CHAPTER

MODEL SELECTION FOR MAXIMUM LIKELIHOOD ESTIMATION

14

CHAPTER CONTENTS

So far, a parametric model was assumed to be given and fixed. However, in practice, it is often necessary to choose a parametric model from some candidates. In this chapter, a data-driven method to choose an appropriate parametric model is explained.

14.1 MODEL SELECTION

As shown in Chapter 12, “Gaussian models” have several different variations.

(A) Variance-covariance matrix $\boldsymbol{\Sigma}$ is generic:

$$q(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

(B) Variance-covariance matrix $\boldsymbol{\Sigma}$ is diagonal:

$$q(\boldsymbol{x};\boldsymbol{\mu},(\sigma^{(1)})^2,\ldots,(\sigma^{(d)})^2) = \prod_{j=1}^{d}\frac{1}{\sqrt{2\pi(\sigma^{(j)})^2}}\exp\left(-\frac{(x^{(j)}-\mu^{(j)})^2}{2(\sigma^{(j)})^2}\right).$$

(C) Variance-covariance matrix $\boldsymbol{\Sigma}$ is proportional to the identity matrix:

$$q(\boldsymbol{x};\boldsymbol{\mu},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})}{2\sigma^2}\right).$$

To obtain a good approximation to the true probability density function by a parametric method, a parametric model that (approximately) contains the true

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00025-X

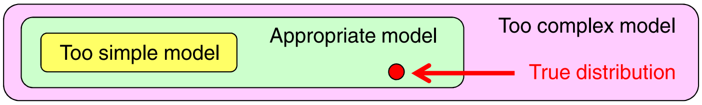

FIGURE 14.1

Model selection. Too simple model may not be expressive enough to represent the true probability distribution, while too complex model may cause unreliable parameter estimation.

probability density function may be required. From this viewpoint, a richer model that contains various probability density functions is preferable. On the other hand, as explained in Chapter 12, the good performance of MLE is guaranteed only when the number of training samples is large (relative to the number of parameters). This implies that a simple model that has a small number of parameters is desirable. Therefore, in practice, a model that fulfills these two conflicting requirements in a balanced way should be selected (Fig. 14.1).

However, since the appropriate balance depends on the unknown true probability distribution, let us consider the following data-driven procedure for model selection:

1. Prepare parametric models $\{q_m(\boldsymbol{x};\boldsymbol{\theta})\}_m$.
2. For each parametric model $q_m(\boldsymbol{x};\boldsymbol{\theta})$, compute maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_m$ and obtain a density estimator $\widehat{p}_m(\boldsymbol{x})$ as

$$\widehat{p}_m(\boldsymbol{x}) = q_m(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_m).$$

3. From the obtained density estimators $\{\widehat{p}_m(\boldsymbol{x})\}_m$, choose the one that is closest to the true probability density function $p(\boldsymbol{x})$.

The problem of finding the most promising model from a set of model candidates is called *model selection*. At a glance, finding the estimator that is closest to the true probability density function $p(\boldsymbol{x})$ in Step 3 is not possible because $p(\boldsymbol{x})$ is unknown. Below, two model selection approaches called the *Akaike information criterion* (AIC) and *cross validation* are introduced for coping with this problem.

14.2 KL DIVERGENCE

To perform model selection following the above procedure, a “closeness” measure between the true probability density function $p(\boldsymbol{x})$ and its estimator $\widehat{p}(\boldsymbol{x})$ is needed. Mathematically, for a set $\mathcal{X}$, function $g : \mathcal{X}\times\mathcal{X} \rightarrow \mathbb{R}$ that satisfies the following four conditions for any $x, y, z \in \mathcal{X}$ is called a *distance function*:

$$\begin{aligned}\text{Non-negativity: } & g(x,y) \geq 0,\\ \text{Symmetry: } & g(x,y) = g(y,x),\\ \text{Identity: } & g(x,y) = 0 \iff x = y,\end{aligned}$$

$$\text{Triangle inequality: } g(x,y) + g(y,z) \geq g(x,z).$$

In this section, a typical closeness measure called the *Kullback-Leibler divergence* (KL divergence) [64] is introduced.

The KL divergence from p to $\widehat{p}$ is defined as

$$\mathrm{KL}(p\|\widehat{p}) = E\left[\log \frac{p(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}\right], \tag{14.1}$$

where E is the expectation operator over $p(\boldsymbol{x})$:

$$E[\bullet] = \int \bullet \, p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{14.2}$$

The KL divergence does not actually satisfy symmetric and triangle inequality, meaning that it is not mathematically a distance. Nevertheless, the KL divergence still satisfies non-negativity and identity. Therefore, even though it is not a proper distance, a smaller KL divergence would imply that $\widehat{p}$ is a better approximator to p.

Since the KL divergence $\mathrm{KL}(p\|\widehat{p})$ contains unknown true probability density function $p(\boldsymbol{x})$, it cannot be directly computed. Let us expand Eq. (14.1) as

$$\mathrm{KL}(p\|\widehat{p}) = E\left[\log p(\boldsymbol{x})\right] - E\left[\log \widehat{p}(\boldsymbol{x})\right],$$

where $E\left[\log p(\boldsymbol{x})\right]$ is the negative *entropy* of p that is a constant independent of $\widehat{p}$. In the context of model selection, the KL divergence is used for comparing different models, and therefore the constant term is irrelevant. For this reason, the first term is ignored and only the second term $E\left[\log \widehat{p}(\boldsymbol{x})\right]$, which is the expected log-likelihood, is considered below.

Even if the first term is ignored, the second term still contains the expectation over unknown p and thus it cannot be directly computed. The most naive approach to estimating the expectation would be the sample average, i.e., the expected log-likelihood is approximated by the average log-likelihood over i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$ with $p(\boldsymbol{x})$:

$$E[\log \widehat{p}(\boldsymbol{x})] \approx \frac{1}{n}\sum_{i=1}^{n} \log \widehat{p}(\boldsymbol{x}_i).$$

Then, a larger log-likelihood approximately implies a smaller KL divergence, and therefore a model with a large log-likelihood would be close to the true probability density function p.

However, the naive use of the log-likelihood for model selection does not work well in practice. To illustrate this, let us consider model selection from the three Gaussian models introduced in the beginning of Section 14.1:

(A) Variance-covariance matrix $\boldsymbol{\Sigma}$ is generic.

(B) Variance-covariance matrix $\boldsymbol{\Sigma}$ is diagonal.

(C) Variance-covariance matrix $\boldsymbol{\Sigma}$ is proportional to the identity matrix.

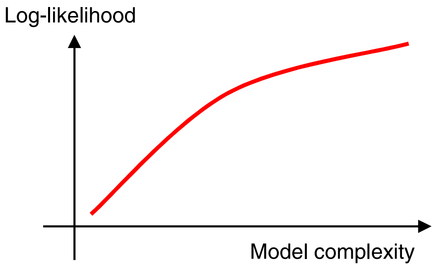

FIGURE 14.2

For nested models, log-likelihood is monotone nondecreasing as the model complexity increases.

These three models are nested, meaning that model (C) is a special case of model (B) and model (B) is a special case of model (A). This implies that the maximum likelihood solution in model (C) is also included in model (B), and therefore model (B) can achieve at least the same log-likelihood as model (C). Similarly, the maximum likelihood solutions in model (B) and model (C) are also included in model (A), and therefore model (A) can achieve at least the same log-likelihood as model (B) and model (C). For this reason, when model selection is performed for nested models based on the log-likelihood, the largest model (e.g., model (A) in the above Gaussian model examples) is always chosen (Fig. 14.2).

14.3 AIC

As explained above, the naive use of log-likelihood for model selection results in just always selecting the most complex model. This is caused by the fact that the average log-likelihood is not an accurate enough estimator of the expected log-likelihood. For appropriate model selection, therefore, a more accurate estimator of the expected log-likelihood is needed. The *Akaike information criterion*AIC gives a better estimator of the expected log-likelihood [1].

AIC is defined as

$$\mathrm{AIC} = -\sum_{i=1}^{n} \log q(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) + b, \tag{14.3}$$

where the first term is the negative log-likelihood and the second term is the number of parameters. Thus, AIC can be regarded as correcting the negative log-likelihood adding by the number of parameters.

Let us intuitively explain why AIC can yield better model selection. As illustrated in Fig. 14.2, the negative log-likelihood is monotone nondecreasing as the number of parameters increases. On the other hand, the number of parameters is monotone

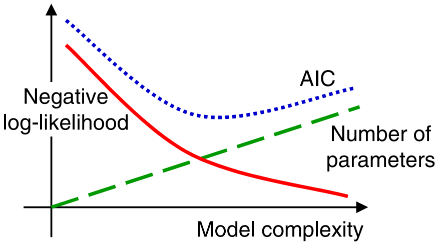

FIGURE 14.3

AIC is the sum of the negative log-likelihood and the number of parameters.

increasing as the number of parameters increases. Therefore, AIC is balancing these conflicting functions by summing them up. As a result, an appropriate model that achieves a reasonably large log-likelihood with a reasonably small number of parameters tends to have a small AIC value (Fig. 14.3).

In various areas of science and engineering, *Occam's razor*, also known as the *principle of parsimony*, is often used as a guiding principle. Occam's razor suggests choosing a simpler hypothesis among similar hypotheses, and AIC can be regarded as justifying the use of Occam's razor in statistical inference. More specifically, among the models that achieve similar log-likelihood values, AIC suggests choosing the one that has the smallest number of parameters.

Next, theoretical validity of AIC is investigated. Let $J(\boldsymbol{\theta})$ be the negative expected log-likelihood of $q(\boldsymbol{x};\boldsymbol{\theta})$,

$$J(\boldsymbol{\theta}) = -E[\log q(\boldsymbol{x};\boldsymbol{\theta})],$$

where E is the expectation operator over $p(\boldsymbol{x})$ (see Eq.(14.2)). Let $\boldsymbol{\theta}^*$ be the minimizer of $J(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, J(\boldsymbol{\theta}).$$

According to *asymptotic theory*, the expectation of $J(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ can be expanded using $\boldsymbol{\theta}^*$ as follows [63, 107]:

$$\mathbb{E}[J] = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log q(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})\right] + \frac{1}{n}\mathrm{tr}\left(\boldsymbol{Q}(\boldsymbol{\theta}^*)\boldsymbol{G}(\boldsymbol{\theta}^*)^{-1}\right) + o(n^{-1}), \qquad (14.4)$$

where $\mathbb{E}$ denotes the expectation over all training samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ following i.i.d. with $p(\boldsymbol{x})$:

$$\mathbb{E}[\bullet] = \int\cdots\int \bullet\, p(\boldsymbol{x}_1)\cdots p(\boldsymbol{x}_n)\mathrm{d}\boldsymbol{x}_1\cdots\mathrm{d}\boldsymbol{x}_n.$$

$f(n) = O(g(n))$ means that

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} < \infty,$$

while $f(n) = o(g(n))$ means that

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

Intuitively, $O(n^{-1})$ denotes a term that has the same size as n^{-1}, while $o(n^{-1})$ denotes a term that is smaller than n^{-1}.

FIGURE 14.4

Big-o and small-o notations.

$o(n^{-1})$ denotes a term that is smaller than n^{-1} asymptotically (see Fig. 14.4). Matrices $\boldsymbol{Q}(\boldsymbol{\theta})$ and $\boldsymbol{G}(\boldsymbol{\theta})$ are defined as

$$\boldsymbol{Q}(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial\boldsymbol{\theta}} \log q(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial\boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}}\right],$$

$$\boldsymbol{G}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}}\right].$$

$\boldsymbol{Q}(\boldsymbol{\theta})$ looks similar to the *Fisher information matrix* $\boldsymbol{F}(\boldsymbol{\theta})$ (see Eq. (13.4)):

$$\boldsymbol{F}(\boldsymbol{\theta}) = \int \left(\frac{\partial}{\partial\boldsymbol{\theta}} \log q(\boldsymbol{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right) q(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}.$$

However, $\boldsymbol{Q}(\boldsymbol{\theta})$ includes the expectation over true $p(\boldsymbol{x})$, while $\boldsymbol{F}(\boldsymbol{\theta})$ includes the expectation over model $q(\boldsymbol{x};\boldsymbol{\theta})$. $\boldsymbol{G}(\boldsymbol{\theta})$ is the negative Hessian matrix of $\log q(\boldsymbol{x};\boldsymbol{\theta})$ expected over true $p(\boldsymbol{x})$.

$\boldsymbol{G}(\boldsymbol{\theta})$ can be expanded as

$$\begin{aligned}
\boldsymbol{G}(\boldsymbol{\theta}) &= -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right] \\
&= -E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})}\right] + E\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})} \frac{\frac{\partial}{\partial\boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})}\right] \\
&= -E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})}\right] + E\left[\frac{\partial}{\partial\boldsymbol{\theta}} \log q(\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial}{\partial\boldsymbol{\theta}^\top} \log q(\boldsymbol{x};\boldsymbol{\theta})\right] \\
&= -\int \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})}{q(\boldsymbol{x};\boldsymbol{\theta})} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \boldsymbol{Q}(\boldsymbol{\theta}).
\end{aligned}$$

Now, suppose that the parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$ contains the true probability density function $p(\boldsymbol{x})$, i.e., there exists $\boldsymbol{\theta}^*$ such that $q(\boldsymbol{x};\boldsymbol{\theta}^*) = p(\boldsymbol{x})$. Then $\boldsymbol{G}(\boldsymbol{\theta}^*)$ can be expressed as

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\theta}^*) &= -\int \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} q(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} + \boldsymbol{Q}(\boldsymbol{\theta}^*) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\int q(\boldsymbol{x};\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} + \boldsymbol{Q}(\boldsymbol{\theta}^*)\\ &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}1 + \boldsymbol{Q}(\boldsymbol{\theta}^*) = \boldsymbol{Q}(\boldsymbol{\theta}^*).\end{aligned}$$

Thus, the second term in Eq. (14.4) can be expressed as

$$\mathrm{tr}\left(\boldsymbol{Q}(\boldsymbol{\theta}^*)\boldsymbol{G}(\boldsymbol{\theta}^*)^{-1}\right) = \mathrm{tr}\,(\boldsymbol{I}_b) = b,$$

which agrees with the second term in AIC, i.e., the number of parameters.

As explained above, AIC assumes that the parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$ contains the true probability density function $p(\boldsymbol{x})$. The *Takeuchi information criterion* (TIC) is a generalization of AIC that removes this assumption [63, 107]. TIC is defined as

$$\mathrm{TIC} = -\sum_{i=1}^{n}\log q(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) + \mathrm{tr}\left(\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})\widehat{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})^{-1}\right),$$

where $\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ and $\widehat{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ are the sample approximations to $\boldsymbol{Q}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ and $\boldsymbol{G}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$, respectively:

$$\begin{aligned}\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log q(\boldsymbol{x}_i;\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log q(\boldsymbol{x}_i;\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}},\\ \widehat{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log q(\boldsymbol{x}_i;\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}}.\end{aligned}$$

The law of large numbers and consistency of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ yield that $\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ and $\widehat{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ converge in probability to $\boldsymbol{Q}(\boldsymbol{\theta}^*)$ and $\boldsymbol{G}(\boldsymbol{\theta}^*)$, respectively:

$$\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) \xrightarrow{\mathrm{p}} \boldsymbol{Q}(\boldsymbol{\theta}^*) \quad\text{and}\quad \widehat{\boldsymbol{G}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) \xrightarrow{\mathrm{p}} \boldsymbol{G}(\boldsymbol{\theta}^*).$$

It is known that the expected TIC divided by n converges to J and its error has smaller order than n^{-1}:

$$\frac{1}{n}\mathbb{E}[\mathrm{TIC}] = \mathbb{E}[J(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})] + o(n^{-1}).$$

On the other hand, if the negative average log-likelihood is naively used as an estimator of $J(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$,

$$\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n}\log q(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})\right] = \mathbb{E}[J(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})] + O(n^{-1})$$

holds, where $O(n^{-1})$ denotes a term that has the same size as n^{-1} (see Fig. 14.4). Thus, TIC is a more accurate estimator of $J(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ than the negative average log-likelihood. Since TIC does not require the assumption that the parametric model $q(\boldsymbol{x};\boldsymbol{\theta})$ contains

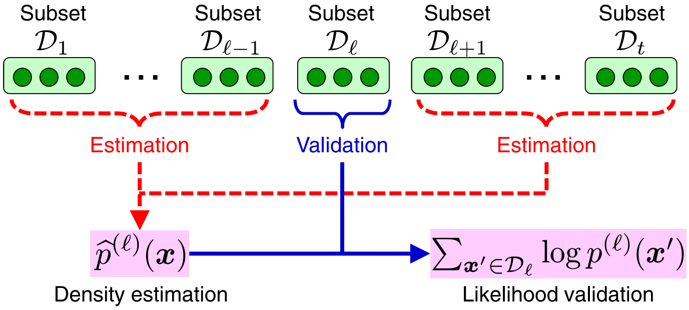
Subset $\mathcal{D}_1$
Subset $\mathcal{D}_{\ell-1}$
Subset $\mathcal{D}_\ell$
Subset $\mathcal{D}_{\ell+1}$
Subset $\mathcal{D}_t$
Estimation
Validation
Estimation
$\widehat{p}^{(\ell)}(\boldsymbol{x})$
Density estimation
$\sum_{\boldsymbol{x}'\in\mathcal{D}_\ell} \log p^{(\ell)}(\boldsymbol{x}')$
Likelihood validation

FIGURE 14.5

Cross validation.

the true probability density function $p(\boldsymbol{x})$, it is more general than AIC. However, since AIC is much simpler to implement, it seems to be more frequently used in practice than TIC.

14.4 CROSS VALIDATION

Although AIC is easy to implement, its validity is only guaranteed when the number of training samples is large. Here, a more flexible model selection method called *cross validation* is introduced.

The basic idea of cross validation is to split training samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ into the estimation samples and the validation samples. The estimation samples are used for estimating the probability density function, and the validation samples are used for evaluating the validity of the estimated probability density function. If the log-likelihood is used as the evaluation measure, the model that gives the largest log-likelihood for the validation samples is chosen as the most promising one.

However, simply splitting the training samples into the estimation subset and the validation subset for model selection can cause strong dependency on data splitting. To mitigate this problem, cross validation splits the training samples into t disjoint subsets of (approximately) the same size (see Fig. 14.5). Then $(t-1)$ subsets are used for estimation and the remaining subset is used for validation. The choice of $(t-1)$ subsets has t possibilities and thus all t possibilities are investigated and the average log-likelihood is regarded as the final score for model selection.

The algorithm of cross validation is summarized in Fig. 14.6.

14.5 DISCUSSION

Let us interpret the statistical meaning of cross validation using the *KL divergence* from true probability density p to its estimator $\widehat{p}_j^{(\ell)}$:

```
1. Prepare candidates of models: $\{\mathcal{M}_j\}_j$.
2. Split training samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ into $t$ disjoint subsets of
   (approximately) the same size: $\{\mathcal{D}_\ell\}_{\ell=1}^t$.
3. For each model candidate $\mathcal{M}_j$
   (a) For each split $\ell = 1, \ldots, t$
       i. Obtain density estimator $\widehat{p}_j^{(\ell)}(\boldsymbol{x})$ using model $\mathcal{M}_j$ from all
          training samples without $\mathcal{D}_\ell$.
       ii. Compute the average log-likelihood $J_j^{(\ell)}$ of $\widehat{p}_j^{(\ell)}(\boldsymbol{x})$ for holdout
           samples $\mathcal{D}_\ell$:
           $$J_j^{(\ell)} = \frac{1}{|\mathcal{D}_\ell|} \sum_{\boldsymbol{x}' \in \mathcal{D}_\ell} \log \widehat{p}_j^{(\ell)}(\boldsymbol{x}'),$$
           where $|\mathcal{D}_\ell|$ denotes the number of elements in set $\mathcal{D}_\ell$.
   (b) Compute the average log-likelihood $J_j$ over all $t$ splits:
       $$J_j = \frac{1}{t} \sum_{\ell=1}^{t} J_j^{(\ell)}.$$
4. Choose the model $\mathcal{M}_{\widehat{j}}$ that maximizes the average log-likelihood:
   $$\widehat{j} = \operatorname*{argmax}_j J_j.$$
5. Obtain the final density estimator using chosen model $\mathcal{M}_{\widehat{j}}$, from all
   training samples $\{\boldsymbol{x}_i\}_{i=1}^n$.
```

FIGURE 14.6

Algorithm of likelihood cross validation.

$$\mathrm{KL}(p\|\widehat{p}) = E\left[\log \frac{p(\boldsymbol{x})}{\widehat{p}_j^{(\ell)}(\boldsymbol{x})}\right] = E\left[\log p(\boldsymbol{x})\right] - E\left[\log \widehat{p}_j^{(\ell)}(\boldsymbol{x})\right].$$

Since the first term is constant, let us ignore it and define the second term as J:

$$J = -E[\log \widehat{p}_j^{(\ell)}(\boldsymbol{x})].$$

The cross validation score $J_j^{(\ell)}$ (see Fig. 14.6) is known to be an almost unbiased estimator of J [118]. Thus, model selection by likelihood cross validation corresponds to finding the model that minimizes the KL divergence, which is the same objective as the AIC explained in Section 14.3.

Indeed, the likelihood cross validation and the AIC are known to perform similarly when the number of training samples is large [97]. However, for a small number of training samples, likelihood cross validation seems to be more reliable in practice. Furthermore, while the AIC is applicable only to model selection of maximum likelihood estimation under the KL divergence, cross validation is applicable to any estimation method and any error metric. For example, as shown in Fig. 16.17, cross validation can be applied to choosing tuning parameters in pattern recognition under the misclassification rate criterion. Thus, cross validation is a useful alternative to the AIC. A practical drawback of cross validation is that it is computationally expensive due to repeated estimation and validation. However, note that the cross validation procedure can be easily parallelized for multiple servers or cores.