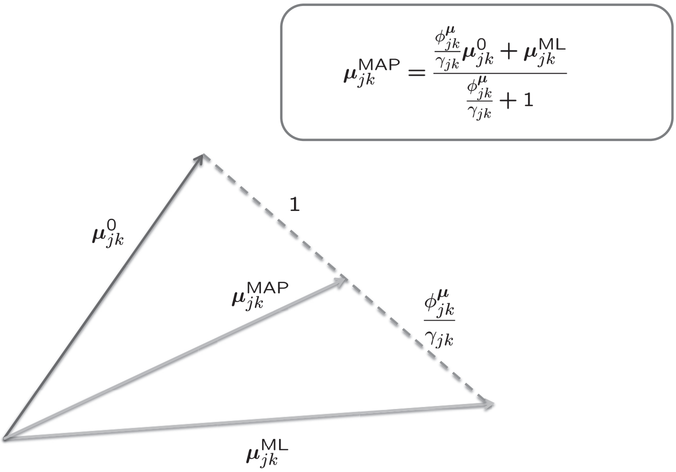
$\mu_{jk}^{\mathsf{MAP}} = \frac{\frac{\phi_{jk}^{\mu}}{\gamma_{jk}}\mu_{jk}^{0} + \mu_{jk}^{\mathsf{ML}}}{\frac{\phi_{jk}^{\mu}}{\gamma_{jk}} + 1}$
μ_{jk}^{0}
1
μ_{jk}^{MAP}
$\frac{\phi_{jk}^{\mu}}{\gamma_{jk}}$
μ_{jk}^{ML}

Figure 4.1 Geometric meaning of the MAP estimate of the Gaussian mean vector. It is represented as the linear interpolation of the prior mean vector $\boldsymbol{\mu}_{jk}^{0}$ and the ML estimate of the mean vector $\boldsymbol{\mu}_{jk}^{\mathrm{ML}}$. The interpolation ratio depends on the hyperparameter $\phi_{jk}^{\boldsymbol{\mu}}$ and the amount of data γ_{jk} assigned to the Gaussian.

performance drastically, the MAP estimate can smooth the incorrect estimation based on the hyperparameter $\boldsymbol{\mu}_{jk}^{0}$. In the practical situation, we also often have a zero count problem (i.e., $\gamma_{jk} = 0$), which makes the ML estimate singular due to the zero divide. However, the MAP solution of the mean vector avoids this problem and provides a reasonable estimate obtained from the hyperparameter $\boldsymbol{\mu}_{jk}^{0}$.

- Covariance matrix

  By using Eqs. (4.93) and (4.79), the MAP estimate of the covariance matrix can be rewritten as follows:

$$\begin{aligned}\boldsymbol{\Sigma}_{jk}^{\mathrm{MAP}} &= (\phi_{jk}^{\mathbf{R}} + \gamma_{jk} - D - 1)^{-1} \\ &\quad \times \left(\boldsymbol{\Gamma}_{jk}^{(2)} + \phi_{jk}^{\boldsymbol{\mu}}\boldsymbol{\mu}_{jk}^{0}(\boldsymbol{\mu}_{jk}^{0})^{\intercal} - \hat{\phi}_{jk}^{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_{jk}\hat{\boldsymbol{\mu}}_{jk}^{\intercal} + \mathbf{R}_{jk}^{0}\right). \qquad (4.100)\end{aligned}$$

  Large sample:

$$\begin{aligned}\boldsymbol{\Sigma}_{jk}^{\mathrm{MAP}} &\approx (\gamma_{jk})^{-1}\left(\boldsymbol{\Gamma}_{jk}^{(2)} - \gamma_{jk}\boldsymbol{\mu}_{jk}^{\mathrm{ML}}(\boldsymbol{\mu}_{jk}^{\mathrm{ML}})^{\intercal}\right) \\ &= \boldsymbol{\Sigma}_{jk}^{\mathrm{ML}}. \qquad (4.101)\end{aligned}$$

  The result is the same when we use the diagonal covariance.

  Small sample (full covariance):

$$\begin{aligned}\boldsymbol{\Sigma}_{jk}^{\mathrm{MAP}} &\approx (\phi_{jk}^{\mathbf{R}} - D - 1)^{-1}\left(\phi_{jk}^{\boldsymbol{\mu}}\boldsymbol{\mu}_{jk}^{0}(\boldsymbol{\mu}_{jk}^{0})^{\intercal} - \phi_{jk}^{\boldsymbol{\mu}}\boldsymbol{\mu}_{jk}^{0}(\boldsymbol{\mu}_{jk}^{0})^{\intercal} + \mathbf{R}_{jk}^{0}\right) \\ &= (\phi_{jk}^{\mathbf{R}} - D - 1)^{-1}\mathbf{R}_{jk}^{0}. \qquad (4.102)\end{aligned}$$

  Unlike the mean vector case, the covariance matrix of the small sample limit is represented by the two hyperparameters $\mathbf{R}_{jk}^{0}$ and $\phi_{jk}^{\mathbf{R}}$. To make the solution meaningful, we need to set $\phi_{jk}^{\mathbf{R}} > D + 1$ to avoid a negative or zero value of the variance.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

- *Small sample* (diagonal covariance):

  By using Eq. (4.83), we can also obtain the variance parameter for dimension d as follows:

$$\Sigma_{jkd}^{\text{MAP}} \approx \frac{r_{jkd}^{0}}{\phi_{jk}^{\mathbf{R}} - 2}. \tag{4.103}$$

  Similarly to the diagonal case, we need to set $\phi_{jk}^{\mathbf{R}} > 2$ to avoid a negative or zero value of the variance.

In summary, the MAP solutions can smooth the estimation values with the hyperparameters when the amount of data is small, and the solutions approach the ML estimates when the amount of data is large. A similar discussion has already been presented in Section 2.1.4, as a general property of the Bayesian approach.

The MAP estimation of the CDHMM parameters can be applied to general CDHMM training. However, since CDHMM is usually trained with a sufficient amount of training data, we do not have to use MAP estimation, and ML estimation is enough in most cases, which corresponds to the case of the large sample limitation in the above discussion. However, we often face the case when the amount of data is small at an adaptation scenario. The following section introduces one of the most successful applications of the MAP estimation for *speaker adaptation*.

4.4 Speaker adaptation

Speaker adaptation is one of the most important techniques in speech recognition, mainly to deal with speaker variations in speech (Lee & Huo 2000, Shinoda 2010). The speech features of a speaker are different from those of another speaker, which degrades the performance of speech recognition. A straightforward solution for this problem is to build a speaker-dependent acoustic model for a specific person. However, it is difficult to collect sufficient training data with labels.

Speaker adaptation aims to solve the problem by first building a speaker-independent (SI) acoustic model Θ^{SI} by using many speakers' data, and updates the model as a speaker-dependent (SD) acoustic model Θ^{SD} with a small amount of data of the target speaker, as shown in Figure 4.2. The speaker-independent acoustic model is usually made by the conventional maximum likelihood procedure, as we discussed in Chapter 3, or discriminative training. It is also obtained by using so-called speaker adaptive training (Anastasakos, McDonough, Schwartz *et al.* 1996) or cluster adaptive training (Gales, Center & Heights 2000), which normalizes the speaker characteristics (or some other characteristics (e.g., noises, speaking styles) obtained from clustering of speech utterances) by using a variant of the maximum likelihood linear regression (MLLR) technique, which is discussed in Section 3.5.1.

4.4.1 Speaker adaptation by a transformation of CDHMM

Once we have SI model parameters Θ^{SI}, the problem is how to estimate the SD model parameters Θ^{SD} without over-training. Basically, the number of SI model parameters

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

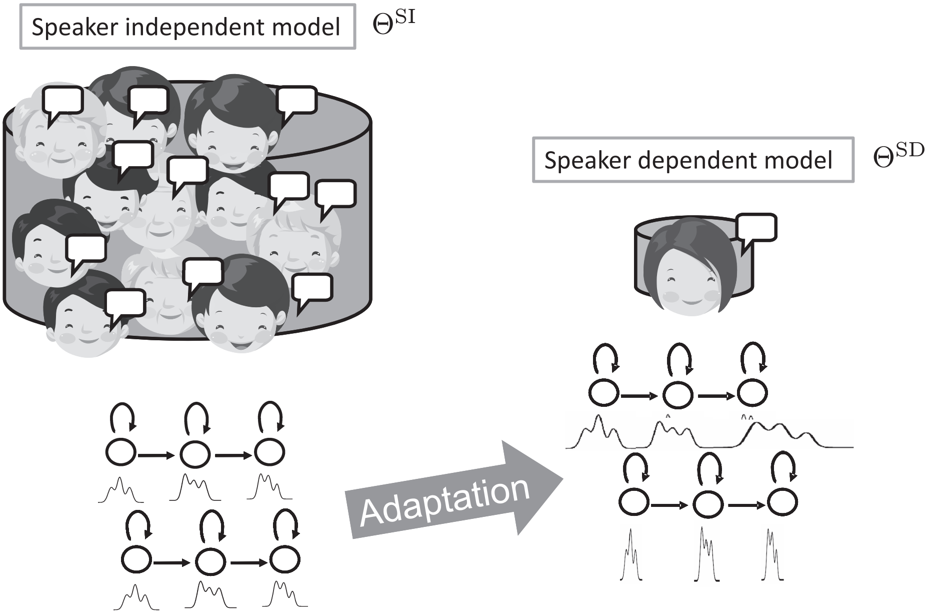

Figure 4.2 Speaker adaptation of HMM parameters. The initial HMMs are trained with many speakers, and then the HMMs are adapted to the target speaker's model with a small amount of adaptation data.

is very large. For example, in the famous speech recognition task using read speech of *Wall Street Journal* (*WSJ*) sentences, the number of CDHMM parameters amounts to several millions or more. On the other hand, the amount of speech data for the target speaker with text labels would be a few minutes at most, and the number of frames corresponds to the order of tens of thousands, and is even smaller than the number of standard CDHMM parameters. The following ML estimate with the EM algorithm, as we discussed in Section 3.4, causes serious over-training:

$$\Theta^{\mathrm{SD,\,ML}} = \arg\max_{\Theta^{\mathrm{SD}'}} Q^{\mathrm{ML}}(\Theta^{\mathrm{SD}'}|\Theta^{\mathrm{SD}}). \tag{4.104}$$

There are several approaches to overcoming the problem by using the maximum likelihood linear regression (MLLR) (Digalakis *et al.* 1995, Leggetter & Woodland 1995, Gales & Woodland 1996), as discussed in Section 3.5, eigenvoice approaches (Kuhn, Junqua, Ngyuen *et al.* 2000), and so on. These approaches set a parametric constraint of fewer CDHMM parameters, and estimate these parameters (Λ) instead of CDHMM parameters with ML indirectly, that is:

$$\Lambda^{\mathrm{ML}} = \arg\max_{\Lambda'} Q^{\mathrm{ML}}(\Lambda'|\Lambda; \Theta^{\mathrm{SI}}). \tag{4.105}$$

Detailed discussions of these adaptation techniques can be found in Lee & Huo (2000) and Shinoda (2010). We can also consider the Bayesian treatment of this indirect estimation of transformation parameters Λ, which is discussed in Section 7.4.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

4.4.2 MAP-based speaker adaptation

In speaker adaptation using the MAP estimation (MAP adaptation), we directly estimate the SD CDHMM parameters Θ^{SD}, unlike the MLLR and eigenvoice techniques.[5] The MAP estimation can avoid the over-training problem. Then, we use SI CDHMM parameters Θ^{SI} as hyperparameters of the prior distributions, i.e.,

$$
\begin{aligned}
\Theta^{\text{SD, MAP}} &= \arg\max_{\Theta^{\text{SD}'}} Q^{\text{MAP}}(\Theta^{\text{SD}'}|\Theta^{\text{SD}}; \Psi(\Theta^{\text{SI}})), && (4.106)\\
&= \arg\max_{\Theta^{\text{SD}'}} Q^{\text{ML}}(\Theta^{\text{SD}'}|^{\text{SD}}) + \log p(\Theta^{\text{SD}'}|\Psi(\Theta^{\text{SI}})), && (4.107)
\end{aligned}
$$

where $p(\Theta^{\text{SD}'}|\Psi(\Theta^{\text{SI}}))$ is a prior distribution with hyperparameters of the prior distribution, and it is set as a conjugate distribution of CDHMM, as discussed in Section 4.3.3. We discuss below how to set Θ^{SI} to hyperparameters Ψ in detail.

Let π_j^{SI}, a_{ij}^{SI}, ω_{jk}^{SI}, $\boldsymbol{\mu}_{jk}^{\text{SI}}$, and Σ_{jkd}^{SI} be the SI CDHMM parameters with diagonal covariance. Although there are several ways to determine hyperparameters from the speaker-independent HMM parameters, we can set the following relationship between hyperparameters and SI parameters by using Eqs. (4.96), (4.99), and (4.103):

$$
\begin{cases}
\dfrac{\phi_j^{\pi}-1}{\sum_{j'=1}^{J}(\phi_{j'}^{\pi}-1)} & = \pi_j^{\text{SI}},\\
\dfrac{\phi_{ij}^{a}-1}{\sum_{j'=1}^{J}(\phi_{ij'}^{a}-1)} & = a_{ij}^{\text{SI}},\\
\dfrac{\phi_{jk}^{\omega}-1}{\sum_{k'=1}^{K}(\phi_{jk'}^{\omega}-1)} & = \omega_{jk}^{\text{SI}},\\
\boldsymbol{\mu}_{jk}^{0} & = \boldsymbol{\mu}_{jk}^{\text{SI}},\\
\dfrac{r_{jkd}^{0}}{\phi_{jk}^{\mathbf{R}}-2} & = \Sigma_{jk}^{\text{SI}}.
\end{cases}
\tag{4.108}
$$

This equation is obtained based on the constraint that we can obtain the SI performance when the amount of adaptation data for the target speaker is zero. To satisfy the above equations, we can use the following hyperparameter setting:

$$
\begin{cases}
\phi_j^{\pi} & = \lambda\pi_j^{\text{SI}} + 1,\\
\phi_{ij}^{a} & = \lambda a_{ij}^{\text{SI}} + 1,\\
\phi_{jk}^{\omega} & = \lambda\omega_{jk}^{\text{SI}} + 1,\\
\phi_{jk}^{\boldsymbol{\mu}} & = \phi,\\
\boldsymbol{\mu}_{jk}^{0} & = \boldsymbol{\mu}_{jk}^{\text{SI}},\\
r_{jkd}^{0} & = \Sigma_{jkd}^{\text{SI}}(\phi - 2).
\end{cases}
\tag{4.109}
$$

[5] There are several approaches combining indirect adaptation via the estimation of transformation parameters and MAP-based direct estimation of CDHMM parameters (Digalakis & Neumeyer 1996, Takahashi & Sagayama 1997).

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Note that the above hyperparameter setting has two additional parameters ϕ and λ. These are often set with fixed values (e.g., $\phi = 10$, $\lambda = 1$). Thus, by substituting Eq. (4.109) into Eq. (4.80) and (4.83), the MAP estimates of SD HMM parameters are obtained as:

$$\begin{cases}
\pi_j^{\text{SD, MAP}} & = \frac{\hat{\phi}_j^{\pi}-1}{\sum_{j'=1}^{J}(\hat{\phi}_{j'}^{\pi}-1)} \\
& = \frac{\lambda\pi_j^{\text{SI}}+\xi_1(j)}{\sum_{j'=1}^{J}(\lambda\pi_{j'}^{\text{SI}}+\xi_1(j'))}, \\
a_{ij}^{\text{SD, MAP}} & = \frac{\hat{\phi}_{ij}^{a}-1}{\sum_{j'=1}^{J}(\hat{\phi}_{ij'}^{a}-1)} \\
& = \frac{\lambda a_{ij}^{\text{SI}}+\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{j'=1}^{J}(\lambda a_{ij'}^{\text{SI}}+\sum_{t=1}^{T}\xi_t(i,j'))}, \\
\omega_{jk}^{\text{SD, MAP}} & = \frac{\hat{\phi}_{jk}^{\omega}-1}{\sum_{k'=1}^{K}(\hat{\phi}_{jk'}^{\omega}-1)} \\
& = \frac{\lambda\omega_{jk}^{\text{SI}}+\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{k'=1}^{K}(\lambda\omega_{jk'}^{\text{SI}}+\sum_{t=1}^{T}\gamma_t(j,k'))}, \\
\boldsymbol{\mu}_{jk}^{\text{SD, MAP}} & = \hat{\boldsymbol{\mu}}_{jk} \\
& = \frac{\phi\boldsymbol{\mu}_{jk}^{\text{SI}}+\sum_{t=1}^{T}\gamma_t(j,k)\mathbf{o}_t}{\phi+\sum_{t=1}^{T}\gamma_t(j,k)}, \\
\Sigma_{jkd}^{\text{SD, MAP}} & = \frac{\hat{r}_{jkd}}{\hat{\phi}_{jk}^{\mathbf{R}}-2} \\
& = \frac{\sum_{t=1}^{T}\gamma_t(j,k)o_{td}^2+\phi(\mu_{jkd}^{\text{SI}})^2-(\phi+\sum_{t=1}^{T}\gamma_t(j,k))(\mu_{jkd}^{\text{SD, MAP}})^2+\Sigma_{jkd}^{\text{SI}}(\phi-2)}{\phi+\sum_{t=1}^{T}\gamma_t(j,k)-2}.
\end{cases} \tag{4.110}$$

Gauvain & Lee (1994) compare speaker adaptation performance by employing ML and MAP estimations of acoustic model parameters using the DARPA Naval Resources Management (RM) task (Price, Fisher, Bernstein *et al.* 1988). With 2 minutes of adaptation data, the ML word error rate was 31.5 % and was worse than the speaker independent word error rate (13.9 %) due to the over-training effect. However, the MAP word error rate was 8.7 %, clearly showing the effectiveness of the MAP approach. MAP estimation has also been used in speaker verification based on universal background models (Reynolds, Quatieri & Dunn 2000), which is described in Section 4.6, and in the discriminative training of acoustic models in speech recognition as a parameter smoothing technique (Povey 2003), which is described in the next section.

4.5 Regularization in discriminative parameter estimation

This section describes another well-known application of MAP estimation in discriminative training of CDHMM parameters. Discriminative training is based on discriminative criteria, which minimizes the ASR errors directly rather than maximizing likelihood values (Juang & Katagiri 1992), and improves the performance further from the ML-based CDHMM. However, discriminative training of CDHMM parameters always has a

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

problem of over-estimation, and the regularization effect of the MAP estimation helps to avoid this problem. Discriminative training has been studied by many researchers, and there are many approaches to realize it for ASR based on different discriminative criteria and optimization techniques (e.g., maximum mutual information (MMI) criterion (Bahl, Brown, de Souza *et al.* 1986), MMI with extended Baum–Welch algorithm (Normandin 1992), minimum classification error (MCE) criterion (Juang & Katagiri 1992), MCE with various gradient methods (McDermott, Hazen, Le Roux *et al.* 2007), minimum phone error (MPE) criterion (Povey & Woodland 2002), and the unified interpretation of these techniques (Schlüter, Macherey, Müller *et al.* 2001, Nakamura, McDermott, Watanabe *et al.* 2009)).

This section explains the regularization effect of the MMI estimation of HMM parameters with the extended Baum–Welch algorithm (Povey & Woodland 2002). In this section, we limit the discussion of discriminative training to focus on introducing the application of MAP estimation.

4.5.1 Extended Baum–Welch algorithm

The MMI estimation of HMM parameters can be performed by the extended Baum–Welch algorithm or variants of gradient based methods. The MMI estimation starts from the following objective function based on the posterior distribution of the word sequence:

$$\begin{aligned}\mathcal{F}^{\mathrm{MMI}}(\Theta) &= \sum_{r=1}^{R} \log p(W_r|\mathbf{O}_r;\Theta)\\ &= \sum_{r=1}^{R} \log \frac{\sum_{S_{W_r}} \left(p\left(\mathbf{O}_r, S_{W_r}|\Theta\right)\right)^{\kappa} p_L(W_r)}{\sum_{W}\sum_{S_W} \left(p\left(\mathbf{O}_r, S_W|\Theta\right)\right)^{\kappa} p_L(W)},\end{aligned} \tag{4.111}$$

where $\mathbf{O}_r = \{\mathbf{o}_t|t = 1, \cdots, T_r\}$ is the rth utterance's acoustic feature sequence whose length is T_r. The total number of the utterances is R. W_r is a correct word sequence of the utterance r, and S_{W_r} is a set of all possible state sequences given W_r.[6] Similarly, W is a word sequence hypothesis, and the summation over W is performed among all possible word sequences. κ is the acoustic score scale, and $p\left(\mathbf{O}_r, S_{W_r}|\Theta\right)$ is an acoustic likelihood, and p_L is the language model probability. Θ is a set of all acoustic model (CDHMM) parameters for all context-dependent phonemes, unlike the definition of the CDHMM parameters for single context-dependent phonemes in Section 4.2. The MMI estimate of Θ can be obtained by optimizing this objective function as follows:

[6] The summation over state sequences S_{W_r} in the numerator in Eq. (4.111) is often approximated by the Viterbi sequence without the summation obtained by the Viterbi algorithm in Section 3.3.2, i.e.,

$$\sum_{S_{W_r}} p\left(\mathbf{O}_r, S_{W_r}|\Theta\right) \approx \max_{S_{W_r}} p\left(\mathbf{O}_r, S_{W_r}|\Theta\right). \tag{4.112}$$

Similarly, the exact summation over W in the denominator is almost impossible in the large-scale ASR, and it is also approximated by the summation over pruned word sequences in a lattice, which is obtained after the ASR decoding process.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$\Theta^{\mathrm{DT}} = \arg\max_{\Theta} \mathcal{F}^{\mathrm{MMI}}(\Theta). \tag{4.113}$$

When we only consider the numerator of Eq. (4.111) in this optimization, that corresponds to the maximum likelihood estimation of the CDHMM parameters.

By using the extended Baum–Welch algorithm, a new mean vector and variance at dimension d are iteratively updated from the previously estimated $\mu_{jkd}^{\mathrm{DT}}[\tau]$ and $\Sigma_{jkd}^{\mathrm{DT}}[\tau]$[7] at the τ iteration step, as follows:

$$\begin{aligned}
\boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau+1] &= \frac{\boldsymbol{\gamma}_{jk}^{(1),\mathrm{num}} - \boldsymbol{\gamma}_{jk}^{(1),\mathrm{den}} + D\boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau]}{\gamma_{jk}^{\mathrm{num}} - \gamma_{jk}^{\mathrm{den}} + D}, \\
\Sigma_{jkd}^{\mathrm{DT}}[\tau+1] &= \frac{\gamma_{jkd}^{(2),\mathrm{num}} - \gamma_{jkd}^{(2),\mathrm{den}} + D\left(\Sigma_{jkd}^{\mathrm{DT}}[\tau] + \left(\mu_{jkd}^{\mathrm{DT}}[\tau]\right)^2\right)}{\gamma_{jk}^{\mathrm{num}} - \gamma_{jk}^{\mathrm{den}} + D} - \left(\mu_{jkd}^{\mathrm{DT}}[\tau+1]\right)^2 .
\end{aligned} \tag{4.114}$$

The derivation of the extended Baum–Welch algorithm can also be found in Section 5.2.8. Here, $\gamma_{jk}^{\mathrm{num}}$, $\boldsymbol{\gamma}_{jk}^{(1),\mathrm{num}}$, and $\gamma_{jkd}^{(2),\mathrm{num}}$ are the Gaussian sufficient statistics defined in Eq. (4.93), but these are obtained from the numerator of the lattice. Similarly, $\gamma_{jk}^{\mathrm{den}}$, $\boldsymbol{\gamma}_{jk}^{(1),\mathrm{den}}$, and $\gamma_{jkd}^{(2),\mathrm{den}}$ are obtained from the denominator of the lattice. D is a smoothing parameter used with the previous estimated parameters.

By comparison with the ML estimates of $\boldsymbol{\mu}$ and Σ in Eq. (3.151), which is only computed from the Gaussian sufficient statistics, the MMI estimates are computed from $\boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau]$ and $\Sigma_{jkd}^{\mathrm{DT}}[\tau]$, and the numerator and denominator statistics. This is the main difference between ML and MMI estimation methods. However, by setting D, $\gamma_{jk}^{\mathrm{den}}$, $\boldsymbol{\gamma}_{jk}^{(1),\mathrm{den}}$, and $\gamma_{jkd}^{(2),\mathrm{den}}$ to 0, Eq. (4.114) is close to the ML estimates if we consider that the numerator statistics can be regarded as the statistics used in the ML–EM, i.e.,

$$\begin{aligned}
\lim_{D,\gamma^{\mathrm{den}}\to 0} \boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau+1] &= \frac{\boldsymbol{\gamma}_{jk}^{(1),\mathrm{num}}}{\gamma_{jk}^{\mathrm{num}}} \approx \boldsymbol{\mu}_{jk}^{\mathrm{ML}}, \\
\lim_{D,\gamma^{\mathrm{den}}\to 0} \Sigma_{jkd}^{\mathrm{DT}}[\tau+1] &= \frac{\gamma_{jkd}^{(2),\mathrm{num}}}{\gamma_{jk}^{\mathrm{num}}} - \left(\frac{\gamma_{jkd}^{(1),\mathrm{num}}}{\gamma_{jk}^{\mathrm{num}}}\right)^2 \approx \Sigma_{jkd}^{\mathrm{ML}}.
\end{aligned} \tag{4.115}$$

Therefore, the MMI estimate can also involve the ML-like solution in a specific limitation.

We can further provide an interesting interpretation of the MMI estimate. First we focus on the following difference statistics between the numerator and denominator statistics:

$$\begin{aligned}
\delta_{jk} &\triangleq \gamma_{jk}^{\mathrm{num}} - \gamma_{jk}^{\mathrm{den}}, \\
\boldsymbol{\delta}_{jk}^{(1)} &\triangleq \boldsymbol{\gamma}_{jk}^{(1),\mathrm{num}} - \boldsymbol{\gamma}_{jk}^{(1),\mathrm{den}}, \\
\delta_{jkd}^{(2)} &\triangleq \gamma_{jkd}^{(2),\mathrm{num}} - \gamma_{jkd}^{(2),\mathrm{den}}.
\end{aligned} \tag{4.116}$$

[7] Note again that Σ means the diagonal component of the covariance matrix, and does not mean the standard deviation σ.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Then, we can rewrite Eq. (4.114) with these difference statistics as follows:

$$\begin{aligned}\boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau+1] &= \frac{\boldsymbol{\delta}_{jk}^{(1)} + D\boldsymbol{\mu}_{jk}^{\mathrm{DT}}[\tau]}{\delta_{jk}+D},\\ \Sigma_{jkd}^{\mathrm{DT}}[\tau+1] &= \frac{\delta_{jkd}^{(2)} + D\left(\Sigma_{jkd}^{\mathrm{DT}}[\tau] + \left(\mu_{jkd}^{\mathrm{DT}}[\tau]\right)^2\right)}{\delta_{jk}+D} - \left(\mu_{jkd}^{\mathrm{DT}}[\tau+1]\right)^2. \qquad (4.117)\end{aligned}$$

Therefore, Eq. (4.117) means that the MMI estimates are represented by the linear interpolation between the difference-based Gaussian statistics and the previously estimated parameters. D plays a role of tuning the linear interpolation ratio.

Note that the MMI estimates are based on the difference statistics, and if the denominator statistics are large, the difference statistics become small, and the MMI estimates would meet an over-training problem. The smoothing based on the D with the previous estimated parameters could mitigate the over-training problem, and the combination of the MMI estimation with the MAP estimation can further mitigate it.

4.5.2 MAP interpretation of i-smoothing

In MMI and MPE training (Povey & Woodland 2002), the following smoothing terms are introduced for the numerator statistics in Eq. (4.114):

$$\begin{aligned}\gamma_{jk}^{'\mathrm{num}} &= \gamma_{jk}^{\mathrm{num}} + \eta,\\ \boldsymbol{\gamma}_{jk}^{'(1),\mathrm{num}} &= \boldsymbol{\gamma}_{jk}^{(1),\mathrm{num}} + \eta\boldsymbol{\mu}_{jk}^{0},\\ \gamma_{jkd}^{'(2),\mathrm{num}} &= \gamma_{jkd}^{(2),\mathrm{num}} + \eta\left((\mu_{jkd}^{0})^2 + \Sigma_{jkd}^{0}\right), \qquad (4.118)\end{aligned}$$

where η is called the i-smoothing factor. This section reviews this statistics update, which can be interpreted as the MAP estimation where η behaves as a hyperparameter in the MAP estimation. μ_{jkd}^0 and Σ_{jkd}^0 are obtained from the maximum likelihood estimation (i.e., $\mu_{jkd}^0 = \mu_{jkd}^{\mathrm{ML}}$ and $\Sigma_{jkd}^0 = \Sigma_{jkd}^{\mathrm{ML}}$), or estimation based on discriminative training.

To derive the above smoothing factor, we first consider the conjugate distribution of the diagonal-covariance Gaussian distribution, which is based on the Gaussian–gamma distribution, as shown in Table 2.1. The Gaussian–gamma distribution is defined in Appendix C.13 as follows:

$$\begin{aligned}&\mathcal{N}\mathrm{Gam}(\mu, r|\mu^0, \phi^\mu, r^0, \phi^r)\\ &\quad = C_{\mathcal{N}\mathrm{Gam}}(\phi^\mu, r^0, \phi^r) r^{\frac{\phi^r-1}{2}} \exp\left(-\frac{r^0 r}{2} - \frac{\phi^\mu r(\mu-\mu^0)^2}{2}\right), \qquad (4.119)\end{aligned}$$

where we omit state index j, mixture component k, and dimension index d for simplicity. By setting hyperparameters μ^0, ϕ^μ, r^0, and ϕ^r with the following variables:

$$\begin{cases} \phi^{\mu} = \eta, \\ r^0 = \Sigma^0 \eta, \\ \phi^r = \eta + 2, \end{cases} \tag{4.120}$$

the prior distribution is represented as follows:

$$\begin{aligned} p(\mu, \Sigma) &= \mathcal{N}\mathrm{Gam}(\mu, r | \mu^0, \eta, \Sigma^0 \eta, \eta + 2) \\ &\propto r^{\frac{\eta + \frac{1}{2}}{2}} \exp\left(-\frac{\Sigma^0 r \eta}{2} - \frac{\eta r (\mu - \mu^0)^2}{2} \right). \end{aligned} \tag{4.121}$$

By using this prior distribution similarly to the MAP auxiliary function in Section 4.2, the MMI objective function with the prior distribution can be obtained as follows:

$$\mathcal{F}^{\mathrm{MMI}}(\mu, \Sigma) + \log p(\mu, \Sigma). \tag{4.122}$$

Based on the extended Baum–Welch calculation with the additional prior distribution, we can obtain the following update equation (Povey & Woodland 2002):

$$\begin{aligned} \mu^{\mathrm{DT}}[\tau + 1] &= \frac{\gamma^{(1),\mathrm{num}} + \eta \mu^0 - \gamma^{(1),\mathrm{den}} + D \mu^{\mathrm{DT}}[\tau]}{\gamma^{\mathrm{num}} + \eta - \gamma^{\mathrm{den}} + D}, \\ \Sigma^{\mathrm{DT}}[\tau + 1] &= \frac{\gamma^{(2),\mathrm{num}} + \eta \left((\mu^0)^2 + \Sigma^0 \right) - \gamma^{(2),\mathrm{den}} + D \left(\Sigma^{\mathrm{DT}}[\tau] + \left(\mu^{\mathrm{DT}}[\tau] \right)^2 \right)}{\gamma^{\mathrm{num}} + \eta - \gamma^{\mathrm{den}} + D} \\ &\quad - \left(\mu^{\mathrm{DT}}[\tau + 1] \right)^2. \end{aligned} \tag{4.123}$$

This equation is based on Eq. (4.117), with the effect of prior distribution $p(\mu, \Sigma)$ through Eq. (4.118).

Below, we discuss this update equation with the MAP solution by following the similar discussion in the previous section based on the ML–EM conversion. By setting D, γ^{den}, $\gamma^{(1),\mathrm{den}}$, and $\gamma^{(2),\mathrm{den}}$ to 0, Eq. (4.123) is represented as follows:

$$\lim_{D, \gamma^{\mathrm{den}} \to 0} \mu^{\mathrm{DT}}[\tau + 1] = \frac{\gamma^{(1),\mathrm{num}} + \eta \mu^0}{\gamma^{\mathrm{num}} + \eta}, \tag{4.124}$$

$$\begin{aligned} \lim_{D, \gamma^{\mathrm{den}} \to 0} \Sigma^{\mathrm{DT}}[\tau + 1] &= \frac{\gamma^{(2),\mathrm{num}} + \eta \left((\mu^0)^2 + \Sigma^0 \right)}{\gamma^{\mathrm{num}} + \eta} \\ &\quad - \left(\frac{\gamma^{(1),\mathrm{num}} + \eta \mu^0}{\gamma^{\mathrm{num}} + \eta} \right)^2. \end{aligned} \tag{4.125}$$

By comparing the MAP solutions for μ and Σ in Eq. (4.83), we find that Eqs. (4.124) and (4.125) correspond to the MAP solutions. Thus, we have found that the i-smoothing in the MMI and MPE solutions can be interpreted as MAP. From Eq. (4.123), the smoothing terms that come from D can also be similarly interpreted as MAP, when we consider the following Gaussian–gamma prior distribution:

$$p(\mu, \Sigma) = \mathcal{N}\mathrm{Gam}(\mu, r | \mu^{\mathrm{DT}}[\tau], D, \Sigma^{\mathrm{DT}}[\tau] D, D + 2). \tag{4.126}$$

Therefore, we could also provide the MAP interpretation of the D-related terms in the extended Baum–Welch algorithm. However, how to provide this prior distribution in the objective function is not trivial, and the theoretical analysis of this interpretation in the discriminative training framework is an interesting open question.

Povey & Woodland (2002) use this i-smoothing technique with MMI and MPE estimation methods, and report 1% absolute WER improvement from the MMI estimation method without the i-smoothing technique. In addition, using the speaker independent HMM parameters as prior hyperparameters (Povey, Gales, Kim *et al.* 2003, Povey, Woodland & Gales 2003) also realizes discriminative acoustic model adaptation based on the MAP estimation. There are several studies of using Bayesian approaches to discriminative training of acoustic models (e.g., based on minimum relative entropy discrimination (Kubo, Watanabe, Nakamura *et al.* 2010)), and Section 5.2 also introduces the Bayesian sensing HMM with discriminative training based on the evidence framework.

4.6 Speaker recognition/verification

In this section we focus on text-independent speaker recognition or speaker verification systems, and show how MAP estimation is used. Speaker recognition is a similar problem to automatic speech recognition. Let $\mathbf{O} \in \mathbb{R}^D$ be a sequence of D dimensional speech feature vectors. Usually $\mathbf{O}$ is one utterance by a specific speaker. Similarly to speech recognition, MFCC is usually used as a feature. The speaker recognition task is to estimate speaker label $\hat{c}$ among a speaker set $\mathcal{C}$ by using the maximum a-posteriori estimation for the posterior distribution:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} p(c|\mathbf{O}), \tag{4.127}$$

where $p(c|\mathbf{O})$ is obtained from a statistical speaker model, and is discussed later. This is similar to a speech recognition problem, as shown in Eq. (3.2) by replacing the estimation target from the word sequence W with the speaker index c. Since the output is not structured compared with W, speaker recognition can be realized by relatively simple models compared with ASR.

Speaker verification is to determine whether $\mathbf{O}$ is spoken by a target speaker s, which is regarded as single-speaker detection. This is reformulated as a basic test between two hypotheses:

- H_0 : $\mathbf{O}$ is from the hypothesized speaker s;
- H_1 : $\mathbf{O}$ is *not* from the hypothesized speaker s.

And the verification is performed by comparing the posterior distributions of H_0 and H_1 as follows:

$$\frac{p(H_0|\mathbf{O})}{p(H_1|\mathbf{O})} \begin{cases} \geq \epsilon & \text{accept} \quad H_0 \\ < \epsilon & \text{reject} \quad H_0 \end{cases}, \tag{4.128}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

where ϵ is a decision threshold. By using the Bayes rule, we can rewrite Eq. (4.128) with a likelihood ratio as follows:

$$\frac{p(H_0|\mathbf{O})}{p(H_1|\mathbf{O})} = \frac{p(\mathbf{O}|H_0)p(H_0)}{p(\mathbf{O}|H_1)p(H_1)} \approx \frac{p(\mathbf{O}|H_0)}{p(\mathbf{O}|H_1)}, \tag{4.129}$$

where we disregard the contribution of the prior distributions of each hypothesis. Thus, by using the generative model of the H_0 and H_1, we can compare the two hypotheses. We use GMM as these generative models. Hence, $p(\mathbf{O}|H_0)$ and $p(\mathbf{O}|H_1)$ are represented from $p(\mathbf{O}|\Theta_{H_0})$ and $p(\mathbf{O}|\Theta_{H_1})$, where Θ_{H_0} and Θ_{H_1} are sets of GMM parameters.

4.6.1 Universal background model

The generative model of H_1 must consider the characteristics of many speakers. That can be achieved by training the GMM parameters Θ_{H_1} from many speakers. The GMM of H_1 is called the *universal background model* (UBM), and Θ^{UBM} denotes its GMM parameters. Therefore, the likelihood for test data $\mathbf{O}$ can be computed by:

$$p(\mathbf{O}|H_1) = p(\mathbf{O}|\Theta^{\text{UBM}}) = \prod_{t=1}^{T}\sum_{k=1}^{K} \omega_k^{\text{UBM}} \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_k^{\text{UBM}}, \mathbf{\Sigma}_k^{\text{UBM}}), \tag{4.130}$$

where Θ^{UBM} are computed from many training data $\mathcal{O}$ in advance, uttered by various speakers to train Θ^{UBM} with the ML training as follows:

$$\Theta^{\text{UBM}} = \arg\max_{\Theta} p(\mathcal{O}|\Theta), \tag{4.131}$$

which can be performed efficiently by using the EM algorithm, as we discussed in Section 3.4.

While we use many data $\mathcal{O}$ to train Θ^{UBM}, the hypothesis speaker model H_0 with model parameters Θ^{HYP} can be trained by using only a small amount of data $\mathbf{O}$ (e.g., one utterance). Although ML has an over-training problem in this setting, MAP estimation can avoid the problem, and estimate Θ^{HYP} as follows:

$$\begin{aligned}\Theta^{\text{HYP}} &= \arg\max_{\Theta} p(\Theta|\mathbf{O}) \\ &= \arg\max_{\Theta} p(\mathbf{O}|\Theta)p(\Theta|\Psi(\Theta^{\text{UBM}})).\end{aligned} \tag{4.132}$$

where $p(\Theta|\Psi(\Theta^{\text{UBM}}))$ is a prior distribution and $\Psi(\Theta^{\text{UBM}})$ are hyperparameters of GMM parameters. Note that some of the hyperparameters are obtained from the UBM parameters Θ^{UBM}. This is a very similar technique to MAP adaptation of CDHMM parameters, as we discussed in Section 4.4.2, where the target model is based on a speaker-dependent CDHMM while the prior model is based on a speaker-independent CDHMM. Equation (4.132) is a subset solution of CDHMM, as we discussed in Section 4.3, and the Θ^{HYP} is obtained as follows:

$$\begin{cases} \omega_k^{\text{HYP}} &= \frac{\hat{\phi}_k^{\omega} - 1}{\sum_{k'=1}^{K}(\hat{\phi}_{k'}^{\omega} - 1)}, \\ \boldsymbol{\mu}_k^{\text{HYP}} &= \hat{\boldsymbol{\mu}}_k, \\ \Sigma_{kd}^{\text{HYP}} &= \frac{\hat{r}_{kd}}{\hat{\phi}_k^{\mathbf{R}} - 2}, \end{cases} \tag{4.133}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

where

$$\begin{cases} \hat{\phi}_k^{\omega} & \triangleq \phi_k^{\omega} + \sum_{t=1}^{T} \gamma_t(k), \\ \hat{\phi}_k^{\boldsymbol{\mu}} & \triangleq \phi_k^{\boldsymbol{\mu}} + \sum_{t=1}^{T} \gamma_t(k), \\ \hat{\boldsymbol{\mu}}_k & \triangleq \frac{\phi_k^{\boldsymbol{\mu}} \boldsymbol{\mu}_k^0 + \sum_{t=1}^{T} \gamma_t(k) \mathbf{o}_t}{\phi_k^{\boldsymbol{\mu}} + \sum_{t=1}^{T} \gamma_t(k)}, \\ \hat{\phi}_k^{\mathbf{R}} & \triangleq \phi_k^{\mathbf{R}} + \sum_{t=1}^{T} \gamma_t(k), \\ \hat{r}_{kd} & \triangleq \sum_{t=1}^{T} \gamma_t(k) o_{td}^2 + \phi_k^{\boldsymbol{\mu}} (\mu_{kd}^0)^2 - \hat{\phi}_k^{\boldsymbol{\mu}} (\hat{\mu}_{kd})^2 + r_{kd}^0. \end{cases} \tag{4.134}$$

Reynolds *et al.* (2000) suggest using specific hyperparameter settings for ϕ_k^{ω}, $\phi_k^{\boldsymbol{\mu}}$, $\boldsymbol{\mu}_k^0$, r_{kd}^0 to obtain the following forms:

$$\begin{cases} \omega_k^{\mathrm{HYP}} = \dfrac{\alpha_k^w \gamma_k / T + (1 - \alpha_k^w) \omega_k^{\mathrm{UBM}}}{\sum_{k'=1}^{K} \alpha_{k'}^w \gamma_{k'} / T + (1 - \alpha_{k'}^w) \omega_{k'}^{\mathrm{UBM}}}, \\ \boldsymbol{\mu}_k^{\mathrm{HYP}} = \alpha_k^m \boldsymbol{\gamma}_k^{(1)} + (1 - \alpha_k^m) \boldsymbol{\mu}_k^{\mathrm{UBM}}, \\ \Sigma_{kd}^{\mathrm{HYP}} = \alpha_k^v \gamma_{kd}^{(2)} + (1 - \alpha_k^v)((\mu_{kd}^{\mathrm{UBM}})^2 + \Sigma_{kd}^{\mathrm{UBM}}) - (\mu_{kd}^{\mathrm{HYP}})^2, \end{cases} \tag{4.135}$$

where α_k^w, α_k^m, and α_k^v are hyperparameters, and can be controlled by a tuning parameter β, as follows:

$$\alpha_k^{w,m,v} = \frac{\gamma_k}{\gamma_k + \beta}. \tag{4.136}$$

Note that this solution also has the MAP property of avoiding sparse data problems.

The hypothesis test in Eq. (4.128) can be performed by considering the likelihood ratio test of UBM and HYP GMMs as

$$\frac{p(\mathbf{O}|\Theta^{\mathrm{HYP}})}{p(\mathbf{O}|\Theta^{\mathrm{UBM}})} \begin{cases} \geq \epsilon & \text{accept} \quad H_0 \\ < \epsilon & \text{reject} \quad H_0. \end{cases} \tag{4.137}$$

Thus, we have shown that MAP estimation plays an important role in speaker verification based on UBM, especially in estimating the hypothesis speaker model.

4.6.2 Gaussian super vector

The MAP estimation of speaker models is further developed by using the Gaussian super vector technique (Campbell, Sturim & Reynolds 2006). The idea of this approach is to consider the MAP estimated GMM parameters as a feature of speaker verification or speaker recognition. The verification/recognition is performed by using a multi-class Support Vector Machine (SVM) (Vapnik 1995), cosine similarity scoring, or other simple classifier. Suppose we have $\mathbf{O}_n$ features, the GMM–UBM process can create the following super vector by concatenating the Gaussian mean vector $\{\boldsymbol{\mu}_{k,n}^{\mathrm{HYP}} | k = 1, \cdots, K\}$, estimated from $\mathbf{O}_n$:

$$\boldsymbol{\mu}_n \triangleq \begin{bmatrix} \boldsymbol{\mu}_{1,n}^{\mathrm{HYP}} \\ \boldsymbol{\mu}_{2,n}^{\mathrm{HYP}} \\ \vdots \\ \boldsymbol{\mu}_{K,n}^{\mathrm{HYP}} \end{bmatrix}. \tag{4.138}$$

The super vector is also obtained by using the vectorized form of the transformation matrix estimated by using the MLLR algorithm (Stolcke *et al.* 2005). The technique is widely used for speaker and language recognition tasks (Kinnunen & Li 2010), and it can usually be used with factor analysis techniques (Kenny 2010, Dehak, Kenny, Dehak *et al.* 2011) by representing the super vector with the speaker-specific and other (channel) factors:

$$\boldsymbol{\mu}_n = \boldsymbol{\mu} + \underbrace{\mathbf{U}_1\mathbf{x}_1}_{\text{speaker}} + \underbrace{\mathbf{U}_2\mathbf{x}_{2n}}_{\text{channel}} + \boldsymbol{\epsilon}_n, \tag{4.139}$$

where the speaker and channel specific factors are also represented by the linear model with the transformation matrices $\mathbf{U}_1$ and $\mathbf{U}_2$. $\boldsymbol{\mu}$ is a bias vector, and $\boldsymbol{\epsilon}_n$ is a noise vector. The approach is also applied to video processing (Shinoda & Inoue 2013). Thus, MAP estimation is still used as an important component of speaker verification tasks, but the techniques have been developed further based on the above factor analysis. Section 7.5 describes a VB solution of this factor analysis.

4.7 *n*-gram adaptation

MAP estimation is also used to obtain the n-gram language model (Federico 1996, Masataki, Sagisaka, Hisaki *et al.* 1997). In the n-gram language model, the generative probability of the word sequence $w_1^N = \{w_i \in \mathcal{V} | i = 1, \cdots, N\}$ with vocabulary $\mathcal{V}$ can be basically represented as a product of multinomial distributions, as discussed in Section 3.6:

$$\begin{aligned} p_\Theta(w_1^N) = \prod_{i=1}^{N} p(w_i|w_1^{i-1}) &\approx \prod_{i=1}^{N} p(w_i|w_{i-n+1}^{i-1}) \\ &\approx \prod_{i=1}^{N} \mathrm{Mult}(w_i|\theta_{w_i|w_{i-n+1}^{i-1}}), \end{aligned} \tag{4.140}$$

where $\Theta = \{\theta_{w_i|w_{i-n+1}^{i-1}}\}$ denotes the n-gram parameters. As discussed in Eq. (3.197), the ML estimate of Θ is obtained by using the number of occurrences $c(w_{i-n+1}^{i})$ of word sequence w_{i-n+1}^{i} in training corpus $\mathcal{D}$:

$$\begin{aligned} \theta_{w_i|w_{i-n+1}^{i-1}}^{\mathsf{ML}} &= \arg \max_{\theta_{w_i|w_{i-n+1}^{i-1}}} p(\mathcal{D}|\theta_{w_i|w_{i-n+1}^{i-1}}) \\ &= \frac{c(w_{i-n+1}^{i})}{\sum_{w_i} c(w_{i-n+1}^{i})}. \end{aligned} \tag{4.141}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Note that we do not consider the smoothing techniques in this section to make the discussion simple.

4.7.1 MAP estimation of *n*-gram parameters

The MAP extension from the above ML framework can be performed by considering the posterior distribution and introducing the prior distribution as follows:

$$\begin{aligned}\theta^{\mathsf{MAP}}_{w_i|w_{i-n+1}^{i-1}} &= \arg\max_{\theta_{w_i|w_{i-n+1}^{i-1}}} p(\theta_{w_i|w_{i-n+1}^{i-1}}|\mathcal{D}) \\ &= \arg\max_{\theta_{w_i|w_{i-n+1}^{i-1}}} p(\mathcal{D}|\theta_{w_i|w_{i-n+1}^{i-1}})p(\theta_{w_i|w_{i-n+1}^{i-1}}).\end{aligned} \tag{4.142}$$

Since $p(\mathcal{D}|\theta_{w_i|w_{i-n+1}^{i-1}})$ is a multinomial distribution, as discussed in Section 2.1.4, we use the following Dirichlet distribution for $p(\theta_{w_i|w_{i-n+1}^{i-1}})$ in Appendix C.4:

$$p(\theta_{w_i|w_{i-n+1}^{i-1}}) = \mathrm{Dir}(\{\theta_{w_i|w_{i-n+1}^{i-1}}\}_{w_i}|\{\phi_{w_i|w_{i-n+1}^{i-1}}\}_{w_i}). \tag{4.143}$$

Thus, we can analytically solve Eq. (4.142) as follows:

$$\theta^{\mathsf{MAP}}_{w_i|w_{i-n+1}^{i-1}} = \frac{\phi_{w_i|w_{i-n+1}^{i-1}} - 1 + c(w_{i-n+1}^{i})}{\sum_{w_i}\phi_{w_i|w_{i-n+1}^{i-1}} - 1 + c(w_{i-n+1}^{i})}. \tag{4.144}$$

This is a similar result to the MAP solutions of mixture weights or transition probabilities in Section 4.3.5. Since the *n*-gram parameter estimation in this setting does not include the latent variables, we can obtain the solution without using the EM algorithm. Therefore, the difference between Eq. (4.144) and those in Section 4.3.5 is between using discrete counts $c(w_{i-n+1}^{i})$ or EM-based expected counts γ and ξ, which are continuous values. Note that the parameters represented by a Dirichlet distribution always satisfy the sum-to-one condition required for *n*-gram language modeling.

4.7.2 Adaptation method

Similarly to MAP estimation based speaker adaptation for HMM parameters, MAP estimation of *n*-gram parameters can be used for speaker/task adaptations (Federico 1996, Masataki *et al*. 1997). Let $\mathcal{D}^{\mathrm{SI}}$ be the speaker (or task) independent corpus, and $\mathcal{D}^{\mathrm{SD}}$ be the speaker (or task) dependent corpus. The following hyperparameter setting is often used:

$$\phi_{w_i|w_{i-n+1}^{i-1}} = \alpha\sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}} + 1. \tag{4.145}$$

Here $\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}}$ is obtained from the ML estimation in Eq. (4.141) by using $\mathcal{D}^{\mathrm{SI}}$. Similarly, $c^{\mathrm{SI}}(w_{i-n+1}^{i})$ is a word count obtained from $\mathcal{D}^{\mathrm{SI}}$. Then, the numerator of the MAP estimation in Eq. (4.144) is rewritten as:

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$\begin{aligned}
\theta^{\mathsf{MAP}}_{w_i|w_{i-n+1}^{i-1}} &\propto \alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}} + c(w_{i-n+1}^{i}) \\
&\propto \alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}} + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})\theta^{\mathrm{SD,\,ML}}_{w_i|w_{i-n+1}^{i-1}} \\
&\propto \frac{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})}{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i}) + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})}\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}} \\
&\quad + \frac{\sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})}{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i}) + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})}\theta^{\mathrm{SD,\,ML}}_{w_i|w_{i-n+1}^{i-1}}. \qquad (4.146)
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})}{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i}) + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})} \\
&\quad + \frac{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})}{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i}) + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})} = 1. \qquad (4.147)
\end{aligned}$$

Therefore, Eq. (4.146) can be regarded as a well-known linear interpolation of two n-gram language model parameters $\theta^{\mathrm{SI,\,ML}}$ and $\theta^{\mathrm{SD,\,ML}}$, i.e.,

$$\begin{aligned}
\theta^{\mathsf{MAP}}_{w_i|w_{i-n+1}^{i-1}} &= \alpha(w_{i-n+1}^{i})\theta^{\mathrm{SI,\,ML}}_{w_i|w_{i-n+1}^{i-1}} \\
&\quad + \left(1 - \alpha(w_{i-n+1}^{i})\right)\theta^{\mathrm{SD,\,ML}}_{w_i|w_{i-n+1}^{i-1}}, \qquad (4.148)
\end{aligned}$$

where $\alpha(w_{i-n+1}^{i})$ is a linear interpolation ratio defined as:

$$\alpha(w_{i-n+1}^{i}) \triangleq \frac{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i})}{\alpha \sum_{w_i} c^{\mathrm{SI}}(w_{i-n+1}^{i}) + \sum_{w_i} c^{\mathrm{SD}}(w_{i-n+1}^{i})}. \qquad (4.149)$$

The linear interpolation ratio depends on the count of each corpus and hyperparameter α.

This linear interpolation based MAP solution can be regarded as an instance of well-known interpolation smoothing techniques in n-gram language modeling (Chen & Goodman 1999, Rosenfeld 2000), as discussed in Section 3.6.2. The analytical result shows that the linear interpolation technique can be viewed as the MAP estimation of n-gram parameters in a Bayesian sense.

4.8 Adaptive topic model

We are facing the era of big data. The volume of data collections grows vastly. Statistical document modeling becomes increasingly important in language processing areas. As addressed in Section 3.7.3, probabilistic latent semantic analysis (PLSA) has been developed to represent a set of documents according to the maximum likelihood (ML) principle. The semantics and statistics can be effectively captured for document representation. However, PLSA is highly sensitive to the target domain, which is continuously changing in real-world applications. Similarly to the adaptation of hidden Markov models to a new speaker in Section 4.3 and the adaptation of n-gram models to a new recognition task in Section 4.7, we are interested in adapting the topic-based document model using PLSA to a new application domain from a set of application-specific documents.

A Bayesian PLSA framework (Chien & Wu 2008) is presented to establish an adaptive topic model to improve document representation by incrementally extracting the up-to-date latent semantic information to match the changing domains at run time. The Dirichlet distribution is introduced to serve as the *conjugate priors* for PLSA parameters, which are multinomial distributed. The reproducible prior/posterior distributions facilitate two kinds of adaptation applications. One is *corrective training* while the other is *incremental learning*. An incremental PLSA is constructed to accomplish the parameter estimation as well as the hyperparameter updating. Differently from standard PLSA using an ML estimate, the Bayesian PLDA is capable of performing dynamic document indexing and modeling. The mechanism of adapting a topic model based on Bayesian PLSA is similar to the mechanisms of folding-in (Berry *et al.* 1995) and SVD updating (Bellegarda 2002) based on latent semantic analysis (LSA)(Berry *et al.* 1995, Bellegarda 2000), which is known as a nonparametric approach. The updating and downdating in an SVD-based LSA framework could not be directly applied for an ML-based PLSA framework. To add up-to-date or remove out-of-date knowledge, the adaptive PLSA is developed for document modeling. The goal of adaptive PLSA aims to use the newly collected documents, called adaptation documents, to adapt an existing PLSA model to match the domains of new queries or documents in information retrieval systems. In Chien & Wu (2008), adaptive PLSA is shown to be superior to adaptive LSA in information retrieval tasks. In what follows, we address the methods of maximum a-posteriori estimation and quasi-Bayes estimation designed for corrective training and incremental learning, respectively.

4.8.1 MAP estimation for corrective training

Corrective training is intended to use batch collection data to correct the ML-based PLSA parameters Θ^{ML} to fit new domain knowledge via the MAP estimation. In a topic model based on PLSA, two sets of multinomial parameters $\Theta = \{p(w_{(v)}|k), p(k|d_m)\}$ have been estimated in the training phase subject to the constraints of multinomial distributions as given in Eq. (3.304). The first one is the topic-dependent unigram probability $p(w_{(v)}|k)$ of a vocabulary word $w_{(v)}$, and the second one is the posterior probability $p(k|d_m)$ of topic k given an observed document d_m. According to MAP estimation, we adapt PLSA parameters $\Theta = \{p(w_{(v)}|k), p(k|d_m)\}$ by maximizing the a-posteriori probability or the sum of logarithms of likelihood function $p(\mathcal{D}|\Theta)$ of adaptation words and documents $\mathcal{D} = \{w_{(v)}, d_m | v = 1, \cdots, |\mathcal{V}|, m = 1, \cdots, M\}$ and prior distribution $p(\Theta)$:

$$\begin{aligned}\Theta^{\text{MAP}} &= \arg\max_{\Theta} p(\Theta|\mathcal{D}) \\ &= \arg\max_{\Theta} \log p(\mathcal{D}|\Theta) + \log p(\Theta). \qquad (4.150)\end{aligned}$$

Here, prior distribution $p(\Theta)$ represents the randomness of multinomial parameters $\{p(w_{(v)}|k)\}$ and $\{p(k|d_m)\}$. Again, it is mathematically attractive to select the *conjugate*

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

prior as the candidate for Bayesian inference. The Dirichlet distribution is known as the conjugate prior for multinomial parameters. Owing to the selection of conjugate prior, two properties of Bayesian learning could be obtained: 1) a closed-form solution for *rapid adaptation*; and 2) a reproducible prior/posterior distribution pair for *incremental learning*. Assuming parameters $\{p(w_{(v)}|k)\}$ and $\{p(k|d_m)\}$ are independent, the prior distribution of the entire parameter set based on Dirichlet density is expressed by

$$p(\Theta|\Psi) \propto \prod_{k=1}^{K} \left[\prod_{v=1}^{|\mathcal{V}|} p(w_{(v)}|k)^{\alpha_{vk}-1} \prod_{m=1}^{M} p(k|d_m)^{\beta_{km}-1} \right], \tag{4.151}$$

where $\Psi = \{\alpha_{vk}, \beta_{km}\}$ denote the hyperparameters of Dirichlet densities. Following the EM algorithm, we implement Eq. (4.151) by calculating the posterior auxiliary function

$$Q^{\text{MAP}}(\Theta'|\Theta) = \mathbb{E}_{(Z)}[\log p(\mathcal{D}, Z|\Theta')|\mathcal{D}, \Theta] + \log p(\Theta'|\Psi). \tag{4.152}$$

By imposing the constraints of multinomial parameters in Eq. (3.304) into the constrained optimization, we form the extended auxiliary function as

$$\begin{aligned}
&\tilde{Q}^{\text{MAP}}(\Theta'|\Theta) \\
&\propto \sum_{k=1}^{K} \sum_{v=1}^{|\mathcal{V}|} \left[\left(\sum_{m=1}^{M} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\alpha_{vk} - 1) \right) \right. \\
&\qquad \left. \times \log p'(w_{(v)}|k) \right] + \eta_w \left(1 - \sum_{v=1}^{|\mathcal{V}|} p'(w_{(v)}|k) \right) \\
&\quad + \sum_{k=1}^{K} \sum_{m=1}^{M} \left[\left(\sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\beta_{km} - 1) \right) \right. \\
&\qquad \left. \times \log p'(k|d_m) \right] + \eta_d \left(1 - \sum_{k=1}^{K} p'(k|d_m) \right),
\end{aligned} \tag{4.153}$$

which is manipulated and extended from the ML auxiliary function $Q^{\text{ML}}(\Theta'|\Theta)$ given in Eq. (3.295). In Eq. (4.153), η_w and η_d denote the Lagrange multipliers for two constraints of multinomial parameters. Then we differentiate Eq. (4.153) with respect to individual multinomial parameters $p'(w_{(v)}|k)$ and set it to zero:

$$\frac{\partial \tilde{Q}^{\text{MAP}}(\Theta'|\Theta)}{\partial p'(w_{(v)}|k)} = \frac{\sum_{m=1}^{M} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\alpha_{vk} - 1)}{p'(w_{(v)}|k)} - \eta_w = 0, \tag{4.154}$$

and obtain

$$p'(w_{(v)}|k) = \frac{1}{\eta_w} \left[\sum_{m=1}^{M} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\alpha_{vk} - 1) \right]. \tag{4.155}$$

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

By substituting this result into the constraint $\sum_{v=1}^{|\mathcal{V}|} p'(w_{(v)}|k) = 1$, we find the Lagrange parameter

$$\eta_w = \sum_{v=1}^{|\mathcal{V}|} \left[\sum_{m=1}^{M} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\alpha_{vk} - 1) \right]. \tag{4.156}$$

Accordingly, we derive the MAP estimates of two PLSA parameters in closed form:

$$p^{\mathrm{MAP}}(w_{(v)}|k) = \frac{\sum_{m=1}^{M} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\alpha_{vk} - 1)}{\sum_{j=1}^{|\mathcal{V}|} \left[\sum_{m=1}^{M} c(w_{(j)}, d_m) p(z_{w_j} = k|w_{(j)}, d_m) + (\alpha_{jk} - 1) \right]}, \tag{4.157}$$

$$\begin{aligned} p^{\mathrm{MAP}}(k|d_m) &= \frac{\sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\beta_{km} - 1)}{\sum_{j=1}^{K} \left[\sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}, d_m) p(z_{w_{(v)}} = j|w_{(v)}, d_m) + (\beta_{jm} - 1) \right]} \\ &= \frac{\sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}, d_m) p(z_{w_{(v)}} = k|w_{(v)}, d_m) + (\beta_{km} - 1)}{c(d_m) + \sum_{j=1}^{K} (\beta_{jm} - 1)}, \end{aligned} \tag{4.158}$$

where the posterior probability $p(z_{w_{(v)}} = k|w_{(v)}, d_m)$ is calculated according to Eq. (3.311) by using adaptation documents $\mathcal{D}$ based on the current estimates $\Theta = \{p(w_{(v)}|k), p(k|d_m)\}$. MAP estimates in Eq. (4.158) are seen as an extension of ML estimates of Eq. (3.308) and Eq. (3.309) by interpolating with the prior statistics $\{\alpha_{vk}\}$ and $\{\beta_{km}\}$, respectively. If prior density is non-informative or adaptation data $\mathcal{D}$ are abundant, MAP estimates are reduced to ML estimates. The MAP PLSA algorithm is developed for corrective training or batch adaptation, and adapts the existing parameters to Θ^{MAP} in a single epoch. In MAP PLSA, the Dirichlet priors and their hyperparameters $\Psi = \{\alpha_{vk}, \beta_{km}\}$ are adopted to characterize the variations of topic-dependent document and word probabilities. These priors are used to express the environmental variations. MAP PLSA involves both word-level $p(w_{(v)}|k)$ and document-level $p(k|d_m)$ parameters. In general, MAP parameters Θ^{MAP} perform better than ML parameters Θ^{ML} when classifying future documents with new terms, topics, and domains.

4.8.2 Quasi-Bayes estimation for incremental learning

Using MAP estimation, only a single learning epoch is performed to correct PLSA parameters. Batch learning is performed. However, batch learning cannot catch the continuously changing domain knowledge or deal with the non-stationary documents collected from real-world applications. An adaptive information system should continuously update system parameters with new words and topics. Out-of-date words or documents should fade away from the system as time moves on. Accordingly, we tackle the updating and downdating problems simultaneously for latent semantic indexing. MAP PLSA could not incrementally accumulate statistics for adaptive topic modeling. It is more interesting to develop an incremental learning algorithm to track the changing

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

topics and domains in test documents. A learning procedure is executed repeatedly in different epochs. Incremental learning is also known as sequential learning or online learning, which is important for speaker adaptation in automatic speech recognition systems where speaker characteristics gradually change with time (Huo & Lee 1997, Chien 1999).

To implement incremental learning for adaptive topic modeling, we continuously estimate PLSA parameters in different learning epochs using the incrementally observed adaptation documents. At the n learning epoch, we estimate PLSA parameters by maximizing the posterior distribution using a sequence of adaptation documents $\mathfrak{D}^n = \{\mathcal{D}_1, \cdots, \mathcal{D}_n\}$:

$$\begin{aligned}(\Theta^{(n)})^{\text{QB}} &= \arg\max_{\Theta} \; p(\Theta|\mathfrak{D}^n) \\ &= \arg\max_{\Theta} \; p(\mathcal{D}_n|\Theta)p(\Theta|\mathfrak{D}^{n-1}) \\ &\approx \arg\max_{\Theta} \; p(\mathcal{D}_n|\Theta)p(\Theta|\Psi^{(n-1)}),\end{aligned} \tag{4.159}$$

where the posterior distribution $p(\Theta|\mathfrak{D}^{n-1})$ is approximated by the closest tractable prior distribution $p(\Theta|\Psi^{(n-1)})$ with *sufficient statistics* or hyperparameters $\Psi^{(n-1)}$ which are evolved from history documents $\mathfrak{D}^{n-1}$. This estimation method is also called the quasi-Bayes (QB) estimation (Huo & Lee 1997, Chien 1999). QB estimation provides recursive learning of PLSA parameters,

$$\Theta^{(1)} \rightarrow \Theta^{(2)} \rightarrow \cdots \rightarrow \Theta^{(n)}, \tag{4.160}$$

from incrementally observed documents,

$$\mathcal{D}_1 \rightarrow \mathcal{D}_2 \rightarrow \cdots \rightarrow \mathcal{D}_n. \tag{4.161}$$

At each epoch, we only use the current block of documents $\mathcal{D}_n = \{w_{(v)}^{(n)}, d_m^{(n)}|v = 1, \cdots, |\mathcal{V}|, m = 1, \cdots, M_n\}$ and the accumulated statistics $\Psi^{(n-1)}$ to update PLSA parameters from $\Theta^{(n-1)}$ to $\Theta^{(n)}$. Current block data $\mathcal{D}_n$ are released after accumulating statistics from $\Psi^{(n-1)}$ to $\Psi^{(n)}$. Memory and computation requirements are reduced at each epoch. The key technique in QB PLSA comes from the introduction of *incremental hyperparameters*. If we substitute the hyperparameters $\Psi^{(n-1)} = \{\alpha_{vk}^{(n-1)}, \beta_{km}^{(n-1)}\}$ into Eq. (4.157) and Eq. (4.158), the QB estimates $(\Theta^{(n)})^{\text{QB}} = \{p^{\text{QB}}(w_{(v)}^{(n)}|k), p^{\text{QB}}(k|d_m^{(n)})\}$ are obtained. This QB PLSA method is geared with the updating mechanism of hyperparameters, which is derived by the E-step for QB estimation in Eq. (4.159). By referring to the auxiliary function of MAP PLSA in Eq. (4.152) and Eq. (4.153), the QB auxiliary function of new estimates $(\Theta^{(n)})' = \{p'(w_{(v)}^{(n)}|k), p'(k|d_m^{(n)})\}$ given current estimates $\Theta^{(n)} = \{p(w_{(v)}^{(n)}|k), p(k|d_m^{(n)})\}$ is defined by

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

$$\begin{aligned}
&Q^{\mathrm{QB}}((\Theta^{(n)})'|\Theta^{(n)}) \\
&\quad\propto \sum_{k=1}^{K}\sum_{v=1}^{|\mathcal{V}|}\Bigg[\Bigg(\sum_{m=1}^{M_n} c(w_{(v)}^{(n)}, d_m^{(n)}) p(k|w_{(v)}^{(n)}, d_m^{(n)}) \\
&\qquad\qquad + (\alpha_{vk}^{(n-1)} - 1)\Bigg) \log p'(w_{(v)}^{(n)}|k)\Bigg] \\
&\quad + \sum_{k=1}^{K}\sum_{m=1}^{M_n}\Bigg[\Bigg(\sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}^{(n)}, d_m^{(n)}) p(k|w_{(v)}^{(n)}, d_m^{(n)}) \\
&\qquad\qquad + (\beta_{km}^{(n-1)} - 1)\Bigg) \log p'(k|d_m^{(n)})\Bigg].
\end{aligned} \tag{4.162}$$

It is important that the exponential of the QB auxiliary function in Eq. (4.162) can be arranged as a new Dirichlet distribution:

$$\begin{aligned}
&\exp\left\{Q^{\mathrm{QB}}((\Theta^{(n)})'|\Theta^{(n)})\right\} \\
&\quad\propto \prod_{k=1}^{K}\left[\prod_{v=1}^{\mathcal{V}} p'(w_{(v)}^{(n)}|k)^{\alpha_{vk}^{(n)}-1} \prod_{m=1}^{M_n} p'(k|d_m^{(n)})^{\beta_{km}^{(n)}-1}\right],
\end{aligned} \tag{4.163}$$

with the updated hyperparameters $\Psi^{(n)} = \{\alpha_{vk}^{(n)}, \beta_{km}^{(n)}\}$ derived thus:

$$\alpha_{vk}^{(n)} = \sum_{m=1}^{M_n} c(w_{(v)}^{(n)}, d_m^{(n)}) p(k|w_{(v)}^{(n)}, d_m^{(n)}) + \alpha_{vk}^{(n-1)}, \tag{4.164}$$

$$\beta_{km}^{(n)} = \sum_{v=1}^{|\mathcal{V}|} c(w_{(v)}^{(n)}, d_m^{(n)}) p(k|w_{(v)}^{(n)}, d_m^{(n)}) + \beta_{km}^{(n-1)}, \tag{4.165}$$

where the posterior probability

$$p(k|w_{(v)}^{(n)}, d_m^{(n)}) = \frac{p(w_{(v)}^{(n)}|k) p(k|d_m^{(n)})}{\sum_{j=1}^{K} p(w_{(v)}^{(n)}|j) p(j|d_m^{(n)})} \tag{4.166}$$

is obtained by using the current block of adaptation data $\mathcal{D}_n = \{w_{(v)}^{(n)}, d_m^{(n)}\}$ based on current QB estimates $\Theta^{(n)} = \{p(w_{(v)}^{(n)}|k), p(k|d_m^{(n)})\}$. Importantly, a *reproducible distribution pair* of a prior in Eq. (4.151) and a posterior in Eq. (4.163) is established. This property is crucial to activate the updating mechanism of hyperparameters for incremental learning. New hyperparameters $\Psi^{(n)} = \{\alpha_{vk}^{(n)}, \beta_{km}^{(n)}\}$ are estimated by combining the previous hyperparameters $\Psi^{(n-1)} = \{\alpha_{vk}^{(n-1)}, \beta_{km}^{(n-1)}\}$ with the accumulated statistics from adaptation documents $\mathcal{D}_n = \{w_{(v)}^{(n)}, d_m^{(n)}\}$ at learning epoch n.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

Table 4.1 Numbers of training, adaptation, and test documents for five populous classes in the Reuters-21578 dataset.

	Acquisitions	Crude	Earn	Money-fx	Trade
Number of training documents	825	196	1447	284	189
Number of added documents per epoch	275	65	475	85	60
Number of test documents	719	189	1087	180	117

Basically, QB estimation finds the point estimate for adaptive topic modeling. This estimation is seen as an extended realization of MAP estimation by activating the mechanism of hyperparameter updating so that incremental learning is established to compensate the non-stationary variations in observation data which may be speech signals, word sequences, or text documents. Incremental learning based on QB estimation is helpful for speech and language applications including speech recognition, information retrieval, and others.

4.8.3 System performance

The performance of the adaptive topic model was evaluated through the tasks of corrective training and incremental learning. The evaluation was performed for the application of document categorization. Table 4.1 shows the set-up of experimental data of the Reuters-21578 dataset. We collected training, adaptation, and test documents from the five most populous categories in the Reuters-21578 dataset for system evaluation. Preprocessing stages of stemming and stop word removal were done. In the task of incremental learning, we used one-third of the adaptation documents at each epoch and investigated the effect of incremental learning in three learning epochs. The performance of corrective training and incremental learning was compared. Training samples of each category were roughly partitioned into half for training and the other half for adaptation. A fivefold cross validation over training and adaptation sets was performed. In the implementation, we determined PLSA probability for each test document. The cosine similarity of feature vectors between a test document and a given class model was calculated for pattern classification. The class feature vector consisted of PLSA probabilities averaging over all documents corresponding to a class. The classification error rate was computed over all test documents in five populous classes. We obtained the classification error rates for PLSA (Hofmann 1999*b*, 2001) (3.47%), SVD updating (Bellegarda 2002) (3.39%), MAP PLSA (Chien & Wu 2008) (3.04%), QB PLSA (Chien & Wu 2008) at 1st learning epoch (3.13%), QB PLSA at 2nd learning epoch (3.04%), and QB PLSA at 3rd learning epoch (3%). Corrective training using SVD updating and MAP PLSA and incremental learning using QB PLSA at different epochs decrease the classification error rates. SVD updating is worse than MAP PLSA and QB PLSA. Incremental learning using QB PLSA performs slightly better than batch learning using MAP PLSA.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press

4.9 Summary

This chapter introduced various applications of MAP approximation for model parameter posteriors. Since the approach can be easily realized from the existing ML based approaches by simply considering the regularization term based on model parameter priors, it is widely used for speech and language processing including acoustic and language modeling in ASR, speaker verification, and document processing. Although the MAP approximation can utilize the most famous Bayesian advantage of “use of prior knowledge,” it does not deal with probabilistic variables explicitly with the marginalization, and it does not fully utilize the Bayesian concept. The following chapters consider more strict Bayesian approaches for speech and language processing.

https://doi.org/10.1017/CBO9781107295360.005 Published online by Cambridge University Press