1
Bayesian Learning

Bayesian learning is an inference method based on the fundamental law of probability, called the Bayes theorem. In this first chapter, we introduce the framework of Bayesian learning with simple examples where Bayesian learning can be performed analytically.

1.1 Framework

Bayesian learning considers the following situation. We have observed a set $\mathcal{D}$ of data, which are subject to a *conditional distribution* $p(\mathcal{D}|\boldsymbol{w})$, called the *model distribution*, of the data given unknown *model parameter* $\boldsymbol{w}$. Although the value of $\boldsymbol{w}$ is unknown, vague information on $\boldsymbol{w}$ is provided as a *prior distribution* $p(\boldsymbol{w})$. The conditional distribution $p(\mathcal{D}|\boldsymbol{w})$ is also called the *model likelihood* when it is seen as a function of the unknown parameter $\boldsymbol{w}$.

1.1.1 Bayes Theorem and Bayes Posterior

Bayesian learning is based on the following basic factorization property of the *joint distribution* $p(\mathcal{D}, \boldsymbol{w})$:

$$\underbrace{p(\boldsymbol{w}|\mathcal{D})}_{\text{posterior}}\ \underbrace{p(\mathcal{D})}_{\text{marginal}} = \underbrace{p(\mathcal{D}, \boldsymbol{w})}_{\text{joint}} = \underbrace{p(\mathcal{D}|\boldsymbol{w})}_{\text{likelihood}}\ \underbrace{p(\boldsymbol{w})}_{\text{prior}}, \tag{1.1}$$

where the marginal distribution is given by

$$p(\mathcal{D}) = \int_{\mathcal{W}} p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w} = \int_{\mathcal{W}} p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}. \tag{1.2}$$

Here, the integration is performed in the domain $\mathcal{W}$ of the parameter $\boldsymbol{w}$. Note that, if the domain $\mathcal{W}$ is discrete, integration should be replaced with

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

summation, i.e., for any function $f(\boldsymbol{w})$,

$$\int_{\mathcal{W}} f(\boldsymbol{w}) d\boldsymbol{w} \to \sum_{\boldsymbol{w}' \in \mathcal{W}} f(\boldsymbol{w}').$$

The *posterior distribution*, the distribution of the unknown parameter $\boldsymbol{w}$ given the observed data set $\mathcal{D}$, is derived by dividing both sides of Eq. (1.1) by the marginal distribution $p(\mathcal{D})$:

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}, \boldsymbol{w})}{p(\mathcal{D})} \propto p(\mathcal{D}, \boldsymbol{w}). \tag{1.3}$$

Here, we emphasized that the posterior distribution is proportional to the joint distribution $p(\mathcal{D}, \boldsymbol{w})$ because the marginal distribution $p(\mathcal{D})$ is a constant (as a function of $\boldsymbol{w}$). In other words, the joint distribution is an *unnormalized posterior distribution*. Eq. (1.3) is called the *Bayes theorem*, and the posterior distribution computed exactly by Eq. (1.3) is called the *Bayes posterior* when we distinguish it from its approximations.

Example 1.1 (Parametric density estimation) Assume that the observed data $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ consist of N *independent and identically distributed (i.i.d.)* samples from the model distribution $p(\boldsymbol{x}|\boldsymbol{w})$. Then, the model likelihood is given by $p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w})$, and therefore, the posterior distribution is given by

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{\prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}) p(\boldsymbol{w})}{\int \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}} \propto \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}) p(\boldsymbol{w}).$$

Example 1.2 (Parametric regression) Assume that the observed data $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$ consist of N i.i.d. input–output pairs from the model distribution $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) p(\boldsymbol{x})$. Then, the likelihood function is given by $p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w}) p(\boldsymbol{x}^{(n)})$, and therefore, the posterior distribution is given by

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{\prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w}) p(\boldsymbol{w})}{\int \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}} \propto \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)}, \boldsymbol{w}) p(\boldsymbol{w}).$$

Note that the input distribution $p(\boldsymbol{x})$ does not affect the posterior, and accordingly is often ignored in practice.

1.1.2 Maximum A Posteriori Learning

Since the joint distribution $p(\mathcal{D}, \boldsymbol{w})$ is just the product of the likelihood function and the prior distribution (see Eq. (1.1)), it is usually easy to

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

compute. Therefore, it is relatively easy to perform *maximum a posteriori (MAP) learning*, where the parameters are point-estimated so that the posterior probability is maximized, i.e.,

$$\widehat{\boldsymbol{w}}^{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D}) = \operatorname*{argmax}_{\boldsymbol{w}} p(\mathcal{D}, \boldsymbol{w}). \tag{1.4}$$

MAP learning includes *maximum likelihood (ML) learning*,

$$\widehat{\boldsymbol{w}}^{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{w}} p(\mathcal{D}|\boldsymbol{w}), \tag{1.5}$$

as a special case with the flat prior $p(\boldsymbol{w}) \propto 1$.

1.1.3 Bayesian Learning

On the other hand, *Bayesian learning* requires integration of the joint distribution with respect to the parameter $\boldsymbol{w}$, which is often computationally hard. More specifically, performing Bayesian learning means computing at least one of the following quantities:

Marginal likelihood **(zeroth moment)**

$$p(\mathcal{D}) = \int p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}. \tag{1.6}$$

This quantity has been already introduced in Eq. (1.2) as the normalization factor of the posterior distribution. As seen in Section 1.1.5 and subsequent sections, the marginal likelihood plays an important role in model selection and hyperparameter estimation.

Posterior mean **(first moment)**

$$\widehat{\boldsymbol{w}} = \langle \boldsymbol{w} \rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int \boldsymbol{w} \cdot p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}, \tag{1.7}$$

where $\langle \cdot \rangle_p$ denotes the expectation value over the distribution p, i.e., $\langle \cdot \rangle_{p(\boldsymbol{w})} = \int \cdot \, p(\boldsymbol{w}) d\boldsymbol{w}$. This quantity is also called the *Bayesian estimator*. The Bayesian estimator or the model distribution with the Bayesian estimator plugged in (see the plug-in predictive distribution (1.10)) can be the final output of Bayesian learning.

Posterior covariance **(second moment)**

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \left\langle (\boldsymbol{w} - \widehat{\boldsymbol{w}})(\boldsymbol{w} - \widehat{\boldsymbol{w}})^\top \right\rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})} \int (\boldsymbol{w} - \widehat{\boldsymbol{w}})(\boldsymbol{w} - \widehat{\boldsymbol{w}})^\top p(\mathcal{D}, \boldsymbol{w}) d\boldsymbol{w}, \tag{1.8}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

where $\top$ denotes the transpose of a matrix or vector. This quantity provides the credibility information, and is used to assess the confidence level of the Bayesian estimator.

Predictive distribution **(expectation of model distribution)**

$$p(\mathcal{D}^{\text{new}}|\mathcal{D}) = \left\langle p(\mathcal{D}^{\text{new}}|\boldsymbol{w})\right\rangle_{p(\boldsymbol{w}|\mathcal{D})} = \frac{1}{p(\mathcal{D})}\int p(\mathcal{D}^{\text{new}}|\boldsymbol{w})p(\mathcal{D},\boldsymbol{w})d\boldsymbol{w}, \tag{1.9}$$

where $p(\mathcal{D}^{\text{new}}|\boldsymbol{w})$ denotes the model distribution on *unobserved* new data $\mathcal{D}^{\text{new}}$. In the i.i.d. case such as Examples 1.1 and 1.2, it is sufficient to compute the predictive distribution for a single new sample $\mathcal{D}^{\text{new}} = \{\boldsymbol{x}\}$.

Note that each of the four quantities (1.6) through (1.9) requires to compute the expectation of some function $f(\boldsymbol{w})$ over the unnormalized posterior distribution $p(\mathcal{D},\boldsymbol{w})$ on $\boldsymbol{w}$, i.e., $\int f(\boldsymbol{w})p(\mathcal{D},\boldsymbol{w})d\boldsymbol{w}$. Specifically, the marginal likelihood, the posterior mean, and the posterior covariance are the zeroth, the first, and the second moments of the unnormalized posterior distribution, respectively. The expectation is analytically intractable except for some simple cases, and numerical computation is also hard when the dimensionality of the unknown parameter $\boldsymbol{w}$ is high. This is the main bottleneck of Bayesian learning, with which many approximation methods have been developed to cope.

It hardly happens that the first moment (1.7) or the second moment (1.8) are computationally tractable but the zeroth moment (1.6) is not. Accordingly, we can say that performing Bayesian learning on the parameter $\boldsymbol{w}$ amounts to obtaining the *normalized* posterior distribution $p(\boldsymbol{w}|\mathcal{D})$. It sometimes happens that computing the predictive distribution (1.9) is still intractable even if the zeroth, the first, and the second moments can be computed based on some approximation. In such a case, the model distribution with the Bayesian estimator plugged in, called the *plug-in predictive distribution*,

$$p(\mathcal{D}^{\text{new}}|\widehat{\boldsymbol{w}}), \tag{1.10}$$

is used for prediction in practice.

1.1.4 Latent Variables

So far, we introduced the observed data set $\mathcal{D}$ as a known variable, and the model parameter $\boldsymbol{w}$ as an unknown variable. In practice, more varieties of known and unknown variables can be involved.

Some probabilistic models have *latent variables* (or *hidden variables*) $\boldsymbol{z}$, which can be involved in the original model, or additionally introduced for

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

computational reasons. They are typically attributed to each of the observed samples, and therefore have large degrees of freedom. However, they are just additional unknown variables, and there is no reason in inference to distinguish them from the model parameters $\boldsymbol{w}$.[1] The joint posterior over the parameters and the latent variables is given by Eq. (1.3) with $\boldsymbol{w}$ and $p(\boldsymbol{w})$ replaced with $\overline{\boldsymbol{w}} = (\boldsymbol{w}, \boldsymbol{z})$ and $p(\overline{\boldsymbol{w}}) = p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})$, respectively.

Example 1.3 (Mixture models) A mixture model is often used for parametric density estimation (Example 1.1). The model distribution is given by

$$p(\boldsymbol{x}|\boldsymbol{w}) = \sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k), \tag{1.11}$$

where $\boldsymbol{w} = \{\alpha_k, \boldsymbol{\tau}_k; \alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1\}_{k=1}^{K}$ is the unknown parameters. The mixture model (1.11) is the weighted sum of K distributions, each of which is parameterized by the component parameter $\boldsymbol{\tau}_k$. The domain of the *mixing weights* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$, also called as the *mixture coefficients*, forms the *standard* $(K-1)$*-simplex*, denoted by $\Delta^{K-1} \equiv \{\boldsymbol{\alpha} \in \mathbb{R}_+^K; \sum_{k=1}^{K} \alpha_k = 1\}$ (see Figure 1.1). Figure 1.2 shows an example of the mixture model with three one-dimensional Gaussian components.

The likelihood,

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{w}) &= \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}), \\ &= \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k) \right), \end{aligned} \tag{1.12}$$

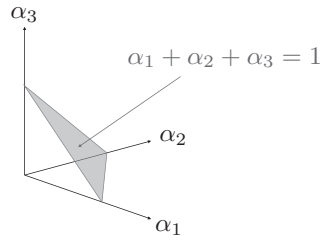

Figure 1.1 $(K-1)$-simplex, Δ^{K-1}, for $K = 3$.

[1] For this reason, the latent variables $\boldsymbol{z}$ and the model parameters $\boldsymbol{w}$ are also called *local latent variables* and *global latent variables*, respectively.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

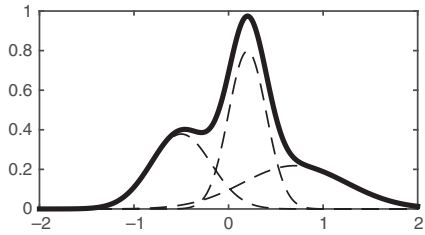

Figure 1.2 Gaussian mixture.

for N observed i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ has $O(K^N)$ terms, which makes even ML learning intractable. This intractability arises from the summation inside the multiplication in Eq. (1.12). By introducing latent variables, we can turn this summation into a multiplication, and make Eq. (1.12) tractable.

Assume that each sample $\boldsymbol{x}$ belongs to a single component k, and is drawn from $p(\boldsymbol{x}|\boldsymbol{\tau}_k)$. To describe the assignment, we introduce a latent variable $\boldsymbol{z} \in \mathcal{Z} \equiv \{\boldsymbol{e}_k\}_{k=1}^K$ associated with each observed sample $\boldsymbol{x}$, where $\boldsymbol{e}_k \in \{0,1\}^K$ is the K-dimensional binary vector, called the *one-of-K representation*, with one at the kth entry and zeros at the other entries:

$$\boldsymbol{e}_k = (\underbrace{0, \ldots, 0, \overbrace{1}^{k\text{th}}, 0, \ldots, 0}_{K})^\top.$$

Then, we have the following model:

$$p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{w}) = p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w}) p(\boldsymbol{z}|\boldsymbol{w}), \tag{1.13}$$

$$\text{where} \qquad p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w}) = \prod_{k=1}^K \{p(\boldsymbol{x}|\boldsymbol{\tau}_k)\}^{z_k}, \qquad p(\boldsymbol{z}|\boldsymbol{w}) = \prod_{k=1}^K \alpha_k^{z_k}.$$

The conditional distribution (1.13) on the observed variable $\boldsymbol{x}$ and the latent variable $\boldsymbol{z}$ given the parameter $\boldsymbol{w}$ is called the *complete likelihood*.

Note that marginalizing the complete likelihood over the latent variable recovers the original mixture model:

$$p(\boldsymbol{x}|\boldsymbol{w}) = \int_{\mathcal{Z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{w}) d\boldsymbol{z} = \sum_{\boldsymbol{z} \in \{\boldsymbol{e}_k\}_{k=1}^K} \prod_{k=1}^K \{\alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k)\}^{z_k} = \sum_{k=1}^K \alpha_k p(\boldsymbol{x}|\boldsymbol{\tau}_k).$$

This means that, if samples are generated from the model distribution (1.13), and only $\boldsymbol{x}$ is recorded, the observed data follow the original mixture model (1.11).

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

In the literature, latent variables tend to be marginalized out even in MAP learning. For example, the *expectation-maximization (EM) algorithm* (Dempster et al., 1977), a popular MAP solver for latent variable models, seeks a (local) maximizer of the posterior distribution with the latent variables marginalized out, i.e.,

$$\widehat{\boldsymbol{w}}^{\mathrm{EM}} = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|\mathcal{D}) = \operatorname*{argmax}_{\boldsymbol{w}} \int_{\mathcal{Z}} p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}) d\boldsymbol{z}. \tag{1.14}$$

However, we can also maximize the posterior jointly over the parameters and the latent variables, i.e.,

$$(\widehat{\boldsymbol{w}}^{\mathrm{MAP-hard}}, \widehat{\boldsymbol{z}}^{\mathrm{MAP-hard}}) = \operatorname*{argmax}_{\boldsymbol{w},\boldsymbol{z}} p(\boldsymbol{w}, \boldsymbol{z}|\mathcal{D}) = \operatorname*{argmax}_{\boldsymbol{w},\boldsymbol{z}} p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}). \tag{1.15}$$

For clustering based on the mixture model in Example 1.3, the EM algorithm (1.14) gives a *soft assignment*, where the expectation value $\widehat{\boldsymbol{z}}^{\mathrm{EM}} \in \Delta^{K-1} \subset [0,1]^K$ is substituted into the joint distribution $p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z})$, while the joint maximization (1.15) gives the *hard assignment*, where the optimal assignment $\widehat{\boldsymbol{z}}^{\mathrm{MAP-hard}} \in \{\boldsymbol{e}_k\}_{k=1}^K \subset \{0,1\}^K$ is looked for in the binary domain.

1.1.5 Empirical Bayesian Learning

In many practical cases, it is reasonable to use a prior distribution parameterized by *hyperparameters* $\boldsymbol{\kappa}$. The hyperparameters can be tuned by hand or based on some criterion outside the Bayesian framework. A popular method of the latter is the *cross validation*, where the hyperparameters are tuned so that an (preferably unbiased) estimator of the performance criterion is optimized. In such cases, the hyperparameters should be treated as *known* variables when Bayesian learning is performed.

On the other hand, the hyperparameters can be estimated within the Bayesian framework. In this case, there is again no reason to distinguish the hyperparameters from the other unknown variables $(\boldsymbol{w}, \boldsymbol{z})$. The joint posterior over all unknown variables is given by Eq. (1.3) with $\boldsymbol{w}$ and $p(\boldsymbol{w})$ replaced with $\overline{\boldsymbol{w}} = (\boldsymbol{w}, \boldsymbol{\kappa}, \boldsymbol{z})$ and $p(\overline{\boldsymbol{w}}) = p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})$, respectively, where $p(\boldsymbol{\kappa})$ is called a *hyperprior*. A popular approach, called *empirical Bayesian (EBayes) learning* (Efron and Morris, 1973), applies Bayesian learning on $\boldsymbol{w}$ (and $\boldsymbol{z}$) and point-estimate $\boldsymbol{\kappa}$, i.e.,

$$\widehat{\boldsymbol{\kappa}}^{\mathrm{EBayes}} = \operatorname*{argmax}_{\boldsymbol{\kappa}} p(\mathcal{D}, \boldsymbol{\kappa}) = \operatorname*{argmax}_{\boldsymbol{\kappa}} p(\mathcal{D}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}),$$
$$\text{where} \quad p(\mathcal{D}|\boldsymbol{\kappa}) = \int p(\mathcal{D}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\kappa}) d\boldsymbol{w} d\boldsymbol{z}.$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Here the marginal likelihood $p(\mathcal{D}|\boldsymbol{\kappa})$ is seen as the likelihood of the hyperparameter $\boldsymbol{\kappa}$, and MAP learning is performed by maximizing the joint distribution $p(\mathcal{D}, \boldsymbol{\kappa})$ of the observed data $\mathcal{D}$ and the hyperparameter $\boldsymbol{\kappa}$, which can be seen as an *unnormalized posterior distribution* of the hyperparameter. The hyperprior is often assumed to be flat: $p(\boldsymbol{\kappa}) \propto 1$.

With an appropriate design of priors, empirical Bayesian learning combined with approximate Bayesian learning is often used for *automatic relevance determination (ARD)*, where irrelevant degrees of freedom of the statistical model are automatically pruned out. Explaining the ARD property of approximate Bayesian learning is one of the main topics of theoretical analysis in Parts III and IV.

1.2 Computation

Now, let us explain how Bayesian learning is performed in simple cases. We start from introducing *conjugacy*, an important notion in performing Bayesian learning.

1.2.1 Popular Distributions

Table 1.1 summarizes several distributions that are frequently used as a model distribution (or likelihood function) $p(\mathcal{D}|\boldsymbol{w})$ or a prior distribution $p(\boldsymbol{w})$ in Bayesian learning. The domain $\mathcal{X}$ of the random variable $\boldsymbol{x}$ and the domain $\mathcal{W}$ of the parameters $\boldsymbol{w}$ are shown in the table.

Some of the distributions in Table 1.1 have complicated function forms, involving Beta or Gamma functions. However, such complications are mostly in the *normalization constant*, and can often be ignored when it is sufficient to find the *shape* of a function. In Table 1.1, the normalization constant is separated by a dot, so that one can find the simple main part. As will be seen shortly, we often refer to the normalization constant when we need to perform integration of a function, which is in the same form as the main part of a popular distribution.

Below we summarize abbreviations of distributions:

$$\mathrm{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{1.16}$$

$$\mathrm{Gamma}(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \exp(-\beta x), \tag{1.17}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Table 1.1 *Popular distributions. The following notation is used:* $\mathbb{R}$ *: The set of all real numbers,* $\mathbb{R}_{++}$ *: The set of all positive real numbers,* $\mathbb{I}_{++}$ *: The set of all positive integers,* $\mathbb{S}_{++}^{M}$ *: The set of all* $M \times M$ *positive definite matrices,* $\mathbb{H}_{N}^{K-1} \equiv \{\boldsymbol{x} \in \{0,\ldots,N\}^{K}; \sum_{k=1}^{K} x_k = N\}$ *: The set of all possible histograms for* N *samples and* K *categories,* $\Delta^{K-1} \equiv \{\boldsymbol{\theta} \in [0,1]^{K}; \sum_{k=1}^{K} \theta_k = 1\}$ *: The standard* $(K-1)$*-simplex,* $\det(\cdot)$ *:Determinant of matrix,* $\mathcal{B}(y,z) \equiv \int_0^1 t^{y-1}(1-t)^{z-1}dt$ *: Beta function,* $\Gamma(y) \equiv \int_0^\infty t^{y-1}\exp(-t)dt$ *: Gamma function, and* $\Gamma_M(y) \equiv \int_{\boldsymbol{T}\in\mathbb{S}_{++}^{M}} \det(\boldsymbol{T})^{y-(M+1)/2}\exp(-\mathrm{tr}(\boldsymbol{T}))d\boldsymbol{T}$ *: Multivariate Gamma function.*

Probability distribution	$p(\boldsymbol{x}\|\boldsymbol{w})$	$\boldsymbol{x} \in \mathcal{X}$	$\boldsymbol{w} \in \mathcal{W}$
Isotropic Gaussian	$\mathrm{Gauss}_M(\boldsymbol{x};\boldsymbol{\mu},\sigma^2\boldsymbol{I}_M) \equiv \frac{1}{(2\pi\sigma^2)^{M/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}\|^2\right)$	$\boldsymbol{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \sigma^2 > 0$
Gaussian	$\mathrm{Gauss}_M(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$	$\boldsymbol{x} \in \mathbb{R}^M$	$\boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\Sigma} \in \mathbb{S}_{++}^M$
Gamma	$\mathrm{Gamma}(x;\alpha,\beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1}\exp(-\beta x)$	$x > 0$	$\alpha > 0, \beta > 0$
Wishart	$\mathrm{Wishart}_M(\boldsymbol{X};\boldsymbol{V},\nu) \equiv \frac{1}{(2^\nu\|\boldsymbol{V}\|)^{M/2}\Gamma_M(\frac{\nu}{2})} \cdot \det(\boldsymbol{X})^{\frac{\nu-M-1}{2}} \exp\left(-\frac{\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right)$	$\boldsymbol{X} \in \mathbb{S}_{++}^M$	$\boldsymbol{V} \in \mathbb{S}_{++}^M, \nu > M-1$
Bernoulli	$\mathrm{Binomial}_1(x;\theta) \equiv \theta^x(1-\theta)^{1-x}$	$x \in \{0,1\}$	$\theta \in [0,1]$
Binomial	$\mathrm{Binomial}_N(x;\theta) \equiv \binom{N}{x}\theta^x(1-\theta)^{N-x}$	$x \in \{0,\ldots,N\}$	$\theta \in [0,1]$
Multinomial	$\mathrm{Multinomial}_{K,N}(\boldsymbol{x};\boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^K (x_k!)^{-1}\theta_k^{x_k}$	$\boldsymbol{x} \in \mathbb{H}_N^{K-1}$	$\boldsymbol{\theta} \in \Delta^{K-1}$
Beta	$\mathrm{Beta}(x;a,b) \equiv \frac{1}{\mathcal{B}(a,b)} \cdot x^{a-1}(1-x)^{b-1}$	$x \in [0,1]$	$a > 0, b > 0$
Dirichlet	$\mathrm{Dirichlet}_K(\boldsymbol{x};\boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)} \cdot \prod_{k=1}^K x_k^{\phi_{k-1}}$	$\boldsymbol{x} \in \Delta^{K-1}$	$\boldsymbol{\phi} \in \mathbb{R}_{++}^K$

$$\text{Wishart}_M(\boldsymbol{X};\boldsymbol{V},\nu) \equiv \frac{1}{(2^\nu|\boldsymbol{V}|)^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)} \cdot \det(\boldsymbol{X})^{\frac{\nu-M-1}{2}} \exp\left(-\frac{\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{X})}{2}\right), \tag{1.18}$$

$$\text{Binomial}_N(x;\theta) \equiv \binom{N}{x} \cdot \theta^x(1-\theta)^{N-x}, \tag{1.19}$$

$$\text{Multinomial}_{K,N}(\boldsymbol{x};\boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^{K}(x_k!)^{-1}\theta_k^{x_k}, \tag{1.20}$$

$$\text{Beta}(x;a,b) \equiv \frac{1}{\mathcal{B}(a,b)} \cdot x^{a-1}(1-x)^{b-1}, \tag{1.21}$$

$$\text{Dirichlet}_K(\boldsymbol{x};\boldsymbol{\phi}) \equiv \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)} \cdot \prod_{k=1}^{K} x_k^{\phi_{k-1}}. \tag{1.22}$$

The distributions in Table 1.1 are categorized into four groups, which are separated by dashed lines. In each group, an upper distribution family is a special case of a lower distribution family. Note that the following hold:

$$\begin{aligned}
\text{Gamma}(x;\alpha,\beta) &= \text{Wishart}_1\left(x;\frac{1}{2\beta},2\alpha\right),\\
\text{Binomial}_N(x;\theta) &= \text{Multinomial}_{2,N}\left((x,N-x)^\top;(\theta,1-\theta)^\top\right),\\
\text{Beta}(x;a,b) &= \text{Dirichlet}_2\left((x,1-x)^\top;(a,b)^\top\right).
\end{aligned}$$

1.2.2 Conjugacy

Let us think about the *function form* of the posterior (1.3):

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}),$$

which is determined by the function form of the product of the model likelihood $p(\mathcal{D}|\boldsymbol{w})$ and the prior $p(\boldsymbol{w})$. Note that we here call the conditional $p(\mathcal{D}|\boldsymbol{w})$ NOT the *model distribution* but the *model likelihood*, since we are interested in the function form of the posterior, a distribution of the parameter $\boldsymbol{w}$.

Conjugacy is defined as the relation between the likelihood $p(\mathcal{D}|\boldsymbol{w})$ and the prior $p(\boldsymbol{w})$.

Definition 1.4 (Conjugate prior) A prior $p(\boldsymbol{w})$ is called *conjugate* with a likelihood $p(\mathcal{D}|\boldsymbol{w})$, if the posterior $p(\boldsymbol{w}|\mathcal{D})$ is in the same distribution family as the prior.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

1.2.3 Posterior Distribution

Here, we introduce computation of the posterior distribution in simple cases where a conjugate prior exists and is adopted.

Isotropic Gaussian Model

Let us compute the posterior distribution for the isotropic Gaussian model:

$$p(\boldsymbol{x}|\boldsymbol{w}) = \mathrm{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_M) = \frac{1}{(2\pi\sigma^2)^{M/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2\right). \quad (1.23)$$

The likelihood for N i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ is written as

$$p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}) = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2\right)}{(2\pi\sigma^2)^{MN/2}}. \quad (1.24)$$

Gaussian Likelihood As noted in Section 1.2.2, we should see Eq. (1.24), which is the distribution of observed data $\mathcal{D}$, as a function of the parameter $\boldsymbol{w}$. Naturally, the function form depends on which parameters are estimated in the *Bayesian way*. The isotropic Gaussian has two parameters $\boldsymbol{w} = (\boldsymbol{\mu}, \sigma^2)$, and we first consider the case where the variance parameter σ^2 is known, and the posterior of the mean parameter $\boldsymbol{\mu}$ is estimated, i.e., we set $\boldsymbol{w} = \boldsymbol{\mu}$. This case contains the case where σ^2 is unknown but point-estimated in the empirical Bayesian procedure or tuned outside the Bayesian framework, e.g., by performing cross-validation (we set $\boldsymbol{w} = \boldsymbol{\mu}, \kappa = \sigma^2$ in the latter case).

Omitting the constant (with respect to $\boldsymbol{\mu}$), the likelihood (1.24) can be written as

$$\begin{aligned}
p(\mathcal{D}|\boldsymbol{\mu}) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\|(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}) + (\overline{\boldsymbol{x}} - \boldsymbol{\mu})\|^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\|^2 + N\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2\right)\right) \\
&\propto \exp\left(-\frac{N}{2\sigma^2}\left\|\boldsymbol{\mu} - \overline{\boldsymbol{x}}\right\|^2\right) \\
&\propto \mathrm{Gauss}_M\left(\boldsymbol{\mu}; \overline{\boldsymbol{x}}, \frac{\sigma^2}{N}\boldsymbol{I}_M\right), \quad (1.25)
\end{aligned}$$

where $\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}^{(n)}$ is the *sample mean*. Note that we omitted the factor $\exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\|^2\right)$ as a constant in the fourth equation.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

The last equation (1.25) implies that, as a function of the mean parameter $\boldsymbol{\mu}$, the model likelihood $p(\mathcal{D}|\boldsymbol{\mu})$ has the same form as the isotropic Gaussian with mean $\overline{\boldsymbol{x}}$ and variance $\frac{\sigma^2}{N}$. Eq. (1.25) also implies that the ML estimator for the mean parameter is given by

$$\widehat{\boldsymbol{\mu}}^{\mathrm{ML}} = \overline{\boldsymbol{x}}.$$

Thus, we found that the likelihood function for the mean parameter of the isotropic Gaussian is in the Gaussian form. This comes from the following facts:

- The isotropic Gaussian model for a single sample $\boldsymbol{x}$ is in the Gaussian form also as a function of the mean parameter, i.e., $\mathrm{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_M) \propto \mathrm{Gauss}_M(\boldsymbol{\mu}; \boldsymbol{x}, \sigma^2 \boldsymbol{I}_M)$.
- The isotropic Gaussians are *multiplicatively closed*, i.e., the product of isotropic Gaussians with different means is a Gaussian: $p(\mathcal{D}|\boldsymbol{\mu}) \propto \prod_{n=1}^{N} \mathrm{Gauss}_M(\boldsymbol{\mu}; \boldsymbol{x}^{(n)}, \sigma^2 \boldsymbol{I}_M) \propto \mathrm{Gauss}_M\left(\boldsymbol{\mu}; \overline{\boldsymbol{x}}, \frac{\sigma^2}{N} \boldsymbol{I}_M\right)$.

Since the isotropic Gaussian is multiplicatively closed and the likelihood (1.25) is in the Gaussian form, the isotropic Gaussian prior must be conjugate. Let us choose the isotropic Gaussian prior,

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \sigma_0^2) = \mathrm{Gauss}_M(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{I}_M) \propto \exp\left(-\frac{1}{2\sigma_0^2} \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2\right),$$

for hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\mu}_0, \sigma_0^2)$. Then, the function form of the posterior is given by

$$\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\mu}_0, \sigma_0^2) &\propto p(\mathcal{D}|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \sigma_0^2) \\
&\propto \mathrm{Gauss}_M\left(\boldsymbol{\mu}; \overline{\boldsymbol{x}}, \frac{\sigma^2}{N}\right) \mathrm{Gauss}_M(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma_0^2) \\
&\propto \exp\left(-\frac{N}{2\sigma^2} \left\|\boldsymbol{\mu} - \overline{\boldsymbol{x}}\right\|^2 - \frac{1}{2\sigma_0^2} \left\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\right\|^2\right) \\
&\propto \exp\left(-\frac{N\sigma^{-2} + \sigma_0^{-2}}{2} \left\|\boldsymbol{\mu} - \frac{N\sigma^{-2}\overline{\boldsymbol{x}} + \sigma_0^{-2}\boldsymbol{\mu}_0}{N\sigma^{-2} + \sigma_0^{-2}}\right\|^2\right) \\
&\propto \mathrm{Gauss}_M\left(\boldsymbol{\mu}; \frac{N\sigma^{-2}\overline{\boldsymbol{x}} + \sigma_0^{-2}\boldsymbol{\mu}_0}{N\sigma^{-2} + \sigma_0^{-2}}, \frac{1}{N\sigma^{-2} + \sigma_0^{-2}}\right).
\end{aligned}$$

Therefore, the posterior is

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\mu}_0, \sigma_0^2) = \mathrm{Gauss}_M\left(\boldsymbol{\mu}; \frac{N\sigma^{-2}\overline{\boldsymbol{x}} + \sigma_0^{-2}\boldsymbol{\mu}_0}{N\sigma^{-2} + \sigma_0^{-2}}, \frac{1}{N\sigma^{-2} + \sigma_0^{-2}}\right). \tag{1.26}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Note that the *equality* holds in Eq. (1.26). We omitted constant factors in the preceding derivation. But once the function form of the posterior is found, the normalization factor is unique. If the function form coincides with one of the well-known distributions (e.g., ones given in Table 1.1), one can find the normalization constant (from the table) without any further computation.

Multiplicative closedness of a function family of the model likelihood is essential in performing Bayesian learning. Such families are called the *exponential family*:

Definition 1.5 (Exponential families) A family of distributions is called the exponential family if it is written as

$$p(\boldsymbol{x}|\boldsymbol{w}) = p(\boldsymbol{t}|\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^\top \boldsymbol{t} - A(\boldsymbol{\eta}) + B(\boldsymbol{t})\right), \tag{1.27}$$

where $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is a function, called *sufficient statistics*, of the random variable $\boldsymbol{x}$, and $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{w})$ is a function, called *natural parameters*, of the parameter $\boldsymbol{w}$.

The essential property of the exponential family is that the interaction between the random variable and the parameter occurs only in the log linear form, i.e., $\exp(\boldsymbol{\eta}^\top \boldsymbol{t})$. Note that, although $A(\cdot)$ and $B(\cdot)$ are arbitrary functions, $A(\cdot)$ does not depend on $\boldsymbol{t}$, and $B(\cdot)$ does not depend on $\boldsymbol{\eta}$.

Assume that N observed samples $\mathcal{D} = (\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(N)}) = (\boldsymbol{t}(\boldsymbol{x}^{(1)}), \ldots, \boldsymbol{t}(\boldsymbol{x}^{(N)}))$ are drawn from the exponential family distribution (1.27). If we use the exponential family prior $p(\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^\top \boldsymbol{t}^{(0)} - A_0(\boldsymbol{\eta}) + B_0(\boldsymbol{t}^{(0)})\right)$, then the posterior is given as an exponential family distribution with the same set of natural parameters $\boldsymbol{\eta}$:

$$p(\boldsymbol{\eta}|\mathcal{D}) = \exp\left(\boldsymbol{\eta}^\top \sum_{n=0}^{N} \boldsymbol{t}^{(n)} - A'(\boldsymbol{\eta}) + B'(\mathcal{D})\right),$$

where $A'(\boldsymbol{\eta})$ and $B'(\mathcal{D})$ are a function of $\boldsymbol{\eta}$ and a function of $\mathcal{D}$, respectively. Therefore, the conjugate prior for the exponential family distribution is the exponential family with the same natural parameters $\boldsymbol{\eta}$.

All distributions given in Table 1.1 are exponential families. For example, the sufficient statistics and the natural parameters for the univariate Gaussian are given by $\boldsymbol{\eta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^\top$ and $\boldsymbol{t} = (x, x^2)^\top$, respectively. The mixture model (1.11) is a common *non*exponential family distribution.

Gamma Likelihood Next we consider the posterior distribution of the variance parameter σ^2 with the mean parameter regarded as a constant, i.e., $w = \sigma^2$.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Omitting the constants (with respect to σ^2) of the model likelihood (1.24), we have

$$p(\mathcal{D}|\sigma^2) \propto (\sigma^2)^{-MN/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2\right).$$

If we see the likelihood as a function of the *inverse* of σ^2, we find that it is proportional to the *Gamma distribution*:

$$\begin{aligned} p(\mathcal{D}|\sigma^{-2}) &\propto (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2\right)\sigma^{-2}\right) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2}+1, \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2\right). \end{aligned} \tag{1.28}$$

Since the mode of the Gamma distribution is known as $\text{argmax}_x \text{Gamma}(x;\alpha,\beta) = \frac{\alpha-1}{\beta}$, Eq. (1.28) implies that the ML estimator for the variance parameter is given by

$$\widehat{\sigma^2}^{\text{ML}} = \frac{1}{\widehat{\sigma^{-2}}^{\text{ML}}} = \frac{\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2}{\frac{MN}{2}+1-1} = \frac{1}{MN}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2.$$

Now we found that the model likelihood of the isotropic Gaussian is in the Gamma form as a function of the inverse variance σ^{-2}. Since the Gamma distribution is in the exponential family and multiplicatively closed, the Gamma prior is conjugate.

If we use the Gamma prior

$$p(\sigma^{-2}|\alpha_0,\beta_0) = \text{Gamma}(\sigma^{-2};\alpha_0,\beta_0) \propto (\sigma^{-2})^{\alpha_0-1}\exp(-\beta_0\sigma^{-2})$$

with hyperparameters $\boldsymbol{\kappa} = (\alpha_0,\beta_0)$, the posterior can be written as

$$\begin{aligned} p(\sigma^{-2}|\mathcal{D},\alpha_0,\beta_0) &\propto p(\mathcal{D}|\sigma^{-2})p(\sigma^{-2}|\alpha_0,\beta_0) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2}+1, \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2\right)\text{Gamma}(\sigma^{-2};\alpha_0,\beta_0) \\ &\propto (\sigma^{-2})^{MN/2+\alpha_0-1}\exp\left(-\left(\frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2+\beta_0\right)\sigma^{-2}\right), \end{aligned}$$

and therefore

$$p(\sigma^{-2}|\mathcal{D},\alpha_0,\beta_0) = \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2}+\alpha_0, \frac{1}{2}\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}\|^2+\beta_0\right). \tag{1.29}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Isotropic Gauss-Gamma Likelihood Finally, we consider the general case where both the mean and variance parameters are unknown, i.e., $\boldsymbol{w} = (\boldsymbol{\mu}, \sigma^{-2})$. The likelihood is written as

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\mu}, \sigma^{-2}) &\propto (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{1}{2}\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2\right)\sigma^{-2}\right) \\
&= (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{N\|\boldsymbol{\mu} - \overline{\boldsymbol{x}}\|^2}{2} + \frac{\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\|^2}{2}\right)\sigma^{-2}\right) \\
&\propto \text{GaussGamma}_M\left(\boldsymbol{\mu}, \sigma^{-2} \middle| \overline{\boldsymbol{x}}, N\boldsymbol{I}_M, \frac{M(N-1)}{2} + 1, \frac{\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\|^2}{2}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
&\text{GaussGamma}_M(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \lambda\boldsymbol{I}_M, \alpha, \beta) \\
&\quad\equiv \text{Gauss}_M(\boldsymbol{x}|\boldsymbol{\mu}, (\tau\lambda)^{-1}\boldsymbol{I}_M) \cdot \text{Gamma}(\tau|\alpha, \beta) \\
&\quad= \frac{\exp\left(-\frac{\tau\lambda}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|^2\right)}{(2\pi(\tau\lambda)^{-1})^{M/2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau) \\
&\quad= \frac{\beta^\alpha}{(2\pi/\lambda)^{M/2}\Gamma(\alpha)} \tau^{\alpha+\frac{M}{2}-1} \exp\left(-\left(\frac{\lambda\|\boldsymbol{x} - \boldsymbol{\mu}\|^2}{2} + \beta\right)\tau\right)
\end{aligned}
$$

is the *isotropic Gauss-Gamma distribution* on the random variable $\boldsymbol{x} \in \mathbb{R}^M$, $\tau > 0$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^M, \lambda > 0, \alpha > 0, \beta > 0$.

Note that, although the isotropic Gauss-Gamma distribution is the product of an isotropic Gaussian distribution and a Gamma distribution, the random variables $\boldsymbol{x}$ and τ are not independent of each other. This is because the isotropic Gauss-Gamma distribution is a *hierarchical model* $p(\boldsymbol{x}|\tau)p(\tau)$, where the variance parameter $\sigma^2 = (\tau\lambda)^{-1}$ for the isotropic Gaussian depends on the random variable τ of the Gamma distribution.

Since the isotropic Gauss-Gamma distribution is multiplicatively closed, it is a conjugate prior. Choosing the isotropic Gauss-Gamma prior

$$
\begin{aligned}
p(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0, \alpha_0, \beta_0) &= \text{GaussGamma}_M(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0\boldsymbol{I}_M, \alpha_0, \beta) \\
&\propto (\sigma^{-2})^{\alpha_0+\frac{M}{2}-1} \exp\left(-\left(\frac{\lambda_0\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{2} + \beta_0\right)\sigma^{-2}\right)
\end{aligned}
$$

with hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\mu}_0, \lambda_0, \alpha_0, \beta_0)$, the posterior is given by

$$
\begin{aligned}
p(\boldsymbol{\mu}, \sigma^{-2}|\mathcal{D}, \boldsymbol{\kappa}) &\propto p(\mathcal{D}|\boldsymbol{\mu}, \sigma^{-2})p(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\kappa}) \\
&\propto \text{GaussGamma}_M\left(\boldsymbol{\mu}, \sigma^{-2} \middle| \overline{\boldsymbol{x}}, N\boldsymbol{I}_M, \frac{M(N-1)}{2} + 1, \frac{\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\|^2}{2}\right) \\
&\quad \cdot \text{GaussGamma}_M(\boldsymbol{\mu}, \sigma^{-2}|\boldsymbol{\mu}_0, \lambda_0\boldsymbol{I}_M, \alpha_0, \beta)
\end{aligned}
$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$
\begin{aligned}
&\propto (\sigma^{-2})^{MN/2} \exp\left(-\left(\frac{N\|\boldsymbol{\mu}-\overline{\boldsymbol{x}}\|^2}{2}+\frac{\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}\|^2}{2}\right)\sigma^{-2}\right)\\
&\qquad\cdot(\sigma^{-2})^{\alpha_0+\frac{M}{2}-1}\exp\left(-\left(\frac{\lambda_0\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2}{2}+\beta_0\right)\sigma^{-2}\right)\\
&\propto (\sigma^{-2})^{M(N+1)/2+\alpha_0-1}\\
&\qquad\cdot\exp\left(-\left(\frac{N\|\boldsymbol{\mu}-\overline{\boldsymbol{x}}\|^2+\lambda_0\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2}{2}+\frac{\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}\|^2}{2}+\beta_0\right)\sigma^{-2}\right)\\
&\propto (\sigma^{-2})^{\widehat{\alpha}+\frac{M}{2}-1}\exp\left(-\left(\frac{\widehat{\lambda}\left\|\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\right\|^2}{2}+\widehat{\beta}\right)\sigma^{-2}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \frac{N\overline{\boldsymbol{x}}+\lambda_0\boldsymbol{\mu}_0}{N+\lambda_0},\\
\widehat{\lambda} &= N+\lambda_0,\\
\widehat{\alpha} &= \frac{MN}{2}+\alpha_0,\\
\widehat{\beta} &= \frac{\sum_{n=1}^{N}\|\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}\|^2}{2}+\frac{N\lambda_0\|\overline{\boldsymbol{x}}-\boldsymbol{\mu}_0\|^2}{2(N+\lambda_0)}+\beta_0.
\end{aligned}
$$

Thus, the posterior is obtained as

$$
p(\boldsymbol{\mu},\sigma^{-2}|\mathcal{D},\boldsymbol{\kappa}) = \mathrm{GaussGamma}_M(\boldsymbol{\mu},\sigma^{-2}|\widehat{\boldsymbol{\mu}},\widehat{\lambda}\boldsymbol{I}_M,\widehat{\alpha},\widehat{\beta}). \tag{1.30}
$$

Although the Gauss-Gamma distribution seems a bit more complicated than the ones in Table 1.1, its moments are known. Therefore, Bayesian learning with a conjugate prior can be analytically performed also when both parameters $\boldsymbol{w} = (\boldsymbol{\mu},\sigma^{-2})$ are estimated.

Gaussian Model

Bayesian learning can be performed for a general Gaussian model in a similar fashion to the isotropic case. Consider the M-dimensional Gaussian distribution,

$$
p(\boldsymbol{x}|\boldsymbol{w}) = \mathrm{Gauss}_M(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}}\cdot\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1.31}
$$

with mean and covariance parameters $\boldsymbol{w} = (\boldsymbol{\mu},\boldsymbol{\Sigma})$. The likelihood for N i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(N)}\}$ is written as

$$
p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\boldsymbol{w}) = \frac{\exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})\right)}{(2\pi)^{NM/2}\det(\boldsymbol{\Sigma})^{N/2}}. \tag{1.32}
$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Gaussian Likelihood Let us first compute the posterior distribution on the mean parameter $\boldsymbol{\mu}$, with the covariance parameter regarded as a known constant. In this case, the likelihood can be written as

$$\begin{aligned}
p(\mathcal{D}|\boldsymbol{\mu}) &\propto \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\left((\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)^{\top}\cdot\boldsymbol{\Sigma}^{-1}\left((\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})+N(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)\right) \\
&\propto \exp\left(-\frac{N}{2}(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}})\right) \\
&\propto \mathrm{Gauss}_M\left(\boldsymbol{\mu};\overline{\boldsymbol{x}},\frac{1}{N}\boldsymbol{\Sigma}\right). \qquad (1.33)
\end{aligned}$$

Therefore, with the conjugate Gaussian prior

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)=\mathrm{Gauss}_M(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\propto\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\top}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right),$$

with hyperparameters $\boldsymbol{\kappa}=(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$, the posterior is written as

$$\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) &\propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&\propto \mathrm{Gauss}_M\left(\boldsymbol{\mu};\overline{\boldsymbol{x}},\frac{1}{N}\boldsymbol{\Sigma}\right)\mathrm{Gauss}_M(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&\propto \exp\left(-\frac{N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}})+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\top}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)}{2}\right) \\
&\propto \exp\left(-\frac{(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^{\top}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})}{2}\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\mu}} &= \left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right), \\
\widehat{\boldsymbol{\Sigma}} &= \left(N\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1}.
\end{aligned}$$

Thus, we have

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)=\mathrm{Gauss}_M\left(\boldsymbol{\mu};\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right). \qquad (1.34)$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Wishart Likelihood If we see the mean parameter $\boldsymbol{\mu}$ as a given constant, the model likelihood (1.32) can be written as follows, as a function of the covariance parameter $\boldsymbol{\Sigma}$:

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\Sigma}^{-1}) &\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})}{2}\right) \\ &\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\operatorname{tr}\left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\ &\propto \text{Wishart}_M\left(\boldsymbol{\Sigma}^{-1}; \left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top\right)^{-1}, M+N+1\right). \end{aligned}$$

Here, as in the isotropic Gaussian case, we computed the distribution on the *inverse* $\boldsymbol{\Sigma}^{-1}$ of the covariance parameter. With the *Wishart distribution*

$$\begin{aligned} p(\boldsymbol{\Sigma}^{-1}|\boldsymbol{V}_0, \nu_0) &= \text{Wishart}_M(\boldsymbol{\Sigma}^{-1}; \boldsymbol{V}_0, \nu_0) \\ &= \frac{1}{(2^{\nu_0}\det(\boldsymbol{V}_0))^{M/2}\,\Gamma_M\left(\frac{\nu_0}{2}\right)} \cdot \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\nu_0-M-1}{2}} \exp\left(-\frac{\operatorname{tr}(\boldsymbol{V}_0^{-1}\boldsymbol{\Sigma}^{-1})}{2}\right) \end{aligned}$$

for hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{V}_0, \nu_0)$ as a conjugate prior, the posterior is computed as

$$\begin{aligned} p(\boldsymbol{\Sigma}^{-1}|\mathcal{D}, \boldsymbol{V}_0, \nu_0) &\propto p(\mathcal{D}|\boldsymbol{\Sigma}^{-1}) p(\boldsymbol{\Sigma}^{-1}|\boldsymbol{V}_0, \nu_0) \\ &\propto \text{Wishart}_M\left(\boldsymbol{\Sigma}^{-1}; \left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top\right)^{-1}, M+N+1\right) \\ &\qquad \cdot \text{Wishart}_M(\boldsymbol{\Sigma}^{-1}; \boldsymbol{V}_0, \nu_0) \\ &\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{N}{2}} \exp\left(-\frac{\operatorname{tr}\left(\left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top\right)\boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\ &\qquad \cdot \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\nu_0-M-1}{2}} \exp\left(-\frac{\operatorname{tr}\left(\boldsymbol{V}_0^{-1}\boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\ &\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\nu_0-M+N-1}{2}} \exp\left(-\frac{\operatorname{tr}\left(\left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top + \boldsymbol{V}_0^{-1}\right)\boldsymbol{\Sigma}^{-1}\right)}{2}\right). \end{aligned}$$

Thus we have

$$\begin{aligned} &p(\boldsymbol{\Sigma}^{-1}|\mathcal{D}, \boldsymbol{V}_0, \nu_0) \\ &\quad = \text{Wishart}_M\left(\boldsymbol{\Sigma}^{-1}; \left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top + \boldsymbol{V}_0^{-1}\right)^{-1}, N+\nu_0\right). \quad (1.35) \end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Note that the Wishart distribution can be seen as a multivariate extension of the Gamma distribution and is reduced to the Gamma distribution for $M = 1$:

$$\text{Wishart}_1\,(x; V, \nu) = \text{Gamma}\,(x; \nu/2, 1/(2V))\,.$$

Gauss-Wishart Likelihood When both parameters $\boldsymbol{w} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ are unknown, the model likelihood (1.32) is seen as

$$\begin{aligned}
p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) &\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})}{2}\right) \\
&\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\text{tr}\left(\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\
&\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\text{tr}\left(\sum_{n=1}^{N}\left((\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)\left((\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})+(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)^\top \boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\
&\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\text{tr}\left(N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top + \sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})^\top)\boldsymbol{\Sigma}^{-1}\right)}{2}\right) \\
&\propto \text{GaussWishart}_M\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}; \overline{\boldsymbol{x}}, N, \left(\textstyle\sum_{n=1}^{N}(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})^\top\right)^{-1}, M + N\Big),
\end{aligned}$$

where

$$\begin{aligned}
&\text{GaussWishart}_M(\boldsymbol{x}, \boldsymbol{\Lambda}|\boldsymbol{\mu}, \lambda, \boldsymbol{V}, \nu) \\
&\quad\equiv \text{Gauss}_M(\boldsymbol{x}|\boldsymbol{\mu}, (\lambda\boldsymbol{\Lambda})^{-1})\text{Wishart}_M(\boldsymbol{\Lambda}|\boldsymbol{V}, \nu) \\
&\quad= \frac{\exp\left(-\frac{\lambda}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}\,(\boldsymbol{x}-\boldsymbol{\mu})\right)}{(2\pi)^{M/2}\det(\lambda\boldsymbol{\Lambda})^{-1/2}} \cdot \frac{\det(\boldsymbol{\Lambda})^{\frac{\nu-M-1}{2}} \exp\left(-\frac{\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{\Lambda})}{2}\right)}{(2^\nu \det(\boldsymbol{V}))^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)} \\
&\quad= \frac{\lambda^{M/2}}{(2^{\nu+1}\pi\det(\boldsymbol{V}))^{M/2}\Gamma_M\left(\frac{\nu}{2}\right)} \det(\boldsymbol{\Lambda})^{\frac{\nu-M}{2}} \exp\left(-\frac{\text{tr}\left((\lambda(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top + \boldsymbol{V}^{-1})\boldsymbol{\Lambda}\right)}{2}\right)
\end{aligned}$$

is the *Gauss–Wishart distribution* on the random variables $\boldsymbol{x} \in \mathbb{R}^M, \boldsymbol{\Lambda} \in \mathbb{S}_{++}^M$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^M, \lambda > 0, \boldsymbol{V} \in \mathbb{S}_{++}^M, \nu > M - 1$.

With the conjugate Gauss–Wishart prior,

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\boldsymbol{\mu}_0, \lambda_0, \alpha_0, \beta_0) &= \text{GaussWishart}_M(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{V}_0, \nu_0) \\
&\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\nu-M}{2}} \exp\left(-\frac{\text{tr}\left((\lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top + \boldsymbol{V}_0^{-1})\boldsymbol{\Sigma}^{-1}\right)}{2}\right)
\end{aligned}$$

with hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{V}_0, \nu_0)$, the posterior is written as

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\mathcal{D}, \boldsymbol{\kappa}) &\propto p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\boldsymbol{\kappa}) \\
&\propto \text{GaussWishart}_M\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}; \overline{\boldsymbol{x}}, N, \left(\textstyle\sum_{n=1}^{N}(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})^\top\right)^{-1}, M + N\Big) \\
&\qquad \cdot \text{GaussWishart}_M(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{V}_0, \nu_0)
\end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{N/2} \exp\left(-\frac{\mathrm{tr}\left(N(\boldsymbol{\mu}-\overline{\boldsymbol{x}})(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top + \sum_{n=1}^N (\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})^\top)\boldsymbol{\Sigma}^{-1}\right)}{2}\right)$$
$$\cdot \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\nu_0-M}{2}} \exp\left(-\frac{\mathrm{tr}\left((\lambda_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top + \boldsymbol{V}_0^{-1})\boldsymbol{\Sigma}^{-1}\right)}{2}\right)$$
$$\propto \det\left(\boldsymbol{\Sigma}^{-1}\right)^{\frac{\widehat{\nu}-M}{2}} \exp\left(-\mathrm{tr}\left(\frac{\left(\widehat{\lambda}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{V}}^{-1}\right)\boldsymbol{\Sigma}^{-1}}{2}\right)\right),$$

where

$$\widehat{\boldsymbol{\mu}} = \frac{N\overline{\boldsymbol{x}} + \lambda_0\boldsymbol{\mu}_0}{N+\lambda_0},$$
$$\widehat{\lambda} = N + \lambda_0,$$
$$\widehat{\boldsymbol{V}} = \left(\sum_{n=1}^N (\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}})^\top + \frac{N\lambda_0}{N+\lambda_0}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}_0)(\overline{\boldsymbol{x}}-\boldsymbol{\mu}_0)^\top + \boldsymbol{V}_0^{-1}\right)^{-1},$$
$$\widehat{\nu} = N + \nu_0.$$

Thus, we have the posterior distribution as the Gauss–Wishart distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\mathcal{D}, \boldsymbol{\kappa}) = \mathrm{GaussWishart}_M\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}|\widehat{\boldsymbol{\mu}}, \widehat{\lambda}, \widehat{\boldsymbol{V}}, \widehat{\nu}\right). \tag{1.36}$$

Linear Regression Model

Consider the *linear regression model*, where an input variable $\boldsymbol{x} \in \mathbb{R}^M$ and an output variable $y \in \mathbb{R}$ are assumed to satisfy the following probabilistic relation:

$$y = \boldsymbol{a}^\top \boldsymbol{x} + \varepsilon, \tag{1.37}$$
$$p(\varepsilon|\sigma^2) = \mathrm{Gauss}_1(\varepsilon; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \tag{1.38}$$

Here $\boldsymbol{a}$ and σ^2 are called the *regression parameter* and the *noise variance parameter*, respectively. By substituting $\varepsilon = y - \boldsymbol{a}^\top \boldsymbol{x}$, which is obtained from Eq. (1.37), into Eq. (1.38), we have

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathrm{Gauss}_1(y; \boldsymbol{a}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y-\boldsymbol{a}^\top \boldsymbol{x})^2}{2\sigma^2}\right).$$

The likelihood function for N observed i.i.d.[2] samples,

$$\mathcal{D} = (\boldsymbol{y}, \boldsymbol{X}),$$

[2] In the context of regression, i.i.d. usually means that the observation noise $\varepsilon^{(n)} = y^{(n)} - \boldsymbol{a}^\top \boldsymbol{x}^{(n)}$ is independent for different samples, i.e., $p(\{\varepsilon^{(n)}\}_{n=1}^N) = \prod_{n=1}^N p(\varepsilon^{(n)})$, and the independence between the input $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})$, i.e., $p(\{\boldsymbol{x}^{(n)}\}_{n=1}^N) = \prod_{n=1}^N p(\boldsymbol{x}^{(n)})$, is not required.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

is given by

$$p(\mathcal{D}|\boldsymbol{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2}{2\sigma^2}\right), \tag{1.39}$$

where we defined

$$\boldsymbol{y} = (y^{(1)}, \ldots, y^{(N)})^\top \in \mathbb{R}^N, \quad \boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})^\top \in \mathbb{R}^{N \times M}.$$

Gaussian Likelihood The computation of the posterior is similar to the isotropic Gaussian case. As in Section 1.2.3, we first consider the case where only the regression parameter $\boldsymbol{a}$ is estimated, with the noise variance parameter σ^2 regarded as a known constant.

One can guess that the likelihood (1.39) is Gaussian as a function of $\boldsymbol{a}$, since it is an exponential of a concave quadratic function. Indeed, by expanding the exponent and completing the square for $\boldsymbol{a}$, we obtain

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{a}) &\propto \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{(\boldsymbol{a}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{a}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})}{2\sigma^2}\right) \\ &\propto \mathrm{Gauss}_M\left(\boldsymbol{a}; (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right). \end{aligned} \tag{1.40}$$

Eq. (1.40) implies that, when $\boldsymbol{X}^\top\boldsymbol{X}$ is *nonsingular* (i.e., its inverse exists), the ML estimator for $\boldsymbol{a}$ is given by

$$\widehat{\boldsymbol{a}}^{\mathrm{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{1.41}$$

Therefore, with the conjugate Gaussian prior

$$p(\boldsymbol{a}|\boldsymbol{a}_0, \boldsymbol{\Sigma}_0) = \mathrm{Gauss}_M(\boldsymbol{a}; \boldsymbol{a}_0, \boldsymbol{\Sigma}_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{a}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{a} - \boldsymbol{a}_0)\right)$$

for hyperparameters $\boldsymbol{\kappa} = (\boldsymbol{a}_0, \boldsymbol{\Sigma}_0)$, the posterior is Gaussian:

$$\begin{aligned} p(\boldsymbol{a}|\mathcal{D}, \boldsymbol{a}_0, \boldsymbol{\Sigma}_0) &\propto p(\mathcal{D}|\boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{a}_0, \boldsymbol{\Sigma}_0) \\ &\propto \mathrm{Gauss}_M\left(\boldsymbol{a}; \boldsymbol{a}_0, \tfrac{1}{N}\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)\mathrm{Gauss}_M(\boldsymbol{a}; \boldsymbol{a}_0, \boldsymbol{\Sigma}_0) \\ &\propto \exp\left(-\frac{\frac{(\boldsymbol{a}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{a}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y})}{\sigma^2} + (\boldsymbol{a}-\boldsymbol{a}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{a}-\boldsymbol{a}_0)}{2}\right) \\ &\propto \exp\left(-\frac{(\boldsymbol{a} - \widehat{\boldsymbol{a}})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}(\boldsymbol{a} - \widehat{\boldsymbol{a}})}{2}\right), \end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

where

$$\widehat{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \left(\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{a}_0\right),$$
$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}.$$

Thus we have

$$p(\boldsymbol{a}|\mathcal{D}, \boldsymbol{a}_0, \boldsymbol{\Sigma}_0) = \text{Gauss}_M\left(\boldsymbol{a}; \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right). \tag{1.42}$$

Gamma Likelihood When only the noise variance parameter σ^2 is unknown, the model likelihood (1.39) is in the Gamma form, as a function of the inverse σ^{-2}:

$$\begin{aligned} p(\mathcal{D}|\sigma^{-2}) &\propto (\sigma^{-2})^{NM/2} \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2}{2}\sigma^{-2}\right) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{NM}{2} + 1, \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2}{2}\right), \end{aligned} \tag{1.43}$$

which implies that the ML estimator is

$$\widehat{\sigma^2}^{\text{ML}} = \frac{1}{\widehat{\sigma^{-2}}^{\text{ML}}} = \frac{1}{MN}\sum_{n=1}^{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2.$$

With the conjugate Gamma prior

$$p(\sigma^{-2}|\alpha_0, \beta_0) = \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \propto (\sigma^{-2})^{\alpha_0 - 1} \exp(-\beta_0 \sigma^{-2})$$

with hyperparameters $\boldsymbol{\kappa} = (\alpha_0, \beta_0)$, the posterior is computed as

$$\begin{aligned} p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) &\propto p(\mathcal{D}|\sigma^{-2}) p(\sigma^{-2}|\alpha_0, \beta_0) \\ &\propto \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + 1, \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2\right) \text{Gamma}(\sigma^{-2}; \alpha_0, \beta_0) \\ &\propto (\sigma^{-2})^{MN/2 + \alpha_0 - 1} \exp\left(-\left(\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2 + \beta_0\right)\sigma^{-2}\right). \end{aligned}$$

Therefore,

$$p(\sigma^{-2}|\mathcal{D}, \alpha_0, \beta_0) = \text{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + \alpha_0, \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|^2 + \beta_0\right). \tag{1.44}$$

Gauss-Gamma Likelihood When we estimate both parameters $\boldsymbol{w} = (\boldsymbol{a}, \sigma^{-2})$, the likelihood (1.39) is written as

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{a},\sigma^{-2}) &\propto (\sigma^{-2})^{NM/2}\exp\left(-\tfrac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}\|^2}{2}\sigma^{-2}\right)\\
&\propto (\sigma^{-2})^{NM/2}\exp\left(-\tfrac{\left(\boldsymbol{a}-\widehat{\boldsymbol{a}}^{\mathrm{ML}}\right)^\top\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{a}-\widehat{\boldsymbol{a}}^{\mathrm{ML}}\right)+\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}\|^2}{2}\sigma^{-2}\right)\\
&\propto \mathrm{GaussGamma}_M\left(\boldsymbol{a},\sigma^{-2};\widehat{\boldsymbol{a}}^{\mathrm{ML}},\boldsymbol{X}^\top\boldsymbol{X},\tfrac{M(N-1)}{2}+1,\tfrac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}\|^2}{2}\right),
\end{aligned}
$$

where $\widehat{\boldsymbol{a}}^{\mathrm{ML}}$ is the ML estimator, given by Eq. (1.41), for the regression parameter, and

$$
\begin{aligned}
&\mathrm{GaussGamma}_M(\boldsymbol{x},\tau|\boldsymbol{\mu},\boldsymbol{\Lambda},\alpha,\beta)\\
&\quad\equiv \mathrm{Gauss}_M(\boldsymbol{x}|\boldsymbol{\mu},(\tau\boldsymbol{\Lambda})^{-1})\cdot\mathrm{Gamma}(\tau|\alpha,\beta)\\
&\quad= \tfrac{\exp\left(-\frac{\tau}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{(2\pi\tau^{-1})^{M/2}\det(\boldsymbol{\Lambda})^{-1/2}}\cdot\tfrac{\beta^\alpha}{\Gamma(\alpha)}\,\tau^{\alpha-1}\exp(-\beta\tau)\\
&\quad= \tfrac{\beta^\alpha}{(2\pi)^{M/2}\det(\boldsymbol{\Lambda})^{-1/2}\Gamma(\alpha)}\,\tau^{\alpha+\frac{M}{2}-1}\exp\left(-\left(\tfrac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})}{2}+\beta\right)\tau\right)
\end{aligned}
$$

is the (general) Gauss-Gamma distribution on the random variable $\boldsymbol{x}\in\mathbb{R}^M$, $\tau>0$ with parameters $\boldsymbol{\mu}\in\mathbb{R}^M,\boldsymbol{\Lambda}\in\mathbb{S}^M_{++},\alpha>0,\beta>0$. With the conjugate Gauss-Gamma prior

$$
\begin{aligned}
p(\boldsymbol{a},\sigma^{-2}|\boldsymbol{\kappa}) &= \mathrm{GaussGamma}_M(\boldsymbol{a},\sigma^{-2}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\alpha_0,\beta_0)\\
&\propto (\sigma^{-2})^{\alpha_0+\frac{M}{2}-1}\exp\left(-\left(\tfrac{(\boldsymbol{a}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Lambda}_0(\boldsymbol{a}-\boldsymbol{\mu}_0)}{2}+\beta_0\right)\sigma^{-2}\right)
\end{aligned}
$$

for hyperparameters $\boldsymbol{\kappa}=(\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\alpha_0,\beta_0)$, the posterior is computed as

$$
\begin{aligned}
p(\boldsymbol{a},\sigma^{-2}|\mathcal{D},\boldsymbol{\kappa}) &\propto p(\mathcal{D}|\boldsymbol{a},\sigma^{-2})p(\boldsymbol{a},\sigma^{-2}|\boldsymbol{\kappa})\\
&\propto \mathrm{GaussGamma}_M\left(\boldsymbol{a},\sigma^{-2};\widehat{\boldsymbol{a}}^{\mathrm{ML}},\boldsymbol{X}^\top\boldsymbol{X},\tfrac{M(N-1)}{2}+1,\tfrac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}\|^2}{2}\right)\\
&\qquad\cdot\mathrm{GaussGamma}_M(\boldsymbol{a},\sigma^{-2}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\alpha_0,\beta_0)\\
&\propto (\sigma^{-2})^{NM/2}\exp\left(-\tfrac{\left(\boldsymbol{a}-\widehat{\boldsymbol{a}}^{\mathrm{ML}}\right)^\top\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{a}-\widehat{\boldsymbol{a}}^{\mathrm{ML}}\right)+\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}\|^2}{2}\sigma^{-2}\right)\\
&\qquad\cdot(\sigma^{-2})^{\alpha_0+\frac{M}{2}-1}\exp\left(-\left(\tfrac{(\boldsymbol{a}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Lambda}_0(\boldsymbol{a}-\boldsymbol{\mu}_0)}{2}+\beta_0\right)\sigma^{-2}\right)\\
&\propto (\sigma^{-2})^{\widehat{\alpha}+\frac{M}{2}-1}\exp\left(-\left(\frac{(\boldsymbol{a}-\widehat{\boldsymbol{\mu}})^\top\widehat{\boldsymbol{\Lambda}}(\boldsymbol{a}-\widehat{\boldsymbol{\mu}})}{2}+\widehat{\beta}\right)\sigma^{-2}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\boldsymbol{\mu}} &= (\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{\Lambda}_0)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}+\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right),\\
\widehat{\boldsymbol{\Lambda}} &= \boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{\Lambda}_0,\\
\widehat{\alpha} &= \tfrac{NM}{2}+\alpha_0,\\
\widehat{\beta} &= \tfrac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{a}}^{\mathrm{ML}}\|^2}{2}+\tfrac{(\widehat{\boldsymbol{a}}^{\mathrm{ML}}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Lambda}_0(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{\Lambda}_0)^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{a}}^{\mathrm{ML}}-\boldsymbol{\mu}_0)}{2}+\beta_0.
\end{aligned}
$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Thus, we obtain

$$p(\boldsymbol{a}, \sigma^{-2}|\mathcal{D}, \boldsymbol{\kappa}) = \text{GaussGamma}_M(\boldsymbol{a}, \sigma^{-2}|\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}, \widehat{\alpha}, \widehat{\beta}). \tag{1.45}$$

Multinomial Model

The *multinomial distribution*, which expresses a distribution over the *histograms* of independent events, is another frequently used basic component in Bayesian modeling. For example, it appears in *mixture models* and *latent Dirichlet allocation*.

Assume that exclusive K events occur with the probability

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Delta^{K-1} \equiv \left\{ \boldsymbol{\theta} \in \mathbb{R}^K; 0 \leq \theta_k \leq 1, \sum_{k=1}^{K} \theta_k = 1 \right\}.$$

Then, the histogram

$$\boldsymbol{x} = (x_1, \ldots, x_K) \in \mathbb{H}_N^{K-1} \equiv \left\{ \boldsymbol{x} \in \mathbb{I}^K; 0 \leq x_k \leq N; \sum_{k=1}^{K} x_k = N \right\}$$

of events after N iterations follows the *multinomial distribution*, defined as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \text{Multinomial}_{K,N}(\boldsymbol{x}; \boldsymbol{\theta}) \equiv N! \cdot \prod_{k=1}^{K} \frac{\theta_k^{x_k}}{x_k!}. \tag{1.46}$$

$\boldsymbol{\theta}$ is called the *multinomial parameter*.

As seen shortly, calculation of the posterior with its conjugate prior is surprisingly easy.

Dirichlet Likelihood As a function of the multinomial parameter $\boldsymbol{w} = \boldsymbol{\theta}$, it is easy to find that the likelihood (1.46) is in the form of the *Dirichlet distribution*:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x} + \mathbf{1}_K),$$

where $\mathbf{1}_K$ is the K-dimensional vector with all elements equal to 1. Since the Dirichlet distribution is an exponential family and hence multiplicatively closed, it is conjugate for the multinomial parameter. With the conjugate Dirichlet prior

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\phi}) \propto \prod_{k=1}^{K} \theta_k^{\phi-1}$$

with hyperparameters $\boldsymbol{\kappa} = \boldsymbol{\phi}$, the posterior is computed as

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\phi}) &\propto p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) \\
&\propto \mathrm{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x} + \mathbf{1}_K) \cdot \mathrm{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{\phi}) \\
&\propto \prod_{k=1}^{K} \theta_k^{x_k} \cdot \theta_k^{\phi_k - 1} \\
&\propto \prod_{k=1}^{K} \theta_k^{x_k + \phi_k - 1}.
\end{aligned}
$$

Thus we have

$$
p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\phi}) = \mathrm{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x} + \boldsymbol{\phi}). \tag{1.47}
$$

Special Cases For $K = 2$, the multinomial distribution is reduced to the *binomial distribution*:

$$
\begin{aligned}
p(x_1|\theta_1) &= \mathrm{Multinomial}_{2,N}\left((x_1, N - x_1)^\top; (\theta_1, 1 - \theta_1)^\top\right) \\
&= \mathrm{Binomial}_N(x_1; \theta_1) \\
&= \binom{N}{x_1} \cdot \theta_1^{x_1} (1 - \theta_1)^{N - x_1}.
\end{aligned}
$$

Furthermore, it is reduced to the *Bernoulli distribution* for $K = 2$ and $N = 1$:

$$
\begin{aligned}
p(x_1|\theta_1) &= \mathrm{Binomial}_1(x_1; \theta_1) \\
&= \theta_1^{x_1} (1 - \theta_1)^{1 - x_1}.
\end{aligned}
$$

Similarly, its conjugate Dirichlet distribution for $K = 2$ is reduced to the *Beta distribution*:

$$
\begin{aligned}
p(\theta_1|\phi_1, \phi_2) &= \mathrm{Dirichlet}_2\left((\theta_1, 1 - \theta_1)^\top; (\phi_1, \phi_2)^\top\right) \\
&= \mathrm{Beta}(\theta_1; \phi_1, \phi_2) \\
&= \frac{1}{\mathcal{B}(\phi_1, \phi_2)} \cdot \theta_1^{\phi_1 - 1} (1 - \theta_1)^{\phi_2 - 1},
\end{aligned}
$$

where $\mathcal{B}(\phi_1, \phi_2) = \frac{\Gamma(\phi_1)\Gamma(\phi_2)}{\Gamma(\phi_1 + \phi_2)}$ is the *Beta function*. Naturally, the Beta distribution is conjugate to the binomial and the Bernoulli distributions, and the posterior can be computed as easily as for the multinomial case.

With a conjugate prior in the form of a popular distribution, the four quantities introduced in Section 1.1.3, i.e., the marginal likelihood, the posterior mean, the posterior covariance, and the predictive distribution, can be obtained analytically. In the following subsections, we show how they are obtained.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Table 1.2 *First and second moments of common distributions.*
$\mathbf{Mean}(\boldsymbol{x}) = \langle \boldsymbol{x} \rangle_{p(\boldsymbol{x}|\boldsymbol{w})}$, $\mathrm{Var}(x) = \left\langle (x - \mathrm{Mean}(x))^2 \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w})}$,
$\mathbf{Cov}(\boldsymbol{x}) = \left\langle (\boldsymbol{x} - \mathbf{Mean}(\boldsymbol{x}))(\boldsymbol{x} - \mathbf{Mean}(\boldsymbol{x}))^\top \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w})}$, $\Psi(z) \equiv \frac{d}{dz} \log \Gamma(z)$:
Digamma function, and $\Psi_m(z) \equiv \frac{d^m}{dz^m} \Psi(z)$*: Polygamma function of order m.*

$p(\boldsymbol{x}\|\boldsymbol{w})$	First moment	Second moment
$\mathrm{Gauss}_M(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\mathbf{Mean}(\boldsymbol{x}) = \boldsymbol{\mu}$	$\mathbf{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$
$\mathrm{Gamma}(x; \alpha, \beta)$	$\mathrm{Mean}(x) = \frac{\alpha}{\beta}$ $\mathrm{Mean}(\log x)$ $= \Psi(\alpha) - \log \beta$	$\mathrm{Var}(x) = \frac{\alpha}{\beta^2}$ $\mathrm{Var}(\log x) = \Psi_1(\alpha)$
$\mathrm{Wishart}_M(\boldsymbol{X}; \boldsymbol{V}, \nu)$	$\mathbf{Mean}(\boldsymbol{X}) = \nu \boldsymbol{V}$	$\mathrm{Var}(x_{m,m'}) = \nu(V_{m,m'}^2 + V_{m,m} V_{m',m'})$
$\mathrm{Multinomial}_{K,N}(\boldsymbol{x}; \boldsymbol{\theta})$	$\mathbf{Mean}(\boldsymbol{x}) = N\boldsymbol{\theta}$	$(\mathbf{Cov}(\boldsymbol{x}))_{k,k'} = \begin{cases} N\theta_k(1-\theta_k) & (k = k') \\ -N\theta_k\theta_{k'} & (k \neq k') \end{cases}$
$\mathrm{Dirichlet}_K(\boldsymbol{x}; \boldsymbol{\phi})$	$\mathbf{Mean}(\boldsymbol{x}) = \frac{1}{\sum_{k=1}^K \phi_k} \boldsymbol{\phi}$ $\mathbf{Mean}(\log x_k)$ $= \Psi(\phi_k) - \Psi(\sum_{k'=1}^K \phi_{k'})$	$(\mathbf{Cov}(\boldsymbol{x}))_{k,k'} = \begin{cases} \frac{\phi_k(\tau - \phi_k)}{\tau^2(\tau+1)} & (k = k') \\ -\frac{\phi_k \phi_{k'}}{\tau^2(\tau+1)} & (k \neq k') \end{cases}$ where $\tau = \sum_{k=1}^K \phi_k$

1.2.4 Posterior Mean and Covariance

As seen in Section 1.2.3, by adopting a conjugate prior having a form of one of the common family distributions, such as the one in Table 1.1, we can have the posterior distribution in the same common family.[3] In such cases, we can simply use the known form of moments, which are summarized in Table 1.2. For example, the posterior (1.42) for the regression parameter $\boldsymbol{a}$ (when the noise variance σ^2 is treated as a known constant) is the following Gaussian distribution:

$$p(\boldsymbol{a}|\mathcal{D}, \boldsymbol{a}_0, \boldsymbol{\Sigma}_0) = \mathrm{Gauss}_M\left(\boldsymbol{a}; \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right),$$
$$\text{where} \quad \widehat{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \left(\frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{a}_0\right),$$
$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}.$$

[3] If we would say that the prior is in the family that contains all possible distributions, this family would be the conjugate prior for any likelihood function, which is however useless. Usually, the notion of the conjugate prior implicitly requires that moments (at least the normalization constant and the first moment) of any family member can be computed analytically.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Therefore, the posterior mean and the posterior covariance are simply given by

$$\langle \boldsymbol{a} \rangle_{p(\boldsymbol{a}|\mathcal{D},\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)} = \widehat{\boldsymbol{a}},$$
$$\left\langle (\boldsymbol{a} - \langle \boldsymbol{a} \rangle)(\boldsymbol{a} - \langle \boldsymbol{a} \rangle)^\top \right\rangle_{p(\boldsymbol{a}|\mathcal{D},\boldsymbol{a}_0,\boldsymbol{\Sigma}_0)} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}},$$

respectively. The posterior (1.29) of the (inverse) variance parameter σ^{-2} of the isotropic Gaussian distribution (when the mean parameter $\boldsymbol{\mu}$ is treated as a known constant) is the following Gamma distribution:

$$p(\sigma^{-2}|\mathcal{D},\alpha_0,\beta_0) = \mathrm{Gamma}\left(\sigma^{-2}; \frac{MN}{2} + \alpha_0, \frac{1}{2}\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2 + \beta_0\right).$$

Therefore, the posterior mean and the posterior variance are given by

$$\left\langle \sigma^{-2} \right\rangle_{p(\sigma^{-2}|\mathcal{D},\alpha_0,\beta_0)} = \frac{\frac{MN}{2} + \alpha_0}{\frac{1}{2}\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2 + \beta_0},$$
$$\left\langle \left(\sigma^{-2} - \left\langle \sigma^{-2} \right\rangle\right)^2 \right\rangle_{p(\sigma^{-2}|\mathcal{D},\alpha_0,\beta_0)} = \frac{\frac{MN}{2} + \alpha_0}{(\frac{1}{2}\sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}\|^2 + \beta_0)^2},$$

respectively.

Also in other cases, the posterior mean and the posterior covariances can be easily computed by using Table 1.2, if the form of the posterior distribution is in the table.

1.2.5 Predictive Distribution

The predictive distribution (1.9) for a new data set $\mathcal{D}^{\mathrm{new}}$ can be computed analytically, if the posterior distribution is in the exponential family, and hence multiplicatively closed. In this section, we show two examplary cases, the linear regression model and the multinomial model.

Linear Regression Model

Consider the linear regression model:

$$p(y|\boldsymbol{x},\boldsymbol{a}) = \mathrm{Gauss}_1(y; \boldsymbol{a}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y - \boldsymbol{a}^\top \boldsymbol{x})^2}{2\sigma^2}\right), \tag{1.48}$$

where only the regression parameter is unknown, i.e., $\boldsymbol{w} = \boldsymbol{a} \in \mathbb{R}^M$, and the noise variance parameter σ^2 is treated as a known constant. We choose the zero-mean Gaussian as a conjugate prior:

$$p(\boldsymbol{a}|\boldsymbol{C}) = \mathrm{Gauss}_M(\boldsymbol{a}; \boldsymbol{0}, \boldsymbol{C}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{a}^\top \boldsymbol{C}^{-1} \boldsymbol{a}\right)}{(2\pi)^{M/2} \det(\boldsymbol{C})^{1/2}}, \tag{1.49}$$

where $\boldsymbol{C}$ is the prior covariance.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

When N i.i.d. samples $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, where

$$\boldsymbol{y} = (y^{(1)}, \ldots, y^{(N)})^\top \in \mathbb{R}^N, \qquad \boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})^\top \in \mathbb{R}^{N \times M},$$

are observed, the posterior is given by

$$\begin{aligned} p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C}) &= \text{Gauss}_M\left(\boldsymbol{a}; \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right) \\ &= \frac{1}{(2\pi)^{M/2} \det\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right)^{1/2}} \cdot \exp\left(-\frac{(\boldsymbol{a} - \widehat{\boldsymbol{a}})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} (\boldsymbol{a} - \widehat{\boldsymbol{a}})}{2}\right), \end{aligned} \tag{1.50}$$

where

$$\widehat{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{C}^{-1}\right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}} \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2}, \tag{1.51}$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}} = \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{C}^{-1}\right)^{-1}. \tag{1.52}$$

This is just a special case of the posterior (1.42) for the linear regression model with the most general Gaussian prior.

Now, let us compute the predictive distribution on the output y^* for a new given input $\boldsymbol{x}^*$. As defined in Eq. (1.9), the predictive distribution is the expectation value of the model distribution (1.48) (for a new input–output pair) over the posterior distribution (1.50):

$$\begin{aligned} p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C}) &= \langle p(y^*|\boldsymbol{x}^*, \boldsymbol{a}) \rangle_{p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C})} \\ &= \int p(y^*|\boldsymbol{x}^*, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C}) d\boldsymbol{a} \\ &= \int \text{Gauss}_1(y^*; \boldsymbol{a}^\top \boldsymbol{x}^*, \sigma^2) \text{Gauss}_M\left(\boldsymbol{a}; \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right) d\boldsymbol{a} \\ &\propto \int \exp\left(-\frac{(y^* - \boldsymbol{a}^\top \boldsymbol{x}^*)^2}{2\sigma^2} - \frac{(\boldsymbol{a} - \widehat{\boldsymbol{a}})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} (\boldsymbol{a} - \widehat{\boldsymbol{a}})}{2}\right) d\boldsymbol{a} \\ &\propto \exp\left(-\frac{y^{*2}}{2\sigma^2}\right) \int \exp\left(-\frac{\boldsymbol{a}^\top \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^* \boldsymbol{x}^{*\top}}{\sigma^2}\right) \boldsymbol{a} - 2\boldsymbol{a}^\top \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} \widehat{\boldsymbol{a}} + \frac{\boldsymbol{x}^* y^*}{\sigma^2}\right)}{2}\right) d\boldsymbol{a} \\ &\propto \exp\left(-\frac{\sigma^{-2} y^{*2} - \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} \widehat{\boldsymbol{a}} + \frac{\boldsymbol{x}^* y^*}{\sigma^2}\right)^\top \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^* \boldsymbol{x}^{*\top}}{\sigma^2}\right)^{-1} \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} \widehat{\boldsymbol{a}} + \frac{\boldsymbol{x}^* y^*}{\sigma^2}\right)}{2}\right) \\ &\qquad \cdot \int \exp\left(-\frac{(\boldsymbol{a} - \breve{\boldsymbol{a}})^\top \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^* \boldsymbol{x}^{*\top}}{\sigma^2}\right) (\boldsymbol{a} - \breve{\boldsymbol{a}})}{2}\right) d\boldsymbol{a}, \end{aligned} \tag{1.53}$$

where

$$\breve{\boldsymbol{a}} = \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^* \boldsymbol{x}^{*\top}}{\sigma^2}\right)^{-1} \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} \widehat{\boldsymbol{a}} + \frac{\boldsymbol{x}^* y^*}{\sigma^2}\right).$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

Note that, although the preceding computation is similar to the one for the posterior distribution in Section 1.2.3, any factor that depends on y^* cannot be ignored even if it does not depend on $\boldsymbol{a}$, since the goal is to obtain the distribution on y^*.

The integrand in Eq. (1.53) coincides with the main part of

$$\mathrm{Gauss}_M\left(\boldsymbol{a}; \breve{\boldsymbol{a}}, \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^*\boldsymbol{x}^{*\top}}{\sigma^2}\right)^{-1}\right)$$

without the normalization factor. Therefore, the integral is the inverse of the normalization factor, i.e.,

$$\int \exp\left(-\frac{(\boldsymbol{a}-\breve{\boldsymbol{a}})^\top\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}+\frac{\boldsymbol{x}^*\boldsymbol{x}^{*\top}}{\sigma^2}\right)(\boldsymbol{a}-\breve{\boldsymbol{a}})}{2}\right) d\boldsymbol{a} = (2\pi)^{M/2}\det\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1} + \frac{\boldsymbol{x}^*\boldsymbol{x}^{*\top}}{\sigma^2}\right)^{-1/2},$$

which is a constant with respect to y^*. Therefore, by using Eqs. (1.51) and (1.52), we have

$$\begin{aligned}
&p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C}) \\
&\propto \exp\left(-\frac{\sigma^{-2}y^{*2}-\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}\widehat{\boldsymbol{a}}+\frac{\boldsymbol{x}^*y^*}{\sigma^2}\right)^\top\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}+\frac{\boldsymbol{x}^*\boldsymbol{x}^{*\top}}{\sigma^2}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}\widehat{\boldsymbol{a}}+\frac{\boldsymbol{x}^*y^*}{\sigma^2}\right)}{2}\right) \\
&\propto \exp\left(-\frac{y^{*2}-\left(\boldsymbol{X}^\top\boldsymbol{y}+\boldsymbol{x}^*y^*\right)^\top\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{y}+\boldsymbol{x}^*y^*\right)}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left\{y^{*2}\left(1-\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{x}^*\right)\right.\right. \\
&\qquad\left.\left.-2y^*\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\right\}\right) \\
&\propto \exp\left(-\frac{1-\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{x}^*}{2\sigma^2}\right. \\
&\qquad\left.\cdot\left(y^*-\frac{\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}}{1-\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{x}^*}\right)^2\right) \\
&\propto \exp\left(-\frac{(y^*-\widehat{y})^2}{2\widehat{\sigma}_y^2}\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{y} &= \frac{\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}}{1-\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{x}^*}, \\
\widehat{\sigma}_y^2 &= \frac{\sigma^2}{1-\boldsymbol{x}^{*\top}\left(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{x}^*\boldsymbol{x}^{*\top}+\sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{x}^*}.
\end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

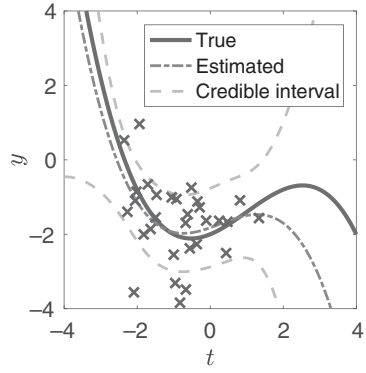

Figure 1.3 Predictive distribution of the linear regression model.

Thus, the predictive distribution has been analytically obtained:

$$p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{C}) = \mathrm{Gauss}_1\left(y^*; \widehat{y}, \widehat{\sigma}_y^2\right). \tag{1.54}$$

Figure 1.3 shows an example of the predictive distribution of the linear regression model. The curve labeled as "True" indicates the mean $y = \boldsymbol{a}^*\boldsymbol{x}$ of the true regression model $y = \boldsymbol{a}^*\boldsymbol{x} + \varepsilon$, where $\boldsymbol{a}^* = (-2, 0.4, 0.3, -0.1)^\top$, $\boldsymbol{x} = (1, t, t^2, t^3)^\top$, and $\varepsilon \sim \mathrm{Gauss}_1(0, 1^2)$. The crosses are $N = 30$ i.i.d. observed samples generated from the true regression model and the input distribution $t \sim \mathrm{Uniform}(-2.4, 1.6)$, where $\mathrm{Uniform}(l, u)$ denotes the uniform distribution on $[l, u]$. The regression model (1.48) with the prior (1.49) for the hyperparameters $\boldsymbol{C} = 10000 \cdot \boldsymbol{I}_M, \sigma^2 = 1$ was trained with the observed samples. The curve labeled as "Estimated" and the pair of curves labeled as "Credible interval" show the mean $\widehat{y}$ and the *credible interval* $\widehat{y} \pm \widehat{\sigma}_y$ of the predictive distribution (1.54), respectively.

Reflecting the fact that the samples are observed only in the middle region ($t \in [-2.4, 1.6]$), the credible interval is large in outer regions. The larger interval implies that the "Estimated" function is less reliable, and we see that the gap from the "True" function is indeed large. Since the true function is unknown in practical situations, the variance of the predictive distribution is important information on the reliability of the estimated result.

Multinomial Model

Let us compute the predictive distribution of the multinomial model:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \mathrm{Multinomial}_{K,N}(\boldsymbol{x}; \boldsymbol{\theta}) \propto \prod_{k=1}^{K} \frac{\theta_k^{x_k}}{x_k!},$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}) = \mathrm{Dirichlet}_K(\boldsymbol{\theta};\boldsymbol{\phi}) \propto \prod_{k=1}^{K} \theta_k^{\phi_k - 1},$$

with the observed data $\mathcal{D} = \boldsymbol{x} = (x_1, \ldots, x_K) \in \mathbb{H}_N^{K-1}$ and the unknown parameter $\boldsymbol{w} = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Delta^{K-1}$.

The posterior was derived in Eq. (1.47):

$$p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\phi}) = \mathrm{Dirichlet}_K(\boldsymbol{\theta};\boldsymbol{x}+\boldsymbol{\phi}) \propto \prod_{k=1}^{K} \theta_k^{x_k + \phi_k - 1}.$$

Therefore, the predictive distribution for a new single sample $\boldsymbol{x}^* \in \mathbb{H}_1^{K-1}$ is given by

$$\begin{aligned}
p(\boldsymbol{x}^*|\boldsymbol{x},\boldsymbol{\phi}) &= \langle p(\boldsymbol{x}^*|\boldsymbol{\theta}) \rangle_{p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\phi})} \\
&= \int p(\boldsymbol{x}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&= \int \mathrm{Multinomial}_{K,1}(\boldsymbol{x}^*;\boldsymbol{\theta}) \mathrm{Dirichlet}_K(\boldsymbol{\theta};\boldsymbol{x}+\boldsymbol{\phi}) d\boldsymbol{\theta} \\
&\propto \int \prod_{k=1}^{K} \theta_k^{x_k^*} \cdot \theta_k^{x_k + \phi_k - 1} d\boldsymbol{\theta} \\
&= \int \prod_{k=1}^{K} \theta_k^{x_k^* + x_k + \phi_k - 1} d\boldsymbol{\theta}. \qquad (1.55)
\end{aligned}$$

In the fourth equation, we ignored the factors that depend neither on $\boldsymbol{x}^*$ nor on $\boldsymbol{\theta}$.

The integrand in Eq. (1.55) is the main part of $\mathrm{Dirichlet}_K(\boldsymbol{\theta};\boldsymbol{x}^*+\boldsymbol{x}+\boldsymbol{\phi})$, and therefore, the integral is equal to the inverse of its normalization factor:

$$\begin{aligned}
\int \prod_{k=1}^{K} \theta_k^{x_k^* + x_k + \phi_k - 1} d\boldsymbol{\theta} &= \frac{\prod_{k=1}^{K} \Gamma(x_k^* + x_k + \phi_k)}{\Gamma(\sum_{k=1}^{K} x_k^* + x_k + \phi_k)} \\
&= \frac{\prod_{k=1}^{K} \Gamma(x_k^* + x_k + \phi_k)}{\Gamma(N + \sum_{k=1}^{K} \phi_k + 1)}.
\end{aligned}$$

Thus, by using the identity $\Gamma(x+1) = x\Gamma(x)$ for the Gamma function, we have

$$\begin{aligned}
p(\boldsymbol{x}^*|\boldsymbol{x},\boldsymbol{\phi}) &\propto \prod_{k=1}^{K} \Gamma(x_k^* + x_k + \phi_k) \\
&\propto \prod_{k=1}^{K} (x_k + \phi_k)^{x_k^*} \Gamma(x_k + \phi_k)
\end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$\begin{aligned}
&\propto \prod_{k=1}^{K} (x_k + \phi_k)^{x_k^*} \\
&\propto \prod_{k=1}^{K} \left(\frac{x_k + \phi_k}{\sum_{k'=1}^{K} x_{k'} + \phi_{k'}} \right)^{x_k^*} \\
&= \text{Multinomial}_{K,1}(\boldsymbol{x}^*; \widehat{\boldsymbol{\theta}}),
\end{aligned} \tag{1.56}$$

where

$$\widehat{\theta}_k = \frac{x_k + \phi_k}{\sum_{k'=1}^{K} x_{k'} + \phi_{k'}}. \tag{1.57}$$

From Eq. (1.47) and Table 1.2, we can easily see that the predictive mean $\widehat{\boldsymbol{\theta}}$, specified by Eq. (1.57), coincides with the posterior mean, i.e., the Bayesian estimator:

$$\widehat{\boldsymbol{\theta}} = \langle \boldsymbol{\theta} \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x}+\boldsymbol{\phi})}.$$

Therefore, in the multinomial model, the predictive distribution coincides with the model distribution with the Bayesian estimator plugged in.

In the preceding derivation, we performed the integral computation and derived the form of the predictive distribution. However, the necessary information to determine the predictive distribution is the probability table on the events $\boldsymbol{x}^* \in \mathbb{H}_1^{K-1} = \{\boldsymbol{e}_k\}_{k=1}^K$, of which the degree of freedom is only K. Therefore, the following simple calculation gives the same result:

$$\begin{aligned}
\text{Prob}(\boldsymbol{x}^* = \boldsymbol{e}_k | \boldsymbol{x}, \boldsymbol{\phi}) &= \langle \text{Multinomial}_{K,1}(\boldsymbol{e}_k; \boldsymbol{\theta}) \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x}+\boldsymbol{\phi})} \\
&= \langle \theta_k \rangle_{\text{Dirichlet}_K(\boldsymbol{\theta}; \boldsymbol{x}+\boldsymbol{\phi})} \\
&= \widehat{\theta}_k,
\end{aligned}$$

which specifies the function form of the predictive distribution, given by Eq. (1.56).

1.2.6 Marginal Likelihood

Let us compute the marginal likelihood of the linear regression model, defined by Eqs. (1.48) and (1.49):

$$\begin{aligned}
p(\mathcal{D}|\boldsymbol{C}) &= p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{C}) \\
&= \langle p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a}) \rangle_{p(\boldsymbol{a}|\boldsymbol{C})} \\
&= \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{C}) d\boldsymbol{a}
\end{aligned}$$

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$
\begin{aligned}
&= \int \text{Gauss}_N(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{a}, \sigma^2 \boldsymbol{I}_N)\text{Gauss}_M(\boldsymbol{a}; \boldsymbol{0}, \boldsymbol{C})d\boldsymbol{a} \\
&= \int \frac{\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{a}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}} \cdot \frac{\exp\left(-\frac{1}{2}\boldsymbol{a}^\top \boldsymbol{C}^{-1}\boldsymbol{a}\right)}{(2\pi)^{M/2}\det(\boldsymbol{C})^{1/2}} d\boldsymbol{a} \\
&= \frac{\exp\left(-\frac{\|\boldsymbol{y}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2}\det(\boldsymbol{C})^{1/2}} \\
&\qquad \cdot \int \exp\left(-\frac{-2\boldsymbol{a}^\top \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2} + \boldsymbol{a}^\top\left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \boldsymbol{C}^{-1}\right)\boldsymbol{a}}{2}\right) d\boldsymbol{a} \\
&= \frac{\exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{y}\|^2}{\sigma^2} - \widehat{\boldsymbol{a}}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}\widehat{\boldsymbol{a}}\right)\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2}\det(\boldsymbol{C})^{1/2}} \\
&\qquad \cdot \int \exp\left(-\frac{(\boldsymbol{a}-\widehat{\boldsymbol{a}})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}(\boldsymbol{a}-\widehat{\boldsymbol{a}})}{2}\right) d\boldsymbol{a}, \qquad (1.58)
\end{aligned}
$$

where $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}$ are, respectively, the posterior mean and the posterior covariance, given by Eqs. (1.51) and (1.52).

By using

$$
\int \exp\left(-\frac{(\boldsymbol{a}-\widehat{\boldsymbol{a}})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}^{-1}(\boldsymbol{a}-\widehat{\boldsymbol{a}})}{2}\right) d\boldsymbol{a} = \sqrt{(2\pi)^M \det\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right)},
$$

and Eq. (1.58), we have

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{C}) &= \frac{\exp\left(-\frac{1}{2}\left(\frac{\|\boldsymbol{y}\|^2}{\sigma^2} - \frac{\boldsymbol{y}^\top \boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^4}\right)\right)}{(2\pi\sigma^2)^{N/2}(2\pi)^{M/2}\det(\boldsymbol{C})^{1/2}} \sqrt{(2\pi)^M \det\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{a}}\right)} \\
&= \frac{\exp\left(-\frac{\|\boldsymbol{y}\|^2 - \boldsymbol{y}^\top \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{y}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{N/2}\det(\boldsymbol{C}\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{I}_M)^{1/2}}, \qquad (1.59)
\end{aligned}
$$

where we also used Eqs. (1.51) and (1.52).

Eq. (1.59) is an explicit expression of the marginal likelihood as a function of the hyperparameter $\boldsymbol{\kappa} = \boldsymbol{C}$. Based on it, we perform EBayes learning in Section 1.2.7.

1.2.7 Empirical Bayesian Learning

In empirical Bayesian (EBayes) learning, the hyperparameter $\boldsymbol{\kappa}$ is estimated by maximizing the marginal likelihood $p(\mathcal{D}|\boldsymbol{\kappa})$. The negative logarithm of the marginal likelihood,

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

$$F^{\text{Bayes}} = -\log p(\mathcal{D}|\boldsymbol{\kappa}), \tag{1.60}$$

is called the *Bayes free energy* or *stochastic complexity*.[4] Since $\log(\cdot)$ is a monotonic function, maximizing the marginal likelihood is equivalent to minimizing the Bayes free energy.

Eq. (1.59) implies that the Bayes free energy of the linear regression model is given by

$$\begin{aligned} 2F^{\text{Bayes}} &= -2\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{C}) \\ &= N\log(2\pi\sigma^2) + \log\det(\boldsymbol{C}\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{I}_M) \\ &\quad + \frac{||\boldsymbol{y}||^2 - \boldsymbol{y}^\top\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}}{\sigma^2}. \end{aligned} \tag{1.61}$$

Let us restrict the prior covariance to be diagonal:

$$\boldsymbol{C} = \mathbf{Diag}(c_1^2, \ldots, c_M^2) \in \mathbb{D}^M. \tag{1.62}$$

The prior (1.49) with diagonal covariance (1.62) is called the *automatic relevance determination (ARD)* prior, which is known to make the EBayes estimator sparse (Neal, 1996). In the following example, we see this effect by setting the design matrix to identity, $\boldsymbol{X} = \boldsymbol{I}_M$, which enables us to derive the EBayes solution analytically.

Under the identity design matrix, the Bayes free energy (1.61) can be decomposed as

$$\begin{aligned} 2F^{\text{Bayes}} &= N\log(2\pi\sigma^2) + \log\det(\boldsymbol{C} + \sigma^2\boldsymbol{I}_M) + \frac{||\boldsymbol{y}||^2 - \boldsymbol{y}^\top\left(\boldsymbol{I}_M + \sigma^2\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{y}}{\sigma^2} \\ &= N\log(2\pi\sigma^2) + \frac{||\boldsymbol{y}||^2}{\sigma^2} + \sum_{m=1}^{M}\left(\log(c_m^2 + \sigma^2) - \frac{y_m^2}{\sigma^2\left(1 + \sigma^2 c_m^{-2}\right)}\right) \\ &= \sum_{m=1}^{M} 2F_m^* + \text{const.}, \end{aligned} \tag{1.63}$$

where

$$2F_m^* = \log\left(1 + \frac{c_m^2}{\sigma^2}\right) - \frac{y_m^2}{\sigma^2}\left(1 + \frac{\sigma^2}{c_m^2}\right)^{-1}. \tag{1.64}$$

In Eq. (1.63), we omitted the constant factors with respect to the hyperparameter $\boldsymbol{C}$. As the remaining terms are decomposed into each component m, we can independently minimize F_m^* with respect to c_m^2.

[4] The logarithm of the marginal likelihood $\log p(\mathcal{D}|\boldsymbol{\kappa})$ is called the *log marginal likelihood* or *evidence*.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

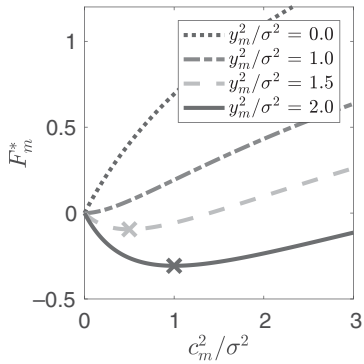

Figure 1.4 The (componentwise) Bayes free energy (1.64) of linear regression model with the ARD prior. The minimizer is shown as a cross if it lies in the positive region of c_m^2/σ^2.

The derivative of Eq. (1.64) with respect to c_m^2 is

$$\begin{aligned} 2\frac{\partial F_m^*}{\partial c_m^2} &= \frac{1}{c_m^2+\sigma^2} - \frac{y_m^2}{\left(1+\sigma^2 c_m^{-2}\right)^2 c_m^4} \\ &= \frac{1}{c_m^2+\sigma^2} - \frac{y_m^2}{\left(c_m^2+\sigma^2\right)^2} \\ &= \frac{c_m^2 - (y_m^2-\sigma^2)}{(c_m^2+\sigma^2)^2}. \end{aligned} \tag{1.65}$$

Eq. (1.65) implies that F_m^* is monotonically increasing over all domain $c_m^2 > 0$ when $y_m^2 \le \sigma^2$, and has the unique minimizer in the region $c_m^2 > 0$ when $y_m^2 > \sigma^2$. Specifically, the minimizer is given by

$$\widehat{c_m^2} = \begin{cases} y_m^2 - \sigma^2 & \text{if } y_m^2 > \sigma^2, \\ +0 & \text{otherwise.} \end{cases} \tag{1.66}$$

Figure 1.4 shows the (componentwise) Bayes free energy (1.64) for different observations, $y_m^2 = 0, \sigma^2, 1.5\sigma^2, 2\sigma^2$. The minimizer is in the positive region of c_m^2 if and only if $y_m^2 > \sigma^2$.

If the EBayes estimator is given by $\widehat{c_m^2} \to +0$, it means that the *prior* distribution for the mth component a_m of the regression parameter is the *Dirac delta function* located at the origin.[5] This formally means that we *a priori*

[5] When $y_m^2 \le \sigma^2$, the Bayes free energy (1.64) decreases as c_m^2 approaches to 0. However, the domain of c_m^2 is restricted to be positive, and therefore, $\widehat{c_m^2} = 0$ is not the solution. We express this solution as $\widehat{c_m^2} \to +0$.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press

knew that $a_m = 0$, i.e., we choose a model that does not contain the mth component.

By substituting Eq. (1.66) into the Bayes posterior mean (1.51), we obtain the EBayes estimator:

$$\begin{aligned}\widehat{a}_m^{\text{EBayes}} &= \widehat{c}_m^2 \left(\widehat{c}_m^2 + \sigma^2\right)^{-1} y_m \\ &= \begin{cases} \left(1 - \frac{\sigma^2}{y_m^2}\right) y_m & \text{if } y_m^2 > \sigma^2, \\ 0 & \text{otherwise.} \end{cases}\end{aligned} \tag{1.67}$$

The form of the estimator (1.67) is called the *James–Stein (JS) estimator* having interesting properties including the *domination* over the ML estimator (Stein, 1956; James and Stein, 1961; Efron and Morris, 1973) (see Appendix A).

Note that the assumption that $\boldsymbol{X} = \boldsymbol{I}_M$ is not practical. For a general design matrix $\boldsymbol{X}$, the Bayes free energy is not decomposable into each component. Consequently, the prior variances $\{c_m^2\}_{m=1}^M$ that minimize the Bayes free energy (1.61) interact with each other. Therefore, the preceding simple mechanism is not applied. However, it is empirically observed that many prior variances tend to go to $\widehat{c}_m^2 \to +0$, so that the EBayes estimator $\widehat{\boldsymbol{a}}^{\text{EBayes}}$ is sparse.

https://doi.org/10.1017/9781139879354.003 Published online by Cambridge University Press