2 Bayesian approach

This chapter describes a general concept and statistics of the Bayesian approach. The Bayesian approach covers wide areas of statistics (Bernardo & Smith 2009, Gelman, Carlin, Stern *et al*. 2013), pattern recognition (Fukunaga 1990), machine learning (Bishop 2006, Barber 2012), and applications of these approaches. In this chapter, we start the discussion from the basic probabilistic theory, and mainly describe the Bayesian approach by aiming to follow a machine learning fashion of constructing and refining statistical models from data. The role of the Bayesian approach in machine learning is very important since the Bayesian approach provides a systematic way to infer unobserved variables (e.g., classification category, model parameters, latent variables, model structure) given data. This chapter limits the discussions considering the speech and language problems in the latter chapters, by providing simple probabilistic rules, and prior and posterior distributions in Section 2.1. The section also provides analytical solutions of posterior distributions of simple models. Based on the basic introduction, Section 2.2 introduces a useful representation of the relationship of probabilistic variables in the Bayesian approach, called the *Graphical model*. The graphical model representation gives us an intuitive view of statistical models even when they have complicated relationships between their variables. Section 2.3 explains the difference between Bayesian and maximum likelihood (ML) approaches. The following chapters extend the general Bayesian approach described in this chapter to deal with statistical models in speech and language processing.

2.1 Bayesian probabilities

This section describes the basic Bayesian framework based on probabilistic theory. Although some of the definitions, equations, and concepts are trivial, this section reviews the basics to assist readers to fully understand the Bayesian approach.

In the Bayesian approach, all the variables that are introduced when models are parameterized, such as model parameters and latent variables, are regarded as probabilistic variables. Thus, let a be a discrete valuable, then the Bayesian approach deals with a as a probabilistic variable, and aims to obtain $p(a)$:

$$a \rightarrow p(a). \tag{2.1}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Hereinafter, we assume that a is a discrete variable, and the expectation is performed by the summation over a for simplicity. Since $p(a)$ is a probabilistic distribution, $p(a)$ always satisfies the following condition:

$$\sum_a p(a) = 1, \quad p(a) \geq 0 \quad \forall a. \tag{2.2}$$

These properties help us to solve some calculations appearing in the following sections. In the continuous variable case, the summation $\sum$ is replaced with the integral $\int$.

2.1.1 Sum and product rules

Since the Bayesian approach treats all variables as probabilistic variables, the probabilistic theory gives us the two important probabilistic rules to govern the relationship between the variables. Let a and b be arbitrary probabilistic variables,

- Sum rule

$$p(b) = \sum_a p(a, b); \tag{2.3}$$

- Product rule

$$p(a, b) = p(a|b)p(b) = p(b|a)p(a). \tag{2.4}$$

Here, $p(a, b)$ is a joint probability that represents a probability of all possible joint events of a and b. $p(a|b)$ or $p(b|a)$ is a conditional probability. These are generalized to N probabilistic variables, for example, if we have $a_1, \cdots, a_N$ probabilistic variables, there rules are represented as:

- Sum rule

$$p(a_i) = \sum_{a_1} \cdots \sum_{a_{i-1}} \sum_{a_{i+1}} \cdots \sum_{a_N} p(a_1, \cdots, a_N); \tag{2.5}$$

- Product rule

$$\begin{aligned} p(a_1, \cdots, a_N) &= p(a_1|a_2, \cdots, a_N)p(a_2, \cdots, a_N) = \cdots \\ &= p(a_N) \prod_{n=1}^{N-1} p(a_n|a_{n+1}, \cdots, a_N). \end{aligned} \tag{2.6}$$

In a Bayesian manner, we formulate the probability distributions based on these rules. For example, the famous Bayes theorem can be derived by reforming the product rule in Eq. (2.4), as follows:

$$p(a|b) = \frac{p(a, b)}{p(b)} = \frac{p(b|a)p(a)}{p(b)} \tag{2.7}$$

$$= \frac{p(b|a)p(a)}{\sum_a p(b|a)p(a)}. \tag{2.8}$$

To derive Eq. (2.8), we use the sum and product rules for $p(a)$. The following discussion provides more practical examples based on this discussion.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

2.1.2 Prior and posterior distributions

The above Bayes theorem has an interesting meaning if we consider a conditional probability distribution of a given an observation x. The conditional distribution $p(a|x)$ is called *posterior distribution*, and the main purpose of the Bayesian approach is to infer the posterior distribution of various valuables. Based on the Bayes theorem in Eq. (2.7), the posterior distribution is decomposed in Eq. (2.9) to the following three distributions:

$$p(a|x) = \frac{p(x|a)p(a)}{p(x)} \tag{2.9}$$

$$= \frac{p(x|a)p(a)}{\sum_a p(x|a)p(a)}, \tag{2.10}$$

where $p(x|a)$ is a likelihood function of x and $p(a)$ is a distribution without considering any observation, and called *prior distribution*. $p(x)$ is a distribution of an observation, and can be computed by using $p(x|a)$ and $p(a)$ based on Eq. (2.10). In most speech processing applications, it is difficult to estimate the posterior distribution directly (Section 3.8 describes it in detail). Therefore, the posterior distribution $p(a|x)$ is indirectly estimated via this Bayes theorem, which is derived from the sum and product rules, which are equivalence equations without approximation.

Since the posterior distribution provides a probability of a given data x, this matches one of the machine learning goals of refining information of a from data x. Therefore, the posterior distribution plays an important role in machine learning, and obtaining an appropriate posterior distribution for our problems in speech and language processing is a main goal of this book.

Once we obtain the posterior distribution $p(a|x)$, we can obtain the values of a via:

- Maximum a-posteriori (MAP) procedure:

$$a^{\mathrm{MAP}} = \arg\max_a p(a|x); \tag{2.11}$$

- Expectation with respect to the posterior distribution:

$$a^{\mathrm{EXP}} = \mathbb{E}_{(a)}[a|x] \triangleq \sum_a a \cdot p(a|x). \tag{2.12}$$

The MAP and expectation are typical ways to obtain meaningful information about a given x in terms of the probabilistic theory. From Eq. (2.10), $p(x)$ is disregarded in the MAP procedure as a constant factor that is independent of a, MAP, and expectation, which makes the calculation simple. The MAP and expectation are generalized to obtain meaningful information $f(a)$ given the posterior distribution $p(a|x)$. More specifically, if we consider a likelihood function of unseen data y given a, i.e., $p(y|a)$, these procedures are rewritten as:

- Maximum a-posteriori (MAP) procedure:

$$p^{\mathrm{MAP}}(y|a) = p(y|\arg\max_a p(a|x)) = p(y|a^{\mathrm{MAP}}); \tag{2.13}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

- Expectation with respect to the posterior distribution:

$$p^{\mathrm{EXP}}(y) = \mathbb{E}_{(a)}[p(y|a)|x] \triangleq \sum_a p(y|a)p(a|x). \tag{2.14}$$

Thus, we can predict y by using these procedures. Note that the MAP procedure decides a deterministic value of a, while the expectation procedure keeps possible a for the expectation. Therefore, the MAP procedure is called *hard decision* and the expectation procedure is called *soft decision*.

The expectation is a more general operation than MAP in terms of considering the distribution shape. For example, if we approximate $p(a|x)$ with a specific Kronecker delta function $\delta(a, a^{\mathrm{MAP}})$ where $a^{\mathrm{MAP}} = \arg\max_a p(a|x)$, Eq. (2.14) is represented as

$$\begin{aligned} p^{\mathrm{EXP}}(y) &= \sum_a p(y|a)\delta(a, a^{\mathrm{MAP}}) = p(y|a^{\mathrm{MAP}}) \\ &= p(y|\arg\max_a p(a|x)) \\ &= p^{\mathrm{MAP}}(y|a), \end{aligned} \tag{2.15}$$

where

$$\delta(a, a') = \begin{cases} 1 & a = a' \\ 0 & \text{otherwise.} \end{cases} \tag{2.16}$$

Thus, the MAP value is obtained from the specific case of the expectation value without considering the distribution shape of $p(a|x)$. However, in many cases, the MAP value is also often used since the expectation needs a complex computation due to the summation over a. Note that the above derivation via a Kronecker delta function (or Dirac delta function when we consider continuous variables) is often used to provide the relationship of the MAP and expectation values.

2.1.3 Exponential family distributions

The previous section introduces the posterior distribution. This section focuses on a specific problem of posterior distributions that consider the model parameter Θ given a set of D dimensional observation vectors, i.e., $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^D | n = 1, \cdots, N\}$. The problem here is to obtain the posterior distribution $p(\Theta|\mathbf{X})$, i.e., it is a general estimation problem of obtaining the distribution of Θ from data $\mathbf{X}$. Once we obtain $p(\Theta|\mathbf{X})$, for example, we can estimate Θ^{MAP} or compute some expectation values, as we discussed in Section 2.1.2.

Then, the Bayes theorem, which provides the relationship between prior and posterior distributions in Eq. (2.9) or (2.10), can be represented as follows:

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{p(\mathbf{X})} \tag{2.17}$$

$$= \frac{p(\mathbf{X}|\Theta)p(\Theta)}{\int p(\mathbf{X}|\Theta)p(\Theta)d\Theta}. \tag{2.18}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Here, we use the integral $\int$ rather than $\sum$ in Eq. (2.18), since model parameters are often represented by continuous variables (e.g., mean and variance parameters in a Gaussian distribution). In this particular case, the Bayes theorem has the more practical meaning that the posterior distribution $p(\Theta|\mathbf{X})$ is represented by the likelihood function $p(\mathbf{X}|\Theta)$, and the prior distribution of the model parameters $p(\Theta)$. Thus

$$p(\mathbf{X}) = \int p(\mathbf{X}|\Theta)p(\Theta)d\Theta, \tag{2.19}$$

which is also called *evidence* function or *marginal likelihood* . The evidence plays an important role in Bayesian inference, which is described in Chapter 5 in detail.

Basically, we can set any distributions (e.g., Gaussian, gamma, Dirichlet, Laplace, Rayleigh distributions, etc.) to prior and posterior distributions. However, a particular family of distributions called *conjugate distribution* makes analytical derivation simpler. Before we describe the conjugate distributions, the following section explains *exponential family* distributions, which are required to explain conjugate distributions.

Exponential family

The *exponential family* is a general distribution family, which contains standard distributions including Gaussian distribution, gamma distribution, and multinomial distribution. Let $\boldsymbol{\theta}$ be a vector form of model parameters. A distribution of a set of observation vectors $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ given $\boldsymbol{\theta}$ (likelihood function), which belongs to the exponential family, is represented by the following exponential form:

$$p(\mathbf{X}|\boldsymbol{\theta}) \triangleq h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}(\boldsymbol{\theta})^\mathsf{T}\mathbf{t}(\mathbf{X}) - g(\boldsymbol{\gamma})\right), \tag{2.20}$$

where $\mathbf{t}(\mathbf{X})$ is a sufficient statistics vector obtained from observation vector $\mathbf{X}$, $g(\boldsymbol{\gamma})$ is a logarithmic normalization factor. $\boldsymbol{\gamma}$ is a transformed vector of $\boldsymbol{\theta}$, and is called a *natural parameter* vector. If $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, it is called the *canonical form*, that simplifies Eq. (2.20) as follows:

$$p(\mathbf{X}|\boldsymbol{\theta}) = h(\mathbf{X}) \exp\left(\boldsymbol{\theta}^\mathsf{T}\mathbf{t}(\mathbf{X}) - g(\boldsymbol{\theta})\right). \tag{2.21}$$

The canonical form makes the calculation of posterior distributions simple.

If we have J multiple parameter vectors, we can represent the exponential form as the factorized form:

$$p(\mathbf{X}|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_J) \triangleq h(\mathbf{X}) \prod_{i=1}^{J} \exp\left(\boldsymbol{\gamma}_i(\boldsymbol{\theta}_i)^\mathsf{T}\mathbf{t}_i(\mathbf{x}) - g_i(\boldsymbol{\gamma}_i)\right). \tag{2.22}$$

When the transformed model parameters are composed of a matrix or a vector, we can also define the exponential family distribution. For example, a multivariate Gaussian distribution $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is parameterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, and the corresponding transformed parameters are also represented by vector $\boldsymbol{\gamma}_1$ and matrix $\boldsymbol{\Gamma}_2$. Then, the exponential family distribution for $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\Theta}_2\}$ is defined as follows:

$$p(\mathbf{X}|\Theta) \triangleq h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}_1^\mathsf{T}\mathbf{t}_1(\mathbf{X}) + \mathrm{tr}[\boldsymbol{\Gamma}_2^\mathsf{T}\mathbf{T}_2(\mathbf{x})] - g(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2)\right). \tag{2.23}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Here, $\mathbf{t}_1$ and $\mathbf{T}_2$ are vector and matrix representations of sufficient statistics, respectively. The rest of this section provides examples of $h(\cdot), g(\cdot), \boldsymbol{\gamma}(\cdot)$, and $\mathbf{t}(\cdot)$ for standard distributions (Gaussian, multivariate Gaussian, and multinomial distributions).

Example 2.1 Gaussian (unknown mean):
We focus on the exponential family form of the Gaussian distribution for scalar observation $X = \{x_n \in \mathbb{R} | n = 1, \cdots, N\}$. As a simple example, we only focus on the Gaussian mean as a model parameter, and regard the precision parameter r as a constant value, i.e., $\mathcal{N}(x_n|\mu; r^{-1})$ where the variables located right after the semicolon; means that these are not treated as probabilistic variables, but specific values. That is $\theta = \mu$ in Eq. (2.20). We use the precision parameter r instead of the variance parameter Σ,[1] which makes the solution simple. Based on the definition in Appendix C.5, the standard form of the Gaussian distribution is represented as

$$\prod_{n=1}^{N} \mathcal{N}(x_n|\mu; r^{-1}) = \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\sum_{n=1}^{N} \frac{r}{2}(x_n - \mu)^2\right). \tag{2.24}$$

We assume that $x_1, \cdots, x_N$ are independent and identically distributed random variables from the Gaussian. The standard form of the Gaussian distribution is rewritten as the following exponential form:

$$\begin{aligned}\prod_{n=1}^{N} \mathcal{N}(x_n|\mu; r^{-1}) &= \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{r}{2}\sum_{n=1}^{N} x_n^2\right) \exp\left(r\mu \sum_{n=1}^{N} x_n - \frac{N\mu^2 r}{2}\right) \\ &= \underbrace{\left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{r}{2}\sum_{n=1}^{N} x_n^2\right)}_{=h(X)} \exp\left(\underbrace{r\sum_{n=1}^{N} x_n}_{=t(X)} \underbrace{\mu}_{=\gamma} - \underbrace{\frac{N\mu^2 r}{2}}_{=g(\gamma)}\right).\end{aligned} \tag{2.25}$$

Thus, the Gaussian distribution is represented by the following exponential form in Eq. (2.20):

$$\begin{cases} t(X) = r\sum_{n=1}^{N} x_n \\ h(X) = \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{r\sum_{n=1}^{N} x_n^2}{2}\right) \\ \gamma(\mu) = \mu \\ g(\gamma) = \frac{N\gamma^2 r}{2}. \end{cases} \tag{2.26}$$

[1] This book regards Σ as the variance parameter (not the standard deviation, which is represented as σ), as shown in Appendix C.5, to make the notation consistent with the covariance matrix $\boldsymbol{\Sigma}$.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Since $\gamma(\mu) = \mu$ in Eq. (2.26), it is regarded as a canonical form, as discussed in Eq. (2.21). Note that the parameterization of $\gamma(\mu)$ and $t(X)$ is not unique. For example, we can obtain the following parameterization from:

$$\begin{cases} t(X) = \sum_{n=1}^{N} x_n \\ h(X) = \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{r\sum_{n=1}^{N} x_n^2}{2}\right) \\ \gamma(\mu) = \mu r \\ g(\gamma) = \frac{N\mu^2 r}{2} = \frac{N\gamma^2}{2r}. \end{cases} \tag{2.27}$$

This is also another exponential form of the Gaussian distribution with unknown mean.

Example 2.2 Gaussian (unknown mean and precision):
Similarly to Example 2.1, we focus on the exponential family form of the Gaussian distribution for scalar observation X, but regard r as also unknown. Therefore, $\boldsymbol{\theta} = [\mu, r]^\intercal$. Thus, unlike the scalar forms of the natural parameter γ and sufficient statistics t in Eq. (2.24), the Gaussian distribution is represented by the vector form of these as

$$\begin{aligned} \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, r^{-1}) &= \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{N\mu^2 r}{2}\right) \exp\left(-\frac{r}{2}\sum_{n=1}^{N} x_n^2 + \mu r \sum_{n=1}^{N} x_n\right) \\ &= \exp\left(\underbrace{\begin{bmatrix} \mu r \\ r \end{bmatrix}^\intercal}_{=\boldsymbol{\gamma}(\boldsymbol{\theta})} \underbrace{\begin{bmatrix} \sum_{n=1}^{N} x_n \\ -\frac{\sum_{n=1}^{N} x_n^2}{2} \end{bmatrix}}_{\mathbf{t}(X)} - \underbrace{\left(\frac{N}{2}\log\left(\frac{2\pi}{r}\right) + \frac{N\mu^2 r}{2}\right)}_{=g(\boldsymbol{\gamma})}\right). \end{aligned} \tag{2.28}$$

Therefore,

$$\begin{cases} \mathbf{t}(X) = \begin{bmatrix} \sum_{n=1}^{N} x_n \\ -\frac{\sum_{n=1}^{N} x_n^2}{2} \end{bmatrix} \\ h(X) = 1 \\ \boldsymbol{\gamma}(\boldsymbol{\theta}) = \begin{bmatrix} \mu r \\ r \end{bmatrix} \\ g(\boldsymbol{\gamma}) = \frac{N}{2}\left(\log\frac{2\pi}{r} + \mu^2 r\right) \\ \qquad = \frac{N}{2}\left(\log\frac{2\pi}{\gamma_2} + \frac{\gamma_1^2}{\gamma_2}\right). \end{cases} \tag{2.29}$$

Again, the parameterization of $\boldsymbol{\gamma}(\boldsymbol{\theta})$ and $\mathbf{t}(x)$ is not unique and the parameterization of $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \left[\mu r, -\frac{r}{2}\right]^\intercal$ and $\mathbf{t}(X) = [\sum_{n=1}^{N} x_n, \sum_{n=1}^{N} x_n^2]^\intercal$ is also possible.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Example 2.3 Multivariate Gaussian (unknown mean and precision):
The next example is to derive an exponential form of the multivariate Gaussian distribution with D dimensional mean vector $\boldsymbol{\mu}$ and $D \times D$ precision matrix $\mathbf{R}$ (we use precision matrix $\mathbf{R}$ instead of covariance matrix $\boldsymbol{\Sigma}$ to make the solution simple). A set of the parameters is $\Theta = \{\boldsymbol{\mu}, \mathbf{R}\}$. This is the most important example in this book, since statistical models in speech and language processing are often represented by multivariate Gaussian distributions, as discussed in Chapter 3. Let $\mathbf{X} = \{\mathbf{x}_n \in \mathbb{R}^D | n = 1, \cdots, N\}$ be independent and identically distributed random variables from the multivariate Gaussian distribution. Again, based on the definition in Appendix C.6, the standard form of the Gaussian distribution is represented as

$$\begin{aligned}\prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \mathbf{R}^{-1}) &= \prod_{n=1}^{N} (2\pi)^{-\frac{D}{2}} |\mathbf{R}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal \mathbf{R}(\mathbf{x}_n - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-\frac{ND}{2}} |\mathbf{R}|^{\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal \mathbf{R}(\mathbf{x}_n - \boldsymbol{\mu})\right). \end{aligned} \tag{2.30}$$

Now we focus on the exponential part in Eq. (2.30), which is rewritten as follows:

$$\begin{aligned}&\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal \mathbf{R}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &\quad = -\boldsymbol{\mu}^\intercal \mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n - \left(\sum_{n=1}^{N}\mathbf{x}_n^\intercal\right)\mathbf{R}\boldsymbol{\mu} + \sum_{n=1}^{N}\mathbf{x}_n^\intercal \mathbf{R}\mathbf{x}_n + N\boldsymbol{\mu}^\intercal \mathbf{R}\boldsymbol{\mu}. \end{aligned} \tag{2.31}$$

To make the observation vector and parameter the inner product form, we first use the trace representation of the quadratic term of $\mathbf{x}_n$ as

$$\begin{aligned}\sum_{n=1}^{N}\mathbf{x}_n^\intercal \mathbf{R}\mathbf{x}_n &= \mathrm{tr}\left[\sum_{n=1}^{N}\mathbf{x}_n^\intercal \mathbf{R}\mathbf{x}_n\right] \\ &= \mathrm{tr}\left[\sum_{n=1}^{N}\mathbf{R}\mathbf{x}_n\mathbf{x}_n^\intercal\right] \\ &= \mathrm{tr}\left[\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\intercal\right], \end{aligned} \tag{2.32}$$

where we use the fact that the trace of the scalar value is equal to the original scalar value, the cyclic property, and the distributive property of the trace as in Appendix B:

$$a = \mathrm{tr}[a], \tag{2.33}$$

$$\mathrm{tr}[\mathbf{ABC}] = \mathrm{tr}[\mathbf{BCA}], \tag{2.34}$$

$$\mathrm{tr}[\mathbf{A}(\mathbf{B} + \mathbf{C})] = \mathrm{tr}[\mathbf{AB} + \mathbf{AC}]. \tag{2.35}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

In addition, we can also use the following equation:

$$\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n+\left(\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\right)\mathbf{R}\boldsymbol{\mu}=2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n. \tag{2.36}$$

Here, since these values are scalar values, we use the following equation to derive Eq. (2.36).

$$\begin{aligned}\left(\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\right)\mathbf{R}\boldsymbol{\mu}&=\left(\left(\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\right)\mathbf{R}\boldsymbol{\mu}\right)^{\mathsf{T}}\\&=\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}^{\mathsf{T}}\left(\sum_{n=1}^{N}\mathbf{x}_n^{\mathsf{T}}\right)^{\mathsf{T}}=\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n,\end{aligned} \tag{2.37}$$

since the transpose of the scalar value is the same as the original scalar value ($a^{\mathsf{T}}=a$) and $\mathbf{R}$ is a symmetric matrix ($\mathbf{R}^{\mathsf{T}}=\mathbf{R}$). Thus, by substituting Eqs. (2.32) and (2.36) into Eq. (2.31), Eq. (2.31) is rewritten as

$$\begin{aligned}&\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{R}(\mathbf{x}_n-\boldsymbol{\mu})\\&\quad=-2\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n+\mathrm{tr}\left[\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\right]+N\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\boldsymbol{\mu}.\end{aligned} \tag{2.38}$$

Note that Eq. (2.38) is a useful form, and it is used in the following sections to calculate the various equations for the multivariate Gaussian distribution.

Therefore, by substituting Eq. (2.38) into Eq. (2.30), we can obtain the exponential form of the multivariate Gaussian distribution as follows:

$$\begin{aligned}&\prod_{n=1}^{N}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu},\mathbf{R}^{-1})\\&\quad=(2\pi)^{-\frac{ND}{2}}|\mathbf{R}|^{\frac{N}{2}}\exp\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n-\frac{1}{2}\mathrm{tr}\left[\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\right]-\frac{N}{2}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\boldsymbol{\mu}\right)\\&\quad=\exp\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n-\frac{1}{2}\mathrm{tr}\left[\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}\right]-\frac{N}{2}\left(\log((2\pi)^D|\mathbf{R}|^{-1})+\boldsymbol{\mu}^{\mathsf{T}}\mathbf{R}\boldsymbol{\mu}\right)\right).\end{aligned} \tag{2.39}$$

Thus, by comparing with Eq. (2.23), we obtain the following parameterization for the multivariate Gaussian distribution:

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

$$\begin{cases} \mathbf{t}_1(\mathbf{X}) = \sum_{n=1}^{N} \mathbf{x}_n \\ \mathbf{T}_2(\mathbf{X}) = -\frac{1}{2} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\intercal} \\ h(x) = 1 \\ \boldsymbol{\gamma}_1(\Theta) = \mathbf{R}\boldsymbol{\mu} \\ \boldsymbol{\Gamma}_2(\Theta) = \mathbf{R} \\ g(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2) = \frac{N}{2} \left(\log((2\pi)^D |\mathbf{R}|^{-1}) + \boldsymbol{\mu}^{\intercal} \mathbf{R} \boldsymbol{\mu} \right) \\ \qquad\qquad = \frac{N}{2} \left(\log((2\pi)^D |\boldsymbol{\Gamma}_2|^{-1}) + \boldsymbol{\gamma}_1^{\intercal} \boldsymbol{\Gamma}_2^{-1} \boldsymbol{\gamma}_1 \right). \end{cases} \tag{2.40}$$

Note that if $D \to 1$, we have $\mathbf{x}_n \to x_n$, $\boldsymbol{\mu} \to \mu$, $\mathbf{R} \to r$, and Eq. (2.40) is equivalent to Eq. (2.29).

Example 2.4 Multinomial distribution:
The standard form of the multinomial distribution (Eq. (C.2)) is represented as follows:

$$\text{Mult}(x_1, \cdots, x_J | \omega_1, \cdots, \omega_J) \triangleq \frac{N!}{\prod_{j=1}^{J} x_j!} \prod_{j=1}^{J} \omega_j^{x_j}, \tag{2.41}$$

where x_j is a non-negative integer, and

$$\sum_{j=1}^{J} x_j = N. \tag{2.42}$$

The parameter $\{\omega_1, \cdots, \omega_J\}$ has the following constraint:

$$\sum_{j=1}^{J} \omega_j = 1, \quad 0 \le \omega_j \le 1 \quad \forall j. \tag{2.43}$$

Therefore, the number of the free parameters is $J-1$. To deal with the constraint, we first consider the $\{\omega_1, \cdots, \omega_{J-1}\}$ as the target vector parameters, i.e., $\boldsymbol{\theta} \triangleq [\omega_1, \cdots, \omega_{J-1}]^{\intercal}$. ω_J is represented by

$$\omega_J = 1 - \sum_{j=1}^{J-1} \omega_j. \tag{2.44}$$

Similarly to the previous Gaussian-based distributions, the multinomial distribution is also represented as the exponential form as follows:

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

$$\begin{aligned}\mathrm{Mult}(x_1,\cdots,x_J|\omega_1,\cdots,\omega_J) &= \frac{N!}{\prod_{j=1}^{J} x_j!}\exp\left(\log\left(\prod_{j=1}^{J}\omega_j^{x_j}\right)\right)\\ &= \frac{N!}{\prod_{j=1}^{J} x_j!}\exp\left(\sum_{j=1}^{J} x_j\log\omega_j\right).\end{aligned} \tag{2.45}$$

By using Eqs. (2.42) and (2.44) for x_J and ω_J, respectively, the exponential part of Eq. (2.45) is rewritten as

$$\begin{aligned}&\mathrm{Mult}(x_1,\cdots,x_J|\omega_1,\cdots,\omega_J)\\ &\propto \exp\left(\sum_{j=1}^{J-1} x_j\log\omega_j + \left(N-\sum_{j=1}^{J-1}x_j\right)\log\left(1-\sum_{j=1}^{J-1}\omega_j\right)\right)\\ &= \exp\left(\sum_{j=1}^{J-1} x_j\log\omega_j - \sum_{j=1}^{J-1}x_j\log\left(1-\sum_{j=1}^{J-1}\omega_j\right) + N\log\left(1-\sum_{j=1}^{J-1}\omega_j\right)\right)\\ &= \exp\left(\underbrace{\sum_{j=1}^{J-1} x_j\log\frac{\omega_j}{1-\sum_{j=1}^{J-1}\omega_j}}_{\triangleq \mathbf{x}^{\mathsf{T}}\boldsymbol{\gamma}} + N\log\left(1-\sum_{j'=1}^{J-1}\omega_{j'}\right)\right),\end{aligned} \tag{2.46}$$

where $\propto$ denotes the proportional relation between left- and right-hand-side equations. Since the probabilistic function has the normalization factor, which can be neglected for most of the calculations, $\propto$ is often used to omit the normalization constant from the equations. Thus, we can derive the linear relationship between x_j and γ_j, which is defined with $\{\omega_j\}_{j=1}^{J-1}$ as follows:

$$\gamma_j \triangleq \log\frac{\omega_j}{1-\sum_{j'=1}^{J-1}\omega_{j'}}. \tag{2.47}$$

Note that ω_j is represented by γ_j by using the following equation:

$$\omega_j = \frac{\exp(\gamma_j)}{1+\sum_{j'=1}^{J-1}\exp(\gamma_{j'})}. \tag{2.48}$$

This is confirmed by substituting Eq. (2.47) into Eq. (2.48) as

$$\begin{aligned}\frac{\exp(\gamma_j)}{1+\sum_{j'=1}^{J-1}\gamma_{j'}} &= \frac{\frac{\omega_j}{1-\sum_{j'=1}^{J-1}\omega_{j'}}}{1+\sum_{j'=1}^{J-1}\frac{\omega_{j'}}{1-\sum_{j''=1}^{J-1}\omega_{j''}}}\\ &= \frac{\omega_j}{1-\sum_{j'=1}^{J-1}\omega_{j'}+\sum_{j'=1}^{J-1}\omega_{j'}} = \omega_j.\end{aligned} \tag{2.49}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Therefore, the canonical form of the multinomial distribution with the parameter $\boldsymbol{\theta} = [\omega_1, \cdots, \omega_{J-1}]^\mathsf{T}$ for $\mathbf{x} = [x_1, \cdots, x_{J-1}]^\mathsf{T}$ is represented as

$$\begin{cases} \mathbf{t}(\mathbf{x}) = \mathbf{x} \\ h(x) = \dfrac{N!}{\prod_{j=1}^{J} x_j!} \\ \boldsymbol{\gamma}(\boldsymbol{\theta}) = \left[\log \frac{\omega_1}{1-\sum_{j=1}^{J-1}\omega_j}, \cdots, \log \frac{\omega_{J-1}}{1-\sum_{j=1}^{J-1}\omega_j}\right]^\mathsf{T} \\ g(\boldsymbol{\gamma}) = -N \log \left(1 - \sum_{j=1}^{J-1} \omega_j\right) \\ \qquad = -N \log \left(1 - \sum_{j=1}^{J-1} \dfrac{\exp(\gamma_j)}{1 + \sum_{j'=1}^{J-1} \exp(\gamma_{j'})}\right) \\ \qquad = N \log \left(1 + \sum_{j=1}^{J-1} \exp(\gamma_j)\right). \end{cases} \tag{2.50}$$

Note that since the multinomial distribution has constraints for the observation x_j in Eq. (2.42) and the parameter ω_j in Eq. (2.44), the obtained canonical form of the multinomial distribution involves these constraints with $J-1$ variables for sufficient statistics $\mathbf{t}$ and the transformed vector $\boldsymbol{\gamma}$.

The obtained exponential family forms for Gaussian, multivariate Gaussian, and multinomial distributions are often used in the Bayesian treatment of statistical models in speech and language processing.

2.1.4 Conjugate distributions

The previous section introduces the exponential family distributions and provides some examples of these distributions. Based on the exponential family distributions, this section explains how to obtain the posterior distributions when we use the exponential family distributions as the likelihood functions. For such a distribution, we can find a nice property to obtain the posterior distribution analytically if we set a particular type of distribution.

Let $p(\mathbf{X}|\boldsymbol{\theta})$ be a likelihood function for a set of observation vectors $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$. We first start the discussion from the simple case that the parameters are represented as a vector form, i.e., $\boldsymbol{\theta}$. An exponential family distribution of $p(\mathbf{X}|\boldsymbol{\theta})$ is defined in Eq. (2.20) as

$$p(\mathbf{X}|\boldsymbol{\theta}) = h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}^\mathsf{T}\mathbf{t}(\mathbf{X}) - g(\boldsymbol{\gamma})\right). \tag{2.51}$$

Here use $\boldsymbol{\gamma}(\boldsymbol{\theta}) \rightarrow \boldsymbol{\gamma}$ for simplicity. Then, we use the following Bayes theorem for $\boldsymbol{\theta}$ based on Eq. (2.18) to calculate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$:

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.52}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

where we disregard the normalization factor $p(\mathbf{X})$. For this calculation, we need to prepare a prior distribution $p(\boldsymbol{\theta})$. Instead of considering the prior distribution of $p(\boldsymbol{\theta})$, we consider the prior distribution of $p(\boldsymbol{\gamma})$. We set a prior distribution for $p(\boldsymbol{\gamma})$, which is parameterized with additional variables $\boldsymbol{\nu}$ and ϕ, where the parameters of prior and posterior distributions are called *hyperparameters*. The hyperparameter appearing in this book is often used as the parameter of prior or posterior distributions. Then, the prior distribution is proportional to the following function form:

$$p(\boldsymbol{\theta}) \to p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \phi) \propto \exp\left(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\nu} - \phi g(\boldsymbol{\gamma})\right). \tag{2.53}$$

Here, $g(\boldsymbol{\gamma})$ is introduced in Eqs. (2.20) and (2.51) as a logarithmic normalization factor of the likelihood function. This form of prior distribution is called *conjugate prior* distribution.

We can calculate the posterior distribution of $p(\boldsymbol{\theta}|\mathbf{X})$ via $\boldsymbol{\gamma}$ by substituting Eqs. (2.51) and (2.53) into Eq. (2.52):

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}) &\to p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \phi) \\ &= h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}^\mathsf{T}\mathbf{t}(\mathbf{X}) - g(\boldsymbol{\gamma})\right) \exp\left(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\nu} - \phi g(\boldsymbol{\gamma})\right) \\ &\propto \exp\left(\boldsymbol{\gamma}^\mathsf{T}(\boldsymbol{\nu} + \mathbf{t}(\mathbf{X})) - (\phi + 1)g(\boldsymbol{\gamma})\right) \\ &= p(\boldsymbol{\gamma}|\boldsymbol{\nu} + \mathbf{t}(\mathbf{X}), \phi + 1), \end{aligned} \tag{2.54}$$

where we use the definition used in the conjugate prior distribution (Eq. (2.53)). This solution means that the conjugate posterior distribution is analytically obtained with the same distribution function as the conjugate prior distribution by just using the simple rule of changing hyperparameters from $(\boldsymbol{\nu}, \phi)$ to $(\boldsymbol{\nu} + \mathbf{t}(\mathbf{X}), \phi + 1)$.

Note that the setting of ϕ is not unique. We consider the case that $g(\boldsymbol{\gamma})$ is decomposed into M functions, i.e.,

$$g(\boldsymbol{\gamma}) \triangleq \sum_{m=1}^{M} g_m(\boldsymbol{\gamma}). \tag{2.55}$$

Then, similarly to Eq. (2.53), we can provide M hyperparameters for a prior distribution as follows:

$$p(\boldsymbol{\theta}) \to p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \phi) \propto \exp\left(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\nu} - \sum_{m=1}^{M} \phi_m g_m(\boldsymbol{\gamma})\right). \tag{2.56}$$

The corresponding posterior distribution is similarly derived by substituting Eqs. (2.51), (2.55), and (2.56) into Eq. (2.52) as:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}) &\to p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \{\phi_m\}_{m=1}^M) \\ &= h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}^\mathsf{T}\mathbf{t}(\mathbf{X}) - \sum_{m=1}^{M} g_m(\boldsymbol{\gamma})\right) \exp\left(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\nu} - \sum_{m=1}^{M} \phi_m g_m(\boldsymbol{\gamma})\right) \\ &\propto \exp\left(\boldsymbol{\gamma}^\mathsf{T}(\boldsymbol{\nu} + \mathbf{t}(\mathbf{X})) - \sum_{m=1}^{M} (\phi_m + 1) g_m(\boldsymbol{\gamma})\right) \\ &= p(\boldsymbol{\gamma}|\boldsymbol{\nu} + \mathbf{t}(\mathbf{X}), \{\phi_m + 1\}_{m=1}^M). \end{aligned} \tag{2.57}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Thus, we can derive the posterior distribution with M hyperparameters. The setting of $\{\phi_m\}$ is an additional flexibility of the prior distribution. If we use many $\{\phi_m\}$, we could precisely represent a prior distribution. However, by using a few $\{\phi_m\}$, we can easily control the shape of a prior distribution with a few free parameters.

If the transformed model parameters are composed of a vector $\boldsymbol{\gamma}_1$ and matrix $\boldsymbol{\Gamma}_2$, as discussed in Eq. (2.23), we also have similar result. A likelihood function of this exponential family distribution is represented by the following general form:

$$p(\mathbf{X}|\Theta) \triangleq h(\mathbf{X}) \exp\left(\boldsymbol{\gamma}_1^\intercal \mathbf{t}_1(\mathbf{X}) + \mathrm{tr}[\boldsymbol{\Gamma}_2^\intercal \mathbf{T}_2(\mathbf{x})] - \sum_{m=1}^{M} g_m(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2)\right). \tag{2.58}$$

Here, similarly to Eq. (2.55), we use the following equation for the $g(\cdot)$ function:

$$g(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2) \triangleq \sum_{m=1}^{M} g_m(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2). \tag{2.59}$$

Therefore, by providing the following prior distribution form as a conjugate prior with hyperparameters $\boldsymbol{\nu}_1, \mathbf{N}_2$ and $\{\phi_m\}_{m=1}^M$:

$$p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2|\boldsymbol{\nu}_1, \mathbf{N}_2, \{\phi\}_{m=1}^M) \propto \exp\left(\boldsymbol{\gamma}_1^\intercal \boldsymbol{\nu}_1 + \mathrm{tr}[\boldsymbol{\Gamma}_2^\intercal \mathbf{N}_2] - \sum_{m=1}^{M} \phi_m g_m(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2)\right). \tag{2.60}$$

We can calculate the posterior distribution by substituting Eqs. (2.60), (2.58), and (2.59) into Eq. (2.52):

$$\begin{aligned}
&p(\Theta|\mathbf{X}) \\
&\quad \to p(\mathbf{X}|\Theta) p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2|\mathbf{N}, \{\phi_m\}_{m=1}^M) \\
&\quad \propto \exp\left(\boldsymbol{\gamma}^\intercal(\boldsymbol{\nu}_1 + \mathbf{t}_1(\mathbf{X})) + \mathrm{tr}[\boldsymbol{\Gamma}_2^\intercal(\mathbf{N}_2 + \mathbf{T}_2(\mathbf{X}))] - \sum_{m=1}^{M}(\phi_m + 1) g_m(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2)\right) \\
&\quad = p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2|\boldsymbol{\nu}_1 + \mathbf{t}_1(\mathbf{X}), \mathbf{N}_2 + \mathbf{T}_2(\mathbf{X}), \{\phi_m + 1\}_{m=1}^M).
\end{aligned} \tag{2.61}$$

Here we use the distributive property of the trace in Appendix B that:

$$\mathrm{tr}[\mathbf{AB}] + \mathrm{tr}[\mathbf{AC}] = \mathrm{tr}[\mathbf{A}(\mathbf{B} + \mathbf{C})]. \tag{2.62}$$

Now, we summarize the conjugate prior and posterior distributions. The exponential family distributions with the vector form parameters $\boldsymbol{\theta}$ have the following relationship:

$$\begin{cases} \text{Prior: } p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \{\phi_m\}_{m=1}^M) \\ \text{Posterior: } p(\boldsymbol{\gamma}|\boldsymbol{\nu} + \mathbf{t}(\mathbf{X}), \{\phi_m + 1\}_{m=1}^M). \end{cases} \tag{2.63}$$

When the distribution has vector and matrix parameters, we have the following relationship:

$$\begin{cases} \text{Prior: } p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2|\boldsymbol{\nu}_1, \mathbf{N}_2, \{\phi_m\}_{m=1}^M) \\ \text{Posterior: } p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2|\boldsymbol{\nu}_1 + \mathbf{t}_1(\mathbf{X}), \mathbf{N}_2 + \mathbf{T}_2(\mathbf{X}), \{\phi_m + 1\}_{m=1}^M). \end{cases} \tag{2.64}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Therefore, the posterior distribution of the natural parameters $(\boldsymbol{\gamma}, \boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2)$ is analytically obtained by using Eqs. (2.63) and (2.64) as a rule. The posterior distribution of the original parameters $p(\Theta|\mathbf{X})$ is obtained by transforming the posterior distribution of the natural parameters.

The rest of this section provides examples of the conjugate prior and posterior distributions for some exponential family distributions.

Example 2.5 Conjugate distributions for Gaussian (unknown mean):
We first describe the case that we only consider a Gaussian mean parameter μ, and the precision parameter $r = \Sigma^{-1}$ is regarded as a constant value. Based on the discussion in Example 2.1, the canonical form of the Gaussian distribution is represented as follows:

$$\prod_{n=1}^{N} \mathcal{N}(x_n|\mu; r^{-1}) = h(X) \exp\left(\gamma t(X) - g(\gamma)\right), \tag{2.65}$$

where

$$\begin{cases} t(X) = \sum_{n=1}^{N} x_n \\ h(X) = \left(\frac{2\pi}{r}\right)^{-\frac{N}{2}} \exp\left(-\frac{r\sum_{n=1}^{N} x_n^2}{2}\right) \\ \gamma = \mu r \\ g(\gamma) = \frac{N\gamma^2}{2r}. \end{cases} \tag{2.66}$$

Therefore, by substituting γ and $g(\gamma)$ in Eq. (2.66) into the general form of the conjugate distribution in Eq. (2.53), we can derive the function of mean μ as follows:

$$\begin{aligned} p(\gamma|\nu, \phi) &\propto \exp\left(\gamma\nu - \phi g(\gamma)\right) = \exp\left(\mu r\nu - \phi\frac{N\mu^2 r}{2}\right) \\ &\propto \exp\left(-\frac{N\phi r}{2}\left(\mu - \frac{\nu}{N\phi}\right)^2\right). \end{aligned} \tag{2.67}$$

Thus, the prior distribution of μ is represented by a Gaussian distribution with $\frac{\nu}{N\phi}$ and $N\phi r$ as the mean and precision parameters, respectively:

$$p(\mu) \propto \mathcal{N}\left(\mu \left| \frac{\nu}{N\phi}, (N\phi r)^{-1}\right.\right). \tag{2.68}$$

Based on the conjugate distribution rule (Eq. (2.63)), the posterior distribution is easily solved by just replacing $\nu \to \nu + t(X)$ and $\phi \to \phi + 1$ in Eq. (2.68) without complex calculations:

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

$$p\left(\gamma \,|\nu + t(X), \phi + 1\right) \propto \exp\left(-\frac{N(\phi+1)r}{2}\left(\mu - \frac{\nu + \sum_{n=1}^{N} x_n}{N(\phi+1)}\right)^2\right)$$
$$\rightarrow \mathcal{N}\left(\mu \left| \frac{\nu + \sum_{n=1}^{N} x_n}{N(\phi+1)}, (N(\phi+1)r)^{-1}\right.\right). \tag{2.69}$$

Therefore, similarly to the prior distribution, the posterior distribution of μ is represented by a Gaussian distribution with $\frac{\nu+\sum_{n=1}^{N} x_n}{N(\phi+1)}$ and $(N(\phi+1)r)^{-1}$ as the mean and variance parameters, respectively:

$$p(\mu|X) \propto \mathcal{N}\left(\mu \left| \frac{\nu + \sum_{n=1}^{N} x_n}{N(\phi+1)}, (N(\phi+1)r)^{-1}\right.\right). \tag{2.70}$$

Thus, both prior and posterior distributions are represented in the same form as a Gaussian distribution with different parameters.

Now we consider the meaning of the solution of Eqs. (2.68) and (2.70). We parameterize the ϕ and ν by newly introducing the following parameters:

$$\begin{aligned} \phi &\triangleq \frac{\phi^{\mu}}{N} \\ \nu &\triangleq \phi^{\mu}\mu^{0}. \end{aligned} \tag{2.71}$$

Then, the prior and posterior distributions of μ in Eqs. (2.68) and (2.70) are rewritten as:

$$\begin{cases} p(\mu) = \mathcal{N}\left(\mu \left| \mu^{0}, (\phi^{0}r)^{-1}\right.\right) \\ p(\mu|X) = \mathcal{N}\left(\mu \left| \hat{\mu}, (\hat{\phi}^{\mu}r)^{-1}\right.\right). \end{cases} \tag{2.72}$$

where

$$\begin{aligned} \hat{\phi}^{\mu} &\triangleq \phi^{\mu} + N \\ \hat{\mu} &\triangleq \frac{\phi^{\mu}\mu^{0} + \sum_{n=1}^{N} x_n}{\phi^{\mu} + N}. \end{aligned} \tag{2.73}$$

These are famous Bayesian solutions of the posterior distribution of the Gaussian mean. We can consider the two extreme cases that the amount of data is zero or very large. Then, the posterior distribution is represented as:

- $N \rightarrow 0$

$$\lim_{N\rightarrow 0} p(\mu|X) = \mathcal{N}\left(\mu \left| \mu^{0}, (\phi^{\mu}r)^{-1}\right.\right) = p(\mu). \tag{2.74}$$

This solution means that we only use the prior information when we don't have data.

- $N \gg 1$

$$\lim_{N\rightarrow\infty} p(\mu|X) \approx \lim_{N\rightarrow\infty} \mathcal{N}\left(\mu \left| \frac{\sum_{n=1}^{N} x_n}{N}, \frac{1}{Nr}\right.\right) \rightarrow \delta(\mu - \mu^{\mathrm{ML}}), \tag{2.75}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

where μ^{ML} is the ML estimate of μ, and the posterior distribution is close to the ML value with small standard deviation, which is similar to the delta function that has a peak value at the ML estimate.

Thus, the solution of Eq. (2.72) approaches the delta function with the ML estimate when the amount of data is very large and approaches the prior distribution when the amount of data is very small. The mean parameter of the posterior distribution,

$$\frac{\phi^{\mu}\mu^{0} + \sum_{n=1}^{N} x_n}{\phi^{\mu} + N}, \tag{2.76}$$

is interpolated by the prior mean parameter μ^0 and the ML estimate, and ϕ^{μ} can control an interpolation ratio.

Example 2.6 Conjugate distributions for Gaussian (unknown mean and precision): Similarly to Example 2.5, we first rewrite a Gaussian distribution. In this situation, the set of the parameters is $\theta = \{\mu, r\}$. From Eq. (2.29), the Gaussian distribution with precision r has the following exponential form:

$$\begin{aligned}\prod_{n=1}^{N} \mathcal{N}(x_n|\mu, r^{-1}) &= h(X)\exp\left(\boldsymbol{\gamma}\mathbf{t}(X)^{\mathsf{T}} - g(\boldsymbol{\gamma})\right) \\ &= h(X)\exp\left(\boldsymbol{\gamma}\mathbf{t}(X)^{\mathsf{T}} - \phi_1 g_1(\boldsymbol{\gamma}) - \phi_2 g_2(\boldsymbol{\gamma})\right),\end{aligned} \tag{2.77}$$

where we introduce ϕ_1 and ϕ_2 that are discussed in Eq. (2.55). The variables in the above equations are represented as follows:

$$\begin{cases} \mathbf{t}(X) = \begin{bmatrix} \sum_{n=1}^{N} x_n \\ -\frac{\sum_{n=1}^{N} x_n^2}{2} \end{bmatrix} \\ h(X) = 1 \\ \boldsymbol{\gamma} = \begin{bmatrix} \mu r \\ r \end{bmatrix} \\ g_1(\boldsymbol{\gamma}) = \dfrac{N}{2}\left(\dfrac{\gamma_1^2}{\gamma_2}\right) \\ g_2(\boldsymbol{\gamma}) = \dfrac{N}{2}\left(\log\dfrac{2\pi}{\gamma_2}\right). \end{cases} \tag{2.78}$$

Therefore, by substituting $\boldsymbol{\gamma}$, $g_1(\boldsymbol{\gamma})$, and $g_2(\boldsymbol{\gamma})$ in Eq. (2.78) into the general form of the conjugate distribution in Eq. (2.56), we can derive the function of mean μ and precision r as follows:

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\nu}, \phi_1, \phi_2) &\propto \exp\left(\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\nu} - \phi_1 g_1(\boldsymbol{\gamma}) - \phi_2 g_2(\boldsymbol{\gamma})\right) \\ &\propto \exp\left([\mu r, r]\begin{bmatrix}\nu_1 \\ \nu_2\end{bmatrix} - \frac{N\phi_1}{2} r\mu^2 - \frac{N\phi_2}{2}\log\left(\frac{2\pi}{r}\right)\right) \\ &\propto r^{\frac{N\phi_2}{2}} \exp\left(\nu_1 r\mu - N\phi_1\frac{r\mu^2}{2} + r\nu_2\right), \end{aligned} \tag{2.79}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

where we omit the factor that does not depend on r and μ. By making a complete square form of μ, we can obtain a Gaussian distribution of μ with mean $\frac{\nu_1}{N\phi_1}$ and precision $N\phi_1 r$:

$$\begin{aligned}
&p(\boldsymbol{\gamma}|\boldsymbol{v},\phi_1,\phi_2) \\
&\quad\propto r^{\frac{N\phi_2}{2}} \exp\left(-\frac{N\phi_1 r}{2}\left(\mu-\frac{\nu_1}{N\phi_1}\right)^2+\frac{r\nu_1^2}{2N\phi_1}+r\nu_2\right) \\
&\quad= r^{\frac{N\phi_2}{2}}\left(\frac{2\pi}{N\phi_1 r}\right)^{\frac{1}{2}}\left(\frac{2\pi}{N\phi_1 r}\right)^{-\frac{1}{2}} \exp\left(-\frac{N\phi_1 r}{2}\left(\mu-\frac{\nu_1}{N\phi_1}\right)^2+\frac{r\nu_1^2}{2N\phi_1}+r\nu_2\right) \\
&\quad\propto \mathcal{N}\left(\mu\left|\frac{\nu_1}{N\phi_1},(N\phi_1 r)^{-1}\right.\right)\underbrace{r^{\frac{N\phi_2}{2}} r^{-\frac{1}{2}} \exp\left(r\frac{\nu_1^2}{2N\phi_1}+r\nu_2\right)}_{\triangleq(*1)}.
\end{aligned} \tag{2.80}$$

Now we consider the rest of the exponential factor $(*1)$. By focusing on r and using the definition of a gamma distribution (Appendix C.11), the factor is rewritten as follows:

$$\begin{aligned}
(*1) &\propto r^{\frac{N\phi_2+1}{2}-1} \exp\left(-\left(-\frac{\nu_1^2}{2N\phi_1}-\nu_2\right)r\right) \\
&\propto \mathrm{Gam}\left(r\left|\frac{N\phi_2+1}{2},-\frac{\nu_1^2}{2N\phi_1}-\nu_2\right.\right),
\end{aligned} \tag{2.81}$$

where the definition of a gamma distribution is as follows:

$$\mathrm{Gam}(r|\alpha,\beta) \triangleq \frac{1}{\Gamma(\alpha)}\beta^\alpha r^{\alpha-1}\exp(-\beta r), \tag{2.82}$$

where $\Gamma(\cdot)$ is a Gamma function (Appendix A.4). Thus, precision $r=\frac{1}{\Sigma}$ is represented by a gamma distribution with $\frac{N\phi_2+1}{2}$ and $-\frac{\nu_1^2}{2N\phi_1}-\nu_2$ as parameters.

This representation can be simplified by using the following definition for the other definition of 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):

$$\begin{aligned}
\mathrm{Gam}_2(y|\phi,r^0) &\triangleq \mathrm{Gam}\left(y\left|\frac{\phi}{2},\frac{r^0}{2}\right.\right) \\
&\propto y^{\frac{\phi}{2}-1}\exp\left(-\frac{r^0 y}{2}\right).
\end{aligned} \tag{2.83}$$

Equation (2.81) is rewritten as

$$(*1) \propto \mathrm{Gam}_2\left(r\left|N\phi_2+1,-\frac{\nu_1^2}{N\phi_1}-2\nu_2\right.\right). \tag{2.84}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Thus, the conjugate prior distribution is represented as the product form of the following Gaussian and gamma distributions:

$$p(\mu, r) = \mathcal{N}\left(\mu \left| \frac{\nu_1}{N\phi_1}, (N\phi_1 r)^{-1} \right.\right) \mathrm{Gam}_2\left(r \left| N\phi_2 + 1, -\frac{\nu_1^2}{N\phi_1} - 2\nu_2 \right.\right). \tag{2.85}$$

This can be also represented as a Gaussian-gamma distribution (or so-called normal-gamma) defined in Appendix C.13, as follows:

$$p(\mu, r) = \mathcal{N}\mathrm{Gam}\left(\mu, r \left| \frac{\nu_1}{N\phi_1}, (N\phi_1 r)^{-1}, -\frac{\nu_1^2}{N\phi_1} - 2\nu_2, N\phi_2 + 1 \right.\right). \tag{2.86}$$

The Gaussian-gamma distribution is a conjugate prior distribution of the joint variable μ and r.

Similarly to the previous example, we introduce the following new parameters:

$$\begin{cases} \phi^\mu \triangleq N\phi_1 \\ \mu^0 \triangleq \dfrac{\nu_1}{N\phi_1} \\ \phi^r \triangleq N\phi_2 + 1 \\ r^0 \triangleq -\dfrac{\nu_1^2}{N\phi_1} - 2\nu_2. \end{cases} \tag{2.87}$$

By using Eq. (2.87), the conjugate prior distribution of Eq. (2.85) is rewritten by using these new parameters as follows:

$$p(\mu, r) = \mathcal{N}\left(\mu \left| \mu^0, (\phi^\mu r)^{-1} \right.\right) \mathrm{Gam}_2\left(r \left| \phi^r, r^0 \right.\right). \tag{2.88}$$

Note that we can also use Gaussian-gamma distribution as:

$$\begin{aligned} p(\mu, r) &= \mathcal{N}(\mu | \mu^0, (r\phi^\mu)^{-1}) \, \mathrm{Gam}_2\left(r \left| \phi^r, r^0 \right.\right) \\ &= \mathcal{N}\mathrm{Gam}(\mu, r | \mu^0, \phi^\mu, r^0, \phi^r). \end{aligned} \tag{2.89}$$

Thus, we can derive the prior distribution of joint variable μ and r as the product of the Gaussian and gamma distributions in Eq. (2.88), or the single Gaussian-gamma distribution in Eq. (2.89).

Now, we focus on the posterior distribution of μ and r. Based on the conjugate distribution theory, the posterior distribution is represented as the same form of the Gaussian-gamma distribution as the prior distribution (2.89) with hyperparameters $\hat{\phi}^\mu, \hat{\mu}, \hat{\phi}^r$, and $\hat{r}$ as follows:

$$p(\mu, r | X) = \mathcal{N}\mathrm{Gam}(\mu, r | \hat{\mu}^0, \hat{\phi}^\mu, \hat{r}^0, \hat{\phi}^r). \tag{2.90}$$

Based on the conjugate distribution rule (Eq. (2.63)), the hyperparameters of the posterior distribution are easily solved by just replacing $\boldsymbol{\nu} \to \boldsymbol{\nu} + \mathbf{t}(X)$ and $\phi_m \to \phi_m + 1$ in Eq. (2.87) without complex calculations, as follows:

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

$$
\begin{aligned}
\hat{\phi}^{\mu} &= N(\phi_1 + 1) = \phi^{\mu} + N \\
\hat{\mu} &= \frac{\nu_1 + \sum_{n=1}^{N} x_n}{N(\phi_1 + 1)} = \frac{\phi^{\mu}\mu^0 + \sum_{n=1}^{N} x_n}{\phi^{\mu} + N} \\
\hat{\phi}^{r} &= N(\phi_2 + 1 + 1) + 1 = N(\phi_2 + 1) + 1 + N \\
&= \phi^{r} + N \\
\hat{r} &= -\frac{\left(\nu_1 + \sum_{n=1}^{N} x_n\right)^2}{N(\phi_1 + 1)} - \left(2\nu_2 - \sum_{n=1}^{N} x_n^2\right) \\
&= -\frac{\left(\phi^{\mu}\mu^0 + \sum_{n=1}^{N} x_n\right)^2}{\phi^{\mu} + N} - \left(-\frac{\nu_1^2}{N\phi_1} - r^0 - \sum_{n=1}^{N} x_n^2\right) \\
&= -\frac{\left(\phi^{\mu}\mu^0 + \sum_{n=1}^{N} x_n\right)^2}{\phi^{\mu} + N} + \phi^{\mu}(\mu^0)^2 + r^0 + \sum_{n=1}^{N} x_n^2.
\end{aligned} \tag{2.91}
$$

Thus, we summarize the result of the hyperparameters of the conjugate posterior distribution as

$$
\begin{cases}
\hat{\phi}^{\mu} = \phi^{\mu} + N \\
\hat{\mu} = \dfrac{\phi^{\mu}\mu^0 + \sum_{n=1}^{N} x_n}{\phi^{\mu} + N} \\
\hat{\phi}^{r} = \phi^{r} + N \\
\hat{r} = -\hat{\phi}^{\mu}(\hat{\mu})^2 + \phi^{\mu}(\mu^0)^2 + r^0 + \sum_{n=1}^{N} x_n^2.
\end{cases} \tag{2.92}
$$

Note that in this representation, the posterior distribution parameters of $\hat{\phi}^{\mu}$ and $\hat{\phi}^{r}$ are obtained by simply adding the number of observations N to the prior distribution parameters of ϕ^{μ} and ϕ^{r}, respectively.

Finally, we summarize the result. The prior and posterior distributions of μ and r in Eqs. (2.89) and (2.90) are also summarized as:

$$
\begin{cases}
p(\mu, r) = \mathcal{NG}\text{am}(\mu, r|\mu^0, \phi^{\mu}, r^0, \phi^{r}) \\
p(\mu, r|X) = \mathcal{NG}\text{am}(\mu, r|\hat{\mu}, \hat{\phi}^{\mu}, \hat{r}, \hat{\phi}^{r}),
\end{cases} \tag{2.93}
$$

or

$$
\begin{cases}
p(\mu, r) = p(\mu|r)p(r) = \mathcal{N}(\mu|\mu^0, (\phi^{\mu} r)^{-1})\ \text{Gam}_2(r|\phi^{r}, r^0) \\
p(\mu, r|X) = p(\mu|r, X)p(r|X) = \mathcal{N}(\mu|\hat{\mu}\ (\hat{\phi}^{\mu} r)^{-1})\ \text{Gam}_2(r|\hat{\phi}^{r}, \hat{r}).
\end{cases} \tag{2.94}
$$

Similarly to the discussion about the mean parameter μ in Example 2.5, we can consider the two extreme cases, that the amount of data is zero or very large, for the behavior of the precision parameter solution r. The posterior distribution of r is represented as:

- $N \to 0$

$$
\lim_{N \to 0} p(r|X) = \text{Gam}_2(r|\phi^{r}, r^0) = p(r). \tag{2.95}
$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

This solution means that we only use the prior information when we don't have data.

- $N \gg 1$

$$\lim_{N\gg 1} p(r|X) \approx \mathrm{Gam}_2\left(r \middle| N, \sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N}\right). \tag{2.96}$$

Since the mean of the gamma distribution r^{Mean} (with $\frac{1}{2}$ factor) is defined in Eq. (C.83), the mean of r in this limit is represented as:

$$\begin{aligned} r^{\mathrm{Mean}} &= \frac{N}{\sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N}} \\ &= \left(\frac{\sum_{n=1}^{N} x_n^2}{N} - \left(\frac{\sum_{n=1}^{N} x_n}{N}\right)^2\right)^{-1}. \end{aligned} \tag{2.97}$$

This is equivalent to the maximum likelihood estimation of r^{ML} represented as follows:

$$r^{\mathrm{ML}} = \left(\mathrm{Mean}[x^2] - (\mathrm{Mean}[x])^2\right)^{-1} = r^{\mathrm{Mean}}. \tag{2.98}$$

Thus, the mean of the posterior distribution approaches the ML estimate of r when the amount of data is large.

Similarly, based on the definition of the variance of the gamma distribution (with $\frac{1}{2}$ factor) in Eq. (C.84), the variance is also represented as

$$\begin{aligned} r^{\mathrm{Variance}} &= \frac{2N}{\left(\sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N}\right)^2} \\ &= \frac{\frac{2}{N}}{\left(\frac{\sum_{n=1}^{N} x_n^2}{N} - \left(\frac{\sum_{n=1}^{N} x_n}{N}\right)^2\right)^2} \\ &= \frac{2\left(r^{\mathrm{ML}}\right)^2}{N} \approx 0. \end{aligned} \tag{2.99}$$

Note that the order of r^{ML} in Eq. (2.98) is a constant for N, and the variance of the precision parameter r^{Variance} approaches 0, i.e., the posterior distribution of r has a strong peak at r^{ML} with a very small variance. Therefore, the posterior distribution of precision parameter $p(r|X)$ in the case of a large amount of data can be approximated as the following Dirac delta function with the ML estimate:

$$\lim_{N\gg 1} p(r|X) \approx \delta(r - r^{\mathrm{ML}}). \tag{2.100}$$

This conclusion is similar to the case of the large amount limitation of the posterior distribution of mean parameter $p(\mu|X)$ in Eq. (2.75).

This Gaussian-gamma distribution is used to model the prior and posterior distributions of Gaussian parameters (μ and r) for scalar continuous observations, or can be used for vector continuous observations when we use a diagonal covariance matrix.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Example 2.7 Conjugate distributions for multivariate Gaussian (unknown mean vector and precision matrix):
Based on the discussion of Eq. (2.39) in Example 2.3, the canonical form of the multivariate Gaussian distribution is represented as follows:

$$\prod_{n=1}^{N}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu},\mathbf{R}^{-1}) \propto \exp\left(\boldsymbol{\mu}^{\intercal}\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n-\frac{1}{2}\mathrm{tr}\left[\mathbf{R}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\intercal}\right]-\frac{N}{2}\left(\log(2\pi|\mathbf{R}|^{-1})+\boldsymbol{\mu}^{\intercal}\mathbf{R}\boldsymbol{\mu}\right)\right), \tag{2.101}$$

where

$$\begin{cases} \mathbf{t}_1(\mathbf{X})=\sum_{n=1}^{N}\mathbf{x}_n \\ \mathbf{T}_2(\mathbf{X})=-\frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\intercal} \\ h(x)=1 \\ \boldsymbol{\gamma}_1(\Theta)=\mathbf{R}\boldsymbol{\mu} \\ \boldsymbol{\Gamma}_2(\Theta)=\mathbf{R} \\ g_1(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)=\frac{N}{2}\boldsymbol{\gamma}_1^{\intercal}\boldsymbol{\Gamma}_2^{-1}\boldsymbol{\gamma}_1 \\ g_2(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)=\frac{N}{2}\log(2\pi|\boldsymbol{\Gamma}_2|^{-1}). \end{cases} \tag{2.102}$$

Therefore, by substituting $\boldsymbol{\gamma}$, $g_1(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)$, and $g_2(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)$ in Eq. (2.102) into the general form of the conjugate distribution in Eq. (2.60), we can derive the function of mean $\boldsymbol{\mu}$ and $\boldsymbol{r}$ as follows:

$$\begin{aligned} &p(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2|\boldsymbol{v}_1,\mathbf{N}_2,\phi_1,\phi_2) \\ &\quad\propto \exp\left(\boldsymbol{\gamma}_1^{\intercal}\boldsymbol{v}_1+\mathrm{tr}[\boldsymbol{\Gamma}_2^{\intercal}\mathbf{N}_2]-\phi_1 g_1(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)-\phi_2 g_2(\boldsymbol{\gamma}_1,\boldsymbol{\Gamma}_2)\right) \\ &\quad\propto \exp\left(\boldsymbol{\mu}^{\intercal}\mathbf{R}\boldsymbol{v}_1+\mathrm{tr}[\mathbf{R}^{\intercal}\mathbf{N}_2]-\frac{N\phi_1}{2}\boldsymbol{\gamma}_1^{\intercal}\boldsymbol{\Gamma}_2^{-1}\boldsymbol{\gamma}_1-\frac{N\phi_2}{2}\log(2\pi|\boldsymbol{\Gamma}_2|^{-1})\right) \\ &\quad\propto \exp\left(\boldsymbol{\mu}^{\intercal}\mathbf{R}\boldsymbol{v}_1+\mathrm{tr}[\mathbf{R}\mathbf{N}_2]-\frac{N\phi_1}{2}\boldsymbol{\mu}^{\intercal}\mathbf{R}\boldsymbol{\mu}-\frac{N\phi_2}{2}\log(|\mathbf{R}|^{-1})\right), \end{aligned} \tag{2.103}$$

where we omit the factor that does not depend on $\mathbf{R}$ and $\boldsymbol{\mu}$. Similarly to Example 2.6, we first use a complete square form of $\boldsymbol{\mu}$ to derive a Gaussian distribution from Eq. (2.103).

In Appendix B.4, we have the following formula for the complete square form of vectors:

$$\mathbf{x}^{\intercal}\mathbf{A}\mathbf{x}-2\mathbf{x}^{\intercal}\mathbf{b}+c=(\mathbf{x}-\mathbf{u})^{\intercal}\mathbf{A}(\mathbf{x}-\mathbf{u})+v, \tag{2.104}$$

where

$$\begin{aligned} \mathbf{u}&\triangleq\mathbf{A}^{-1}\mathbf{b} \\ v&\triangleq c-\mathbf{b}^{\intercal}\mathbf{A}^{-1}\mathbf{b}. \end{aligned} \tag{2.105}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Therefore, by $\mathbf{x} \to \boldsymbol{\mu}$, $\mathbf{A} \to N\phi_1\mathbf{R}$, and $\mathbf{b} \to \mathbf{R}\boldsymbol{v}_1$ in Eqs. (2.104) and (2.105), Eq. (2.103) is rewritten as follows:

$$p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2 | \boldsymbol{v}_1, \mathbf{N}_2, \phi_1, \phi_2) \propto |\mathbf{R}|^{\frac{N\phi_2}{2}} \exp\left(-\frac{N\phi_1}{2} \left(\boldsymbol{\mu} - \frac{\boldsymbol{v}_1}{N\phi_1} \right)^{\mathsf{T}} \mathbf{R} \left(\boldsymbol{\mu} - \frac{\boldsymbol{v}_1}{N\phi_1} \right) + \underbrace{\frac{\boldsymbol{v}_1^{\mathsf{T}} \mathbf{R} \boldsymbol{v}_1}{2N\phi_1} + \mathrm{tr}[\mathbf{R}\mathbf{N}_2]}_{(*)} \right). \tag{2.106}$$

Now we focus on the $(*)$ term in Eq. (2.106). By using the matrix formula in Appendix B, $(*)$ is rewritten as

$$\begin{aligned} (*) &= \mathrm{tr}\left[\frac{\boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}} \mathbf{R}}{2N\phi_1} + \mathbf{N}_2 \mathbf{R} \right] \\ &= \mathrm{tr}\left[\left(\frac{\boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}}}{2N\phi_1} + \mathbf{N}_2 \right) \mathbf{R} \right]. \end{aligned} \tag{2.107}$$

Thus, the conjugate prior distribution is rewritten as:

$$p(\boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_2 | \boldsymbol{v}_1, \mathbf{N}_2, \phi_1, \phi_2) \propto |\mathbf{R}|^{\frac{N\phi_2}{2}} \exp\left(-\frac{N\phi_1}{2} \left(\boldsymbol{\mu} - \frac{\boldsymbol{v}_1}{N\phi_1} \right)^{\mathsf{T}} \mathbf{R} \left(\boldsymbol{\mu} - \frac{\boldsymbol{v}_1}{N\phi_1} \right) + \mathrm{tr}\left[\left(\frac{\boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}}}{2N\phi_1} + \mathbf{N}_2 \right) \mathbf{R} \right] \right). \tag{2.108}$$

Therefore, Eq. (2.108) is represented as the following Gaussian–Wishart distribution in Appendix C.15:

$$\begin{aligned} &\mathcal{NW}(\boldsymbol{\mu}, \mathbf{R} | \boldsymbol{\mu}^0, \phi^{\mu}, \mathbf{R}^0, \phi^{\mathbf{R}}) \\ &\triangleq C_{\mathcal{NW}}(\phi^{\mu}, \mathbf{R}^0, \phi^{\mathbf{R}}) |\mathbf{R}|^{\frac{\phi^{\mathbf{R}} - D}{2}} \\ &\quad \times \exp\left(-\frac{1}{2} \mathrm{tr}\left[\mathbf{R}^0 \mathbf{R} \right] - \frac{\phi^{\mu}}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}^0)^{\mathsf{T}} \mathbf{R} (\boldsymbol{\mu} - \boldsymbol{\mu}^0) \right), \end{aligned} \tag{2.109}$$

where

$$\begin{cases} \phi^{\boldsymbol{\mu}} = N\phi_1 \\ \boldsymbol{\mu}^0 = \dfrac{\boldsymbol{v}_1}{N\phi_1} \\ \phi^{\mathbf{R}} = N\phi_2 + D \\ \mathbf{R}^0 = -\dfrac{\boldsymbol{v}_1 \boldsymbol{v}_1^{\mathsf{T}}}{N\phi_1} - 2\mathbf{N}_2. \end{cases} \tag{2.110}$$

Thus, we can derive the prior distribution as the Gaussian–Wishart distribution.

Now, we focus on the posterior distribution of $\boldsymbol{\mu}$ and $\mathbf{R}$. Similarly, the posterior distribution is represented as the same form of the Gaussian–Wishart distribution as the prior distribution (2.109), with hyperparameters $\hat{\phi}^{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\phi}^{\mathbf{R}}$, and $\hat{\mathbf{R}}$ as follows:

$$p(\boldsymbol{\mu}, \mathbf{R} | X) = \mathcal{NW}(\boldsymbol{\mu}, \mathbf{R} | \hat{\boldsymbol{\mu}}, \hat{\phi}^{\boldsymbol{\mu}}, \hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}). \tag{2.111}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Based on the conjugate distribution rule, Eq. (2.64), the hyperparameters of the posterior distribution are easily solved by just replacing $\boldsymbol{\nu}_1 \to \boldsymbol{\nu}_1 + \mathbf{t}(\mathbf{X})$, $\mathbf{N}_2 \to \mathbf{N}_2 + \mathbf{T}(\mathbf{X})$, and $\phi_m \to \phi_m + 1$ in Eq. (2.110) without complex calculations, as follows:

$$\begin{aligned}
\hat{\phi}^{\boldsymbol{\mu}} &= N(\phi_1 + 1) = \phi^{\boldsymbol{\mu}} + N, \\
\hat{\boldsymbol{\mu}} &= \frac{\boldsymbol{\nu}_1 + \sum_{n=1}^{N} \mathbf{x}_n}{N(\phi_1 + 1)} = \frac{\phi^{\boldsymbol{\mu}} \boldsymbol{\mu}^0 + \sum_{n=1}^{N} \mathbf{x}_n}{\phi^{\boldsymbol{\mu}} + N}, \\
\hat{\phi}^{\mathbf{R}} &= N(\phi_2 + 1) + D = \phi^{\mathbf{R}} + N, \\
\hat{\mathbf{R}} &= -\frac{\left(\boldsymbol{\nu}_1 + \sum_{n=1}^{N} \mathbf{x}_n\right)\left(\boldsymbol{\nu}_1 + \sum_{n=1}^{N} \mathbf{x}_n\right)^{\mathsf{T}}}{N(\phi_1 + 1)} - 2\left(\mathbf{N}_2 - \frac{1}{2}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}}\right) \\
&= -\hat{\phi}^{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^{\mathsf{T}} + \frac{\boldsymbol{\nu}_1 \boldsymbol{\nu}_1^{\mathsf{T}}}{N\phi_1} + \mathbf{R}^0 + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} \\
&= -\hat{\phi}^{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^{\mathsf{T}} + \phi^{\boldsymbol{\mu}} \boldsymbol{\mu} \boldsymbol{\mu}^{\mathsf{T}} + \mathbf{R}^0 + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}}.
\end{aligned} \tag{2.112}$$

Thus, we derive the posterior distribution that is also represented as a Gaussian–Wishart distribution. Finally, the prior and posterior distributions of $\boldsymbol{\mu}$ and $\mathbf{R}$ in Eqs. (2.109) and (2.111) are also summarized as:

$$\begin{cases}
p(\boldsymbol{\mu}, \mathbf{R}) = \mathcal{NW}(\boldsymbol{\mu}, \mathbf{R} | \boldsymbol{\mu}^0, \phi^{\mu}, \mathbf{R}^0, \phi^{\mathbf{R}}) \\
p(\boldsymbol{\mu}, \mathbf{R} | X) = \mathcal{NW}(\boldsymbol{\mu}, \mathbf{R} | \hat{\boldsymbol{\mu}}, \hat{\phi}^{\boldsymbol{\mu}}, \hat{\mathbf{R}}, \hat{\phi}^{\mathbf{R}}),
\end{cases} \tag{2.113}$$

or

$$\begin{cases}
p(\boldsymbol{\mu}, \mathbf{R}) = p(\boldsymbol{\mu}|\mathbf{R}) p(\mathbf{R}) = \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, (\phi^{\mu} \mathbf{R})^{-1}) \mathcal{W}(\mathbf{R} | \phi^{\mathbf{R}}, \mathbf{R}^0) \\
p(\boldsymbol{\mu}, \mathbf{R} | X) = p(\boldsymbol{\mu} | \mathbf{R}, X) p(\mathbf{R} | X) = \mathcal{N}(\boldsymbol{\mu} | \hat{\boldsymbol{\mu}}, (\hat{\phi}^{\mu} \mathbf{R})^{-1}) \mathcal{W}(\mathbf{R} | \hat{\phi}^{\mathbf{R}}, \hat{\mathbf{R}}).
\end{cases} \tag{2.114}$$

Example 2.8 Conjugate distributions for multinomial distribution:
Based on the discussion of Eq. (2.50) in Example 2.4, the canonical form of the multivariate Gaussian distribution is represented as follows:

$$\text{Mult}(x_1, \cdots, x_J | \omega_1, \cdots, \omega_J) = h(x) \exp\left(\boldsymbol{\gamma}^{\mathsf{T}} \mathbf{t}(\mathbf{x})\right), \tag{2.115}$$

where

$$\begin{cases}
\mathbf{t}(\mathbf{x}) = \mathbf{x} \\
h(x) = \dfrac{N!}{\prod_{j=1}^{J} x_j!} \\
\boldsymbol{\gamma} = \left[\log \frac{\omega_1}{1 - \sum_{j=1}^{J-1} \omega_j}, \cdots, \log \frac{\omega_{J-1}}{1 - \sum_{j=1}^{J-1} \omega_j}\right]^{\mathsf{T}} \\
g(\boldsymbol{\gamma}) = N \log\left(1 + \sum_{j=1}^{J-1} \exp(\gamma_j)\right).
\end{cases} \tag{2.116}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Note that we have the following constraints:

$$\sum_{j=1}^{J} x_j = N$$
$$\sum_{j=1}^{J} \omega_j = 1. \tag{2.117}$$

Therefore, by substituting $\boldsymbol{\gamma}$, $g_1(\boldsymbol{\gamma})$, and $g_2(\boldsymbol{\gamma})$ in Eq. (2.78) into the general form of the conjugate distribution in Eq. (2.56), we can derive the function of $\boldsymbol{\gamma}$ as follows:

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\nu},\phi) &\propto \exp\left(\boldsymbol{\gamma}^{\intercal}\boldsymbol{\nu} - \phi g(\boldsymbol{\gamma})\right) \\ &= \exp\left(\left[\log\frac{\omega_1}{1-\sum_{j=1}^{J-1}\omega_1}, \cdots, \log\frac{\omega_{J-1}}{1-\sum_{j=1}^{J-1}\omega_j}\right]\boldsymbol{\nu} + N\phi\log\left(1-\sum_{j=1}^{J-1}\omega_j\right)\right) \\ &= \exp\left(\left[\log\omega_1, \cdots, \log\omega_J\right]\left[\boldsymbol{\nu}^{\intercal}, N\phi - \textstyle\sum_{j=1}^{J-1}\nu_j\right]^{\intercal}\right) \\ &= \prod_{j=1}^{J}(\omega_j)^{\phi_j^{\omega}-1}, \end{aligned} \tag{2.118}$$

where hyperparameters $\{\phi_j^{\omega}\}_{j=1}^{J}$ are defined as follows:

$$\begin{aligned} \phi_j^{\omega} &\triangleq \nu_j + 1 \text{ for } j = 1, \cdots, J-1 \\ \phi_J^{\omega} &\triangleq N\phi - \sum_{j=1}^{J-1}\nu_j + 1. \end{aligned} \tag{2.119}$$

Thus, the conjugate prior distribution is represented as a Dirichlet distribution defined in Appendix C.4 as

$$\mathrm{Dir}(\{\omega_j\}_{j=1}^{J}|\{\phi_j^{\omega}\}_{j=1}^{J}) \triangleq \frac{\Gamma(\sum_{j=1}^{J}\phi_j^{\omega})}{\prod_{j=1}^{J}\Gamma(\phi_j^{\omega})}\prod_{j=1}^{J}(\omega_j)^{\phi_j^{\omega}-1}. \tag{2.120}$$

Based on the conjugate distribution rule, Eq. (2.63), the hyperparameters of the posterior distribution are easily solved by just replacing $\boldsymbol{\nu} \rightarrow \boldsymbol{\nu} + \mathbf{t}(\mathbf{x})$ and $\phi \rightarrow \phi + 1$ in Eq. (2.118) without complex calculations, as follows:

$$p(\boldsymbol{\gamma}|X) \rightarrow \mathrm{Dir}(\{\omega_j\}_{j=1}^{J}|\{\hat{\phi}_j^{\omega}\}_{j=1}^{J}), \tag{2.121}$$

where hyperparameters $\{\hat{\phi}_j^{\omega}\}_{j=1}^{J}$ are obtained as follows:

$$\begin{aligned} \hat{\phi}_j^{\omega} &\triangleq \nu_j + x_j + 1 = \phi_j^{\omega} + x_j \\ \hat{\phi}_J^{\omega} &\triangleq N(\phi+1) - \sum_{j=1}^{J-1}(\nu_j + x_j) + 1 \\ &= N\phi + x_J - \sum_{j=1}^{J-1}\nu_j + 1 \\ &= \phi_J^{\omega} + x_J. \end{aligned} \tag{2.122}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Thus, we derive the posterior distribution that is represented as a Dirichlet distribution. Finally, the prior and posterior distributions of ω are given as:

$$\begin{cases} p(\{\omega_j\}_{j=1}^J) = \mathrm{Dir}\left(\{\omega_j\}_{j=1}^J \middle| \{\phi_j^\omega\}_{j=1}^J\right) \\ p(\{\omega_j\}_{j=1}^J|X) = \mathrm{Dir}\left(\{\omega_j\}_{j=1}^J \middle| \{\hat{\phi}_j^\omega\}_{j=1}^J\right). \end{cases} \tag{2.123}$$

Table 2.1 shows a recipe of the kind of distributions we use as a conjugate prior.

This section provides a solution of the posterior distribution for rather simple statistical models. However, in practical applications, we still face the problems of solving the equations, and often require the approximation to solve them efficiently. The next section explains a powerful approximation method, conditional independence, in Bayesian probabilities.

2.1.5 Conditional independence

Another important mathematical operation of the Bayesian approach, as well as the product and sum rules (Section 2.1.1), is called *conditional independence*. Let a, b, and c be probabilistic variables, the conditional independence of a and b on c is represented as follows:

$$p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|a, c), \tag{2.124}$$

$$\approx p(a|c)p(b|c). \tag{2.125}$$

This is a useful assumption for the Bayesian approach when factorizing the joint probability distribution. For example, Eq. (2.124) based on the product rule needs to consider $p(b|a, c)$ or $p(a|b, c)$. Suppose a, b, and c are discrete elements of sets, i.e., $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $c \in \mathcal{C}$, $p(b|a, c)$ or $p(a|b, c)$ considers the probability of all combinations of a, b, and c, which correspond to $|\mathcal{A}| \times |\mathcal{B}| \times |\mathcal{C}|$. The number of combinations is increased exponentially, if the number of valuables is increased. Therefore, it is computationally very expensive to obtain the conditional distribution, and almost impossible to consider

Table 2.1 Conjugate priors.

Likelihood function	Unknown variable	Conjugate prior
Gaussian	$\mu \in \mathbb{R}$	Gaussian C.5
Gaussian	$r \in \mathbb{R}_{>0}$	Gamma C.11
Gaussian	μ, r	Gaussian–gamma C.13
Multivariate Gaussian	$\boldsymbol{\mu} \in \mathbb{R}^D$	Multivariate Gaussian C.6
Multivariate Gaussian	$\mathbf{R} \in \mathbb{R}^{D\times D}$	Wishart C.14
Multivariate Gaussian	$\boldsymbol{\mu}, \mathbf{R}$	Gaussian–Wishart C.15
Multinomial	$\omega_i \in [0, 1], \sum_i \omega_i = 1$	Dirichlet C.4

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

large amounts of data as probabilistic valuables. Thus, the conditional independence approximation in Eq. (2.125) greatly reduces the computational complexity, and makes the Bayesian treatment of speech and language processing tractable.

By using the product rule, the conditional independence equation is rewritten as follows:

$$p(a|c) \approx \frac{p(a,b|c)}{p(b|c)} = \frac{p(a|b,c)p(b|c)}{p(b|c)} = p(a|b,c). \tag{2.126}$$

Thus,

$$p(a|b,c) \approx p(a|c) \tag{2.127}$$

is also equivalently used as the conditional independence assumption of Eq. (2.125).

The conditional independence is often used in the following sections to make the complicated relationship between probabilistic variables simple. For example, speech recognition has many probabilistic variables which come from acoustic and language models. It is very natural and effective to assume conditional independence between acoustic model and language model variables because these do not depend on each other explicitly.

Example 2.9 Naive Bayes classifier:
One of the simplest classifiers in the machine learning approach is the naive Bayes classifier. The approach is used for many applications including document classification (Lewis 1998, McCallum & Nigam 1998). For example, if we have N data $(x_1, x_2, \cdots x_N)$, and want to classify the data to a specific category $\hat{c}$, this can be performed by using the posterior distribution of category c as follows:

$$\hat{c} = \arg\max_c p(c|\{x_n\}_{n=1}^N). \tag{2.128}$$

The naive Bayes classifier approximates this posterior distribution with the product rule and conditional independence assumption as follows:

$$\begin{aligned} p(c|\{x_n\}_{n=1}^N) &\propto p(\{x_n\}_{n=1}^N|c)p(c) \\ &\approx \prod_{n-1}^{N} p(x_n|c)p(c). \end{aligned} \tag{2.129}$$

This approach approximates the posterior distribution $p(c|\{x_n\}_{n=1}^N)$ with the product of likelihood $p(x_n|c)$ for all samples and prior distribution $p(c)$. Since the naive Bayes classifier is very simple and easy to implement, it is often used as an initial attempt of the machine learning approach if we have training data with labels to obtain $p(x_n|c)$ for all c. For example, in document classification, a multinomial distribution is used to represent the likelihood function $p(x_n|c)$.

---

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

2.2 Graphical model representation

The previous sections (especially Sections 2.1.1 and 2.1.5) discuss how to provide the mathematical relationship between probabilistic variables in a Bayesian manner. This section briefly introduces a graphical model representation that visualizes the relationship between these probabilistic valuables to provide a more intuitive way of understanding the model. A graphical model framework is also widely used in Bayesian machine learning studies, and this book introduces basic graphical model descriptions, which are used in the following sections.

2.2.1 Directed graph

First, we simply consider the following joint distribution of a and b, which can be rewritten as the following two factorization forms based on the product rule:

$$p(a, b) = p(b|a)p(a), \tag{2.130}$$

$$= p(a|b)p(b). \tag{2.131}$$

Therefore, to obtain the joint distribution, we compute either Eq. (2.130) or (2.131) depending on the problem. The graphical model can separately represent these factorization forms intuitively. Figure 2.1 represents the graphical models of $p(b|a)p(a)$ and $p(a|b)p(b)$, respectively. The node represents a probabilistic variable, and the directed link represents the conditional dependency of two probabilistic variables. For example, Eq. (2.130) is composed of the conditional distribution $p(b|a)$ and then the corresponding graphical representation provides the directed link from node a to node b in Figure 2.1(a). Conversely, the conditional distribution $p(a|b)$ in Eq. (2.131) is represented by the directed link from node b to node a in Figure 2.1(b).

Thus, the graphical model specifies a unique factorization form of a joint distribution, intuitively. The graph composed of the directed link, which represents the conditional distribution, is called a *directed graph*. The graphical model can also deal with an *undirected graph*, which is a graphical representation of a Markov random field, but this book focuses on the directed graph representation, which is often used in the later applications.

2.2.2 Conditional independence in graphical model

As we discussed in Section 2.1.5, practical applications often need some approximations in the dependency of probabilistic variables to avoid a complicated dependency of the

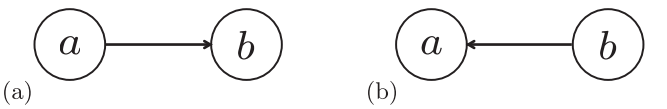

Figure 2.1 Graphical models of $p(b|a)p(a)$ and $p(a|b)p(b)$.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

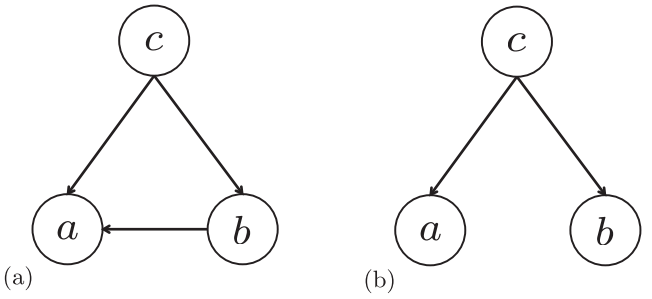

Figure 2.2 Graphical models of $p(a|b,c)p(b|c)p(c)$ and $p(a|c)p(b|c)p(c)$.

factorized distribution. We can represent this approximation in the graphical model representation. If we consider the joint distribution of a, b, and c, the joint distribution is, for example, represented as the following factorization form based on the product rule:

$$p(a,b,c) = p(a|b,c)p(b|c)p(c). \tag{2.132}$$

The graphical model of Eq. (2.132) is represented in Figure 2.2(a). Note that all nodes are connected to each other by directed links. This graph is called a *full connected graph*.

On the other hand, the joint distribution with the following conditional independence can also be represented as a graphical model in Figure 2.2(b):

$$p(a,b,c) = p(a,b|c)p(c) \approx p(a|c)p(b|c)p(c). \tag{2.133}$$

Note that the link between a and b has disappeared from Figure 2.2(b). Thus, the conditional independence in the graphical model is represented by pruning links in the graphs, which corresponds to reducing the dependencies in probabilistic variables, and leads to reduced computational cost.

In real applications, we need to consider large numbers of variables. For example, the naive Bayes classifier introduced in Example 2.9 has to consider $N+1$ probabilistic variables ($\{x_n\}_{n=1}^N$ and c):

$$p(x_1,\cdots,x_N|c)p(c) \approx p(x_1|c)\cdots p(x_N|c)p(c) = \prod_{n=1}^{N} p(x_n|c)p(c). \tag{2.134}$$

The graphical model of this case can be simplified from Figure 2.3(a) to 2.3(b) by using the plate. Based on the plate, we can represent a complicated relationship of probabilistic variables intuitively. In Section 8.2, we also consider the case when the number of probabilistic variables is dealt with as infinite in Bayesian nonparametrics. Then, the number of variables can be represented by using ∞ in a graphical model, as shown in Figure 2.4.

Thus, the graphical model can represent the dependencies of variables based on the product rule and conditional independence graphically. This dependency-network-based Bayesian method is also called a Bayesian network. In particular the Bayesian treatment that considers the dynamical relationship between probabilistic variables is also called a dynamic Bayesian network (Ghahramani 1998, Murphy 2002). A dynamic Bayesian

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

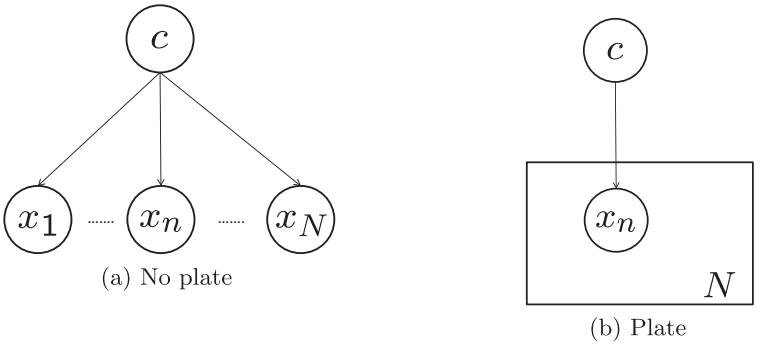

Figure 2.3 Graphical model of $p(x_1|c)\cdots p(x_N|c) = \prod_{n=1}^{N} p(x_n|c)$.

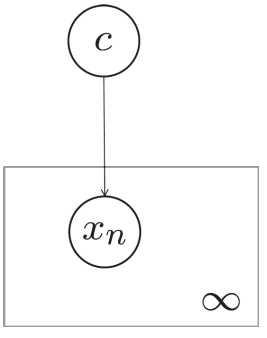

Figure 2.4 Graphical model of $p(x_1|c)\cdots p(x_\infty|c) = \prod_{n=1}^{\infty} p(x_n|c)$.

network provides efficient solutions to the time-series statistical models similarly to HMM and Kalman filters, which are also used in speech recognition (Zweig & Russell 1998, Nefian, Liang, Pi *et al*. 2002, Livescu, Glass & Bilmes 2003). It is helpful to understand probabilistic models, even when they are very complicated in the equation form.

2.2.3 Observation, latent variable, non-probabilistic variable

Previous sections deal with the graphical model of all probabilistic variables. However, our machine learning problems for speech and language processing have three types of variables: observation, latent variables, and non-probabilistic variables. For example, let x be an observation, z is a latent variable, and θ is a model parameter, which we don't deal with as a probabilistic variable in this section, unlike the full Bayesian approach. The probability distribution of x is represented as follows:

$$p(x|\theta) = \sum_z p(x, z|\theta) = \sum_z p(x|z, \theta)p(z|\theta). \tag{2.135}$$

The corresponding graphical model is represented in Figure 2.5(a). Note that three variables x, z, θ have different roles in this equation. For example, x is a final output of this

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

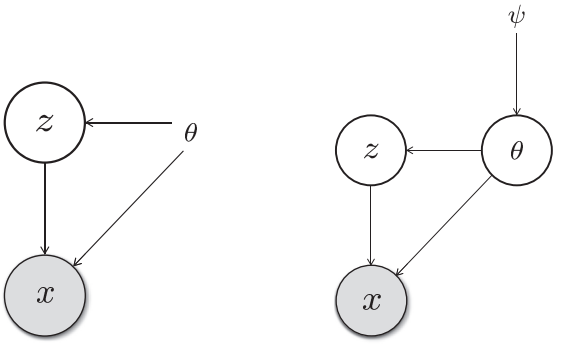

Figure 2.5 Graphical models that have observation x, latent variable z, model parameter θ, and hyperparameter ψ. Part (a) treats θ as a non-probabilistic variable, and (b) treats θ as a probabilistic variable to be marginalized.

equation as an observation, which is not marginalized, while z is a latent variable and should be marginalized. To distinguish the observation and latent variables, the node representing an observation is tinted. θ is not a probabilistic variable in this explanation, and so it is put in the graph *without a circle*.

Similarly, if we consider the same model, but treat θ as a probabilistic variable, θ is marginalized by the prior distribution of θ with hyperparameter ψ. The probability distribution of x is represented as follows:

$$\begin{aligned} p(x|\psi) &= \int \sum_z p(x, z, \theta|\psi) d\theta \\ &= \int \sum_z p(x|z, \theta) p(z|\theta) p(\theta|\psi) d\theta. \end{aligned} \tag{2.136}$$

Here we assume θ to be a continuous variable, and use the integral instead of the summation. We can regard θ as a latent variable in a broad sense, but the other sections distinguish the model parameters and latent variables. The corresponding graphical model is represented in Figure 2.5(b). Thus, by using the representations of observation, latent variables, and non-probabilistic variables, we can provide graphical models of various distributions other than joint distributions. These are basic rules of providing a directed graphical model from the corresponding probabilistic equation.

The directed graph basically describes how observation variables are generated conditioned on the other probabilistic variables. This statistical model of describing the generation of observation variables is called a *Generative model*. HMM, GMM, Kalman filter, n-gram, latent topic model, and deep belief network are typical examples of generative models that can generate speech feature vectors and word sequences. The next section also introduces another way of intuitively understanding our complicated statistical models by describing how observation variables are generated from the distributions in our models.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

2.2.4 Generative process

This section also explains another representation of the Bayesian approach based on the generative process. This representation is used to generate the probabilistic variables in an algorithmic way. The generative process is used to express the joint distribution. The basic syntax of the generative process is as follows:

- Non-probabilistic variables: placed as "require";
- Latent variables: "drawn" from their probability distribution;
- Model parameters: "drawn" from their (prior) probability distribution;
- Observations: finally "drawn" from their probability distribution given sampled latent variables and model parameters.

If we also want to represent the marginalization of a probabilistic variable, we can use an additional syntax "Average" for the marginalization.

As an example of Eq. (2.136), Algorithm 1 represents the generative process of the joint distribution $p(x, z, \theta|\psi)$, which is represented as:

$$p(x, z, \theta|\psi) = p(x|z, \theta)p(z|\theta)p(\theta|\psi), \tag{2.137}$$

where x, z, θ, and ψ are observations, latent variables, model parameters, and hyperparameter (non-probabilistic variables), respectively.

Algorithm 1 Generative process of $p(x, z, \theta|\psi) = p(x|z, \theta)p(z|\theta)p(\theta|\psi)$

Require: ψ

1: Draw θ from $p(\theta|\psi)$
2: Draw z from $p(z|\theta)$
3: Draw x from $p(x|z, \theta)$

This generative process also helps us to understand models intuitively by understanding how probabilistic variables are generated algorithmically. Therefore, both the generative process and graphical model are often provided in the Bayesian approach to represent a complicated generative model. In some of the statistical models used in this book, we provide the generative process and graphical model to allow readers to understand the models intuitively.

2.2.5 Undirected graph

Another example of a graphical model is called an undirected graph (Figure 2.6), that represents the relationship of probabilistic variables but does not have explicit parent–child relationships compared with the directed graph. The network is called a Markov random field, and the probabilistic distribution is usually expressed by a potential function $\psi(a, b, c)$ (a positive, but otherwise arbitrary, real-valued function):

$$p(a, b, c) = \frac{1}{Z}\psi(a, b, c), \tag{2.138}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

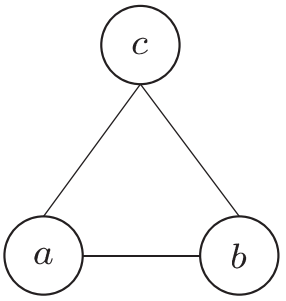

Figure 2.6 Undirected graph.

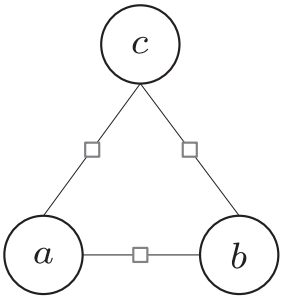

Figure 2.7 Factor graph.

where Z is a normalization constant of this distribution, and is called the *partition function* for this special case. This approach is often used as a context of a log linear discriminative model, where $\psi(a, b, c)$ is a linear function of a feature obtained by a, b, and c and the corresponding weight.

A *factor graph* is another class of graphical model representing the conditional independence relationship between variables. Actually, the factor graph can provide a more concrete representation of the joint distribution of variables than that of the undirected graph. The factor graph introduces additional square nodes to a graph, which can explicitly represent the dependency of several variables.

For example, the partition function can be represented by several cases, as shown in Figures 2.7 and 2.8. Both graphs are fully connected and can represent the joint distribution of a, b, and c. However, the partition function of Figure 2.7 is computed by using the three pairs of partition functions as follows:

$$p(a, b, c) = \frac{1}{Z}\psi(a, b)\psi(b, c)\psi(c, a). \tag{2.139}$$

The possible partition functions are $|\mathcal{A}| \times |\mathcal{B}| + |\mathcal{B}| \times |\mathcal{C}| + |\mathcal{A}| \times |\mathcal{C}|$. On the other hand, Figure 2.8 considers the potential function of the joint event for a, b, and c:

$$p(a, b, c) = \frac{1}{Z}\psi(a, b, c). \tag{2.140}$$

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

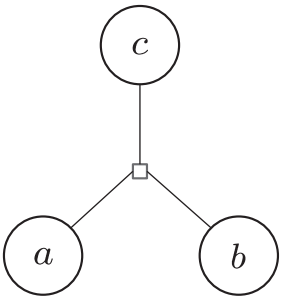

Figure 2.8 Factor graph.

The possible partition functions are $|\mathcal{A}| \times |\mathcal{B}| \times |\mathcal{C}|$. Therefore, if the number of possible variables ($|\mathcal{A}|$, $|\mathcal{B}|$, and $|\mathcal{C}|$) is very large, Figure 2.7 is a more compact representation since the number of possible functions would be smaller.

Thus, factor graphs are more specific about the precise form of the factorization of undirected graphs, and can be used to mainly represent some discriminative models (logistic regression, conditional random field (Lafferty, McCallum & Pereira 2001)). This book generally deals with generative models (HMM, GMM, n-gram, and latent topic model), and does not deal with these discriminative models. However, there are several important applications of discriminative models to speech and language processing (e.g., Gunawardana, Mahajan, Acero *et al*. 2005, Fosler & Morris 2008, Zweig & Nguyen 2009, Gales *et al*. 2012) in addition to the recent trend of deep neural networks (Hinton *et al*. 2012). We present an example of a Bayesian treatment of neural network acoustic models in Section 6.4. The fully Bayesian treatment of the other discriminative models in speech and language processing is an interesting future direction.

2.2.6 Inference on graphs

One of the powerful advantages of the graphical model representation is that once we fix a graphical model, we can infer all variables in the graph efficiently by using belief propagation if the graph does not have a loop.

For example, belief propagation provides a sum product algorithm that can efficiently compute the distribution $p(x_i)$ of the probabilistic variable in an arbitrary node by using message passing. In the HMM case, this sum product algorithm corresponds to the forward–backward algorithm, as discussed in Section 3.3.1. Similarly, belief propagation provides a max sum algorithm that can efficiently compute the arg max value ($\hat{x}_i = \arg\max_{x_i} p(x_i)$) in an arbitrary node by using message passing. Similarly to the sum product algorithm, the max sum algorithm corresponds to the Viterbi algorithm, as discussed in Section 3.3.2. A detailed discussion about the relationship between the forward–backward/Viterbi algorithms in the HMM and these algorithms can be found in Bishop (2006).

However, most of our applications have a loop in a graph, and we cannot use the exact inference based on the above algorithms. The following chapters introduce the approximations of the Bayesian inferences, and especially variational Bayes (VB), as discussed in Chapter 7, and Markov chain Monte Carlo (MCMC), as discussed in

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

Chapter 8, these being promising approaches to obtain approximate inferences in a graphical model. Actually, progress of the graphical model approach has been linked to the progress of these Bayesian inference techniques.

The other approximated approach to inference in a graphical model that contains cycles or loops is to use the sum-product algorithm for the graph even though there is no guarantee of convergence. This approach is called loopy belief propagation, and it is empirically known that it is convergent in some applications.

2.3 Difference between ML and Bayes

As discussed in previous sections, the Bayesian approach deals with all variables introduced for modeling as probabilistic variables. This is the unique difference between the Bayesian approach and the other standard statistical framework, the Maximum Likelihood (ML) approach. Actually this difference can yield various advantages over ML. This section overviews the advantage of the Bayesian approach over the ML approach in general. We discuss this, along with a general pattern recognition problem, as we consider practical speech and language processing issues in the following chapters.

Let $\mathbf{O}$, $\mathbf{Z}, \Theta$, $\boldsymbol{M}$, and $\boldsymbol{W}$ be a set of observation features, latent variables, model parameters, model structure (hyperparameter) variables, and classification categories, respectively, details of which will be introduced in the following chapters. For comparison, we summarize the difference between the approaches in terms of model setting, training, and classification.

- **Model setting**
  - ML:
    Generative model distribution $p(\mathbf{O}, Z|\Theta, M)$.
  - Bayes:
    Generative model distribution $p(\mathbf{O}, Z|\Theta, M)$
    Prior distributions $p(\Theta|M)$ and $p(M)$.

  In addition to the generative model distribution, the Bayesian approach needs to set prior distributions.
- **Training**
  - ML: Point estimation
    $\hat{\Theta}$.
  - Bayes: Distribution estimation
    $p(\Theta|M, \mathbf{O})$ and $p(M|\mathbf{O})$.

  ML point-estimates are given by the optimal values $\hat{\Theta}$ by using the EM algorithm generally when the model has latent variables, while the Bayesian approach estimates posterior distributions. In addition, ML only focuses on model parameters Θ, but the Bayesian approach focuses on both model parameters Θ and model M.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

- **Classification**
  - ML:

$$\arg\max_{W'} \sum_{Z'} p(\mathbf{O}', Z'|\hat{\Theta}, \hat{M}, W')p(W'). \tag{2.141}$$

  - Bayes:

$$\arg\max_{W'} \int \sum_{M,Z'} p(\mathbf{O}', Z'|\Theta, M, W')p(\Theta|M, \mathbf{O}, W)p(M|\mathbf{O}, W)p(W')d\Theta. \tag{2.142}$$

  Here, $\hat{\Theta}$ is obtained in the ML training step, and $\hat{M}$ is usually set in advance by an expert or optimized by evaluating the performance of model M using a development set. Equation (2.142) is obtained by the probabilistic sum and product rules and conditional independence, as discussed in the previous sections. Compared with ML, the Bayes approach marginalizes Θ and M through the expectations of the posterior distributions $p(\Theta|M, \mathbf{O}, W)$ and $p(M|\mathbf{O}, W)$, respectively.

Thus, the main differences between ML and Bayes are (i) use of prior distributions, (ii) use of distributions of model M, (iii) expectation with respect to probabilistic variables based on posterior distributions. These differences yield several advantages of the Bayesian approach over ML. The following sections describe the three main advantages.

2.3.1 Use of prior knowledge

First, we describe the most famous Bayesian advantage over ML based on the use of prior knowledge. Figure 2.9 depicts this advantage focusing on the estimation of model parameters (the mean and variance of a Gaussian). The dashed line shows the true distribution, and the solid line shows the estimated Gaussian distributions based on ML and Bayes. If data to be used to estimate parameters are not sufficient and biased to the

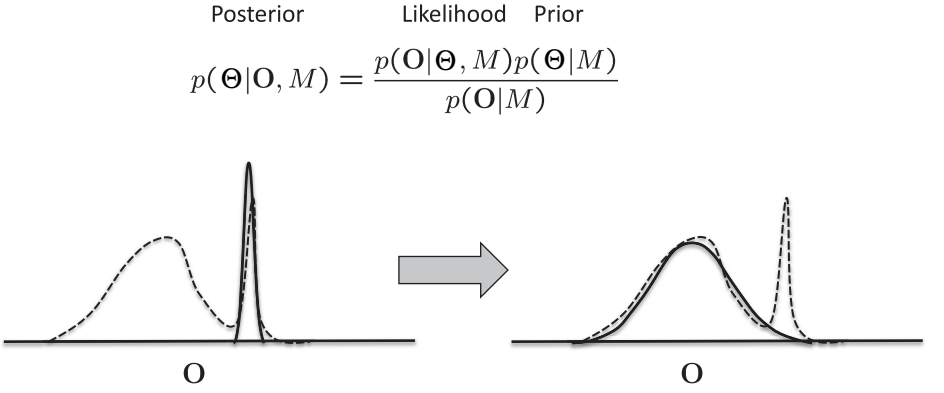

Figure 2.9 Use of prior knowledge.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

small peak of the true distribution, ML tends to estimate the wrong parameter by using the biased data. This is because ML estimates only consider the likelihood function, which leads it to estimate the parameters that generate the observed data:

$$\Theta^{\mathrm{ML}} = \arg\max_{\Theta} p(\mathbf{O}|\Theta, M). \tag{2.143}$$

Thus, ML can correctly estimate parameters only when the amount of data is sufficient.

On the other hand, the Bayesian approach also considers the prior distribution of model parameters $p(\Theta|M)$. Now, we consider point estimation using the maximum a-posteriori (MAP) value, instead of considering the Bayesian distribution estimation for simply comparing the prior effect with ML. For example, based on Eq. (3.345), the MAP estimate of Θ is represented as follows:

$$\begin{aligned}\Theta^{\mathrm{MAP}} &= \arg\max_{\Theta} p(\Theta|\mathbf{O}, M) \\ &= \arg\max_{\Theta} p(\mathbf{O}|\Theta, M)p(\Theta|M).\end{aligned} \tag{2.144}$$

The result considers the prior distribution as a regularization term. So if we set a constraint on a distribution form by appropriate prior knowledge, we can recover a wrong estimation due to the sparse data problem in ML, and we can estimate the parameter correctly. Details are discussed in Chapter 4.

2.3.2 Model selection

The model selection is a unique function of the Bayesian approach, which determines a model structure from data automatically. For example, Figure 2.10 shows how many Gaussians we use to estimate the parameters. It is well known that likelihood values always increase as the number of parameters increases. Therefore, if we enforce use of the ML criterion for the model selection, ML tends to select too many Gaussians, which results in over-fitting. There are some extensions of ML to deal with model selection based on information criteria (e.g., Akaike 1974, Rissanen 1984). However, in most cases of speech and language processing, the ML framework usually optimizes model structure by evaluating the performance of the model using a development set. The development set is usually obtained from a part of training/test data. Although this

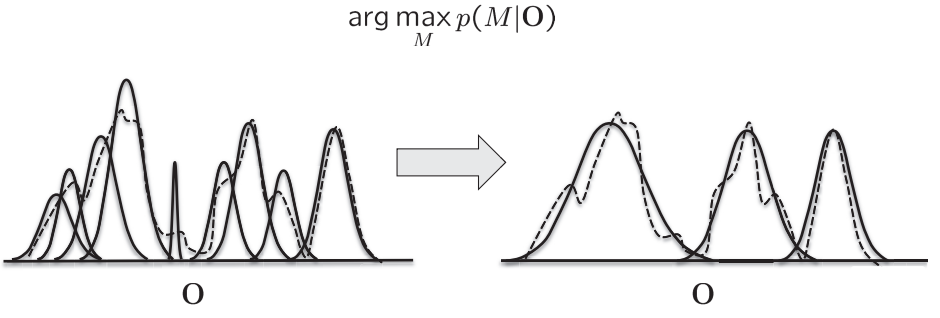

Figure 2.10 Model selection.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

optimization is straightforward, it is very difficult to use in some of the applications when the performance evaluation is difficult (e.g., it has a large computational cost for evaluation or there is no objective performance measure).

The Bayesian approach can deal with model selection within the framework. For example, our Bayesian approach to acoustic modeling involves the posterior distribution of model $p(M|\mathbf{O})$, which will be described in Section 3.8.7. Once we obtain the posterior disquisition, we can select an appropriate model structure in terms of the MAP procedure introduced in Section 2.1.2, as follows:

$$M^{\mathrm{MAP}} = \arg\max_{M} p(M|\mathbf{O}). \tag{2.145}$$

Thus, an appropriate model structure (e.g., the topology of HMMs (Stolcke & Omohundro 1993, Watanabe, Minami, Nakamura *et al*. 2004)) can be selected according to training data, without splitting them to create development data.

Instead of using the MAP procedure, we can use expectation based on $p(M|\mathbf{O})$. This is stricter in the Bayesian sense, and Eq. (2.142) actually includes the expectation over the posterior distribution of models. This approach corresponds to using multiple models with different model structures to classify unseen categories. However, in terms of the computational costs (needs large memory and computational time for the multiple model case), people usually carry out model selection by using the MAP procedure.

We also note that M involves other model variations than model structure as elements. For example, hyperparameters introduced in the model can be optimized by using the same MAP procedure or marginalized out by using the expectation. In particular, the optimization of hyperparameters through the posterior distributions of M is also a powerful example of the Bayesian advantage over ML.

2.3.3 Marginalization

The final Bayesian advantage over ML is the marginalization effect, which was discussed in the expectation effect of the Bayesian approach in the previous section. Since the stochastic fluctuation in the expectation absorbs estimation errors, the marginalization improves the robustness in estimation, classification, and regression over unknown data. In Figure 2.11, the left figure shows the maximum likelihood based distribution with $\hat{\Theta}$, while the right figure shows the example of marginalization over model

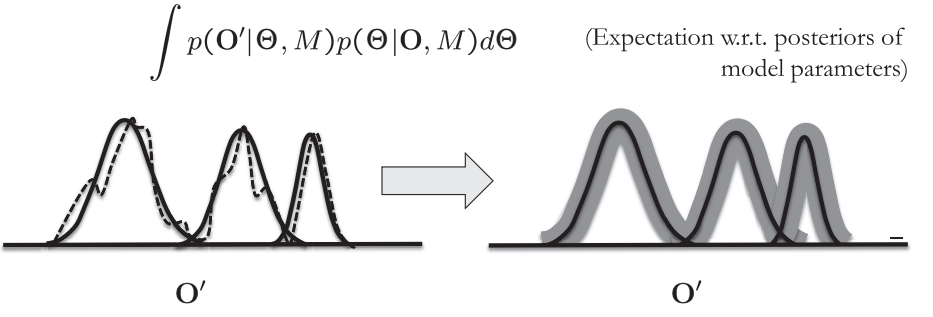

Figure 2.11 Marginalization over a set of model parameters Θ.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

parameters Θ in the likelihood function $p(\mathbf{O}'|\Theta, M)$ given a model structure M. Since the right figure considers the probabilistic fluctuations of Θ by the posterior distribution $p(\Theta|\mathbf{O}, M)$, the marginalized function (expected with respect to Θ) can mitigate the error effects in estimating Θ with the variance, and make the likelihood function robust to unseen data. The marginalization can be performed to all probabilistic variables in a model including latent variables Z, model structure M (and hyperparameters), in addition to the model parameter Θ example in Figure 2.11, if we obtain the prior/posterior distributions of these probabilistic variables.

The marginalization is another unique advantage of the Bayesian approach over ML, whereby incorporating the uncertainty of variables introduced in a model based on probabilistic theory achieves robustness for unseen data. However, it requires an expectation with respect to variables that essentially needs to consider the integral or summations over the variables. Again, this is the main difficulty of the practical Bayesian approach, and it needs some approximations especially to utilize the Bayesian advantage of this marginalization effect.

Although marginalization is not usually performed for observation $\mathbf{O}$, observation features in speech and language processing often include noises, and marginalization over observation features is effective for some applications. For example, if we use a speech enhancement technique as a front-end denoising process of automatic speech recognition, the process inevitably includes noise estimation errors, and the errors can be propagated to speech recognition, which degrades the performance greatly. The approach called *uncertainty techniques* tries to mitigate the errors by using the following Bayesian marginalization of the conventional continuous-density HMM (CDHMM) likelihood function over observation features $\mathbf{O}$ (Droppo, Acero & Deng 2002, Delcroix, Nakatani & Watanabe 2009, Kolossa & Haeb-Umbach 2011):

$$p(\Theta, \Psi_{\mathbf{O}'}^{\text{uns}}, M) \approx \int p(\mathbf{O}'|\Theta, M) p(\mathbf{O}'|\Psi_{\mathbf{O}'}^{\text{uns}}) d\mathbf{O}'. \tag{2.146}$$

The main challenges of the uncertainty techniques are how to estimate feature uncertainties $\Psi_{\mathbf{O}'}^{\text{uns}}$ (the distribution of observation features $p(\mathbf{O}|\Psi_{\mathbf{O}'}^{\text{uns}})$ with hyperparameter $\Psi_{\mathbf{O}'}^{\text{uns}}$) and how to integrate the marginal likelihood function with the decoding algorithm of the HMM. The approaches have been successfully applied to noisy speech recognition tasks, and show improvements by mitigating the error effects in speech enhancement techniques (Barker, Vincent, Ma *et al*. 2013, Vincent, Barker, Watanabe *et al*. 2013).

Thus, we have explained the three main practical advantages of the Bayesian approaches. Note that all of the advantages are based on the posterior distributions, and obtaining the posterior distributions for our target applications is a main issue of the Bayesian approaches. Once we obtain the posterior distributions, Bayesian inference allows us to achieve robust performance for our applications.

2.4 Summary

This chapter introduces the selected Bayesian approaches used for speech and language processing by starting from the basic Bayesian probabilistic theory with graphical

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press

models, and concludes with a summarization of the Bayesian advantages over ML. The discussion is rather general, and to apply Bayesian approaches to our practical problems in speech and language processing, we still need to bridge a gap between the theoretical Bayesian approaches and these practical problems. This is the main goal of this book. The next chapter deals with basic statistical models used in speech and language processing based on ML, and it will be extended in the latter chapters toward this main goal.

https://doi.org/10.1017/CBO9781107295360.003 Published online by Cambridge University Press