Statistic Learning Theory

Weixin Wang and Xiaoke Zhang

University of Science and Technology of China

2021.10.13

Table of Contents

Table of Contents

Preliminaries

In classical statistics, we usually need to compare different statistical procedures for the same problem.
Examples:
For Poisson variable $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$, we can estimate λ by point estimator: $\hat{\lambda_1} = \overline{X_n}$ or $\hat{\lambda_2} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X_n}\right)^2$, which are both unbiased.

Preliminaries

- Q:How do we choose one among them? Which is better?
  $\Rightarrow$ Decision Theory
  (a formal theory for comparing statistical procedures)

We first collect the data $X \in \mathcal{X}$. Consider the parameter θ which lies in parameter space Θ. Let $\hat{\theta}$ be an estimator of θ, which is a map $\mathcal{X} \to \Theta$.
Remark: $\hat{\theta}(X)$ is random but not $\hat{\theta}$, but we usually abbreviate $\hat{\theta}(X)$ as $\hat{\theta}$.

*In the language of decision theory, an estimator is sometimes called a decision rule and the possible values of the decision rule are called actions.

Preliminaries

Goal: find a $\hat{\theta}$ which is close to the true value of θ.

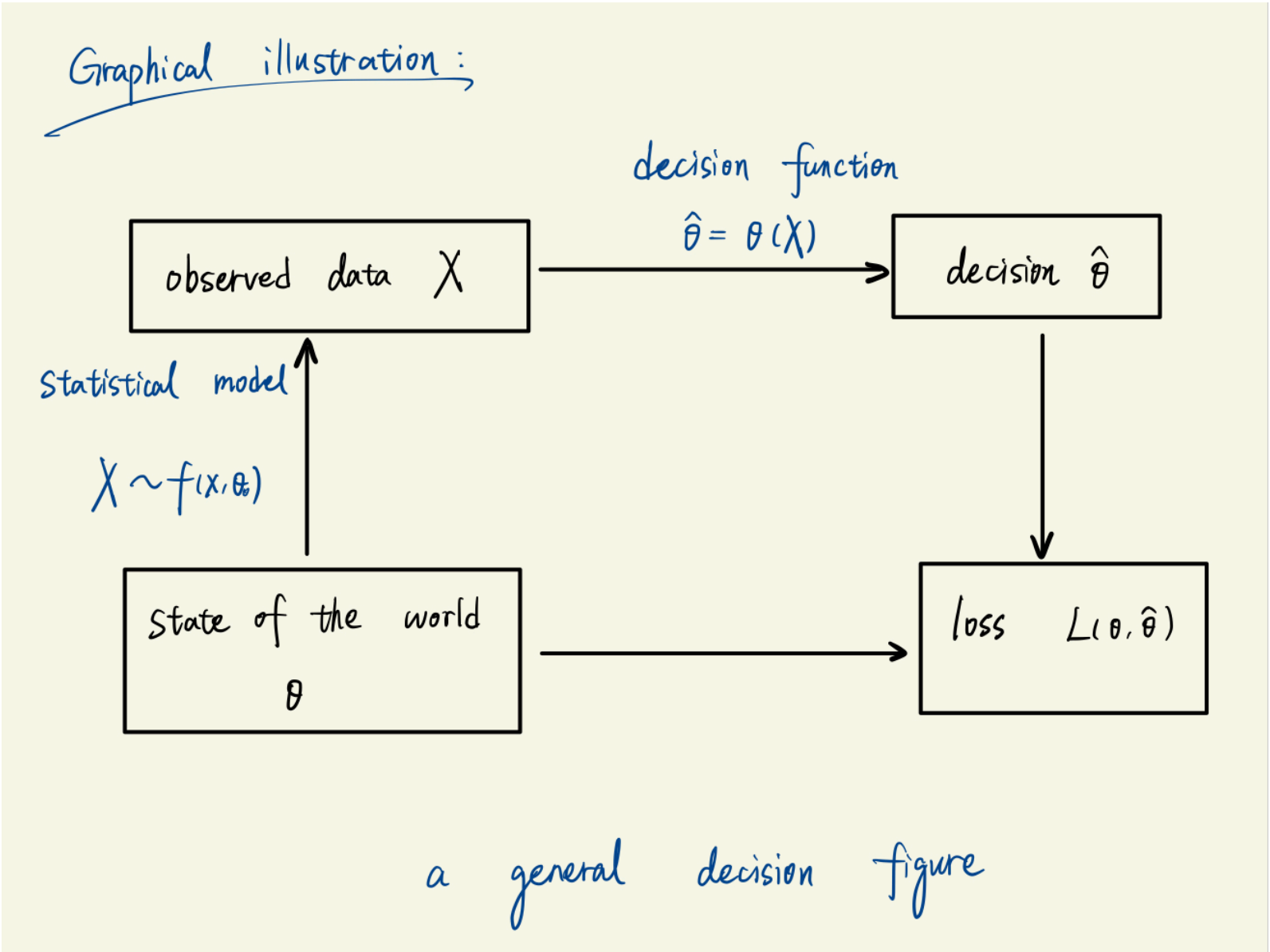
We need a function to measure the discrepancy between θ and $\hat{\theta}$

- loss function: $L(\theta, \hat{\theta}) : \Theta \times \Theta \to \mathbb{R}^+$.

Loss is larger if the difference between our estimate and the true value is larger.

- $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ square error loss
- $L(\theta, \hat{\theta}) = 0$ if $\theta = \hat{\theta}$ or 1 if $\theta \neq \hat{\theta}$ zero-one loss

Preliminaries



Risk Function

- The risk of $\hat{\theta}$: $R(\theta, \hat{\theta}) = E_\theta\left(L(\theta, \hat{\theta})\right) = \int L\left(\theta, \hat{\theta}\,(x)\right) f(x;\theta)\, dx$
- In particular, when $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

$$R\left(\theta, \hat{\theta}\right) = E_\theta(\theta - \hat{\theta})^2 = V_\theta(\hat{\theta}) + bias_\theta^2(\hat{\theta}).$$

Risk Function $\rightarrow$ intermediate object in evaluating a decision function

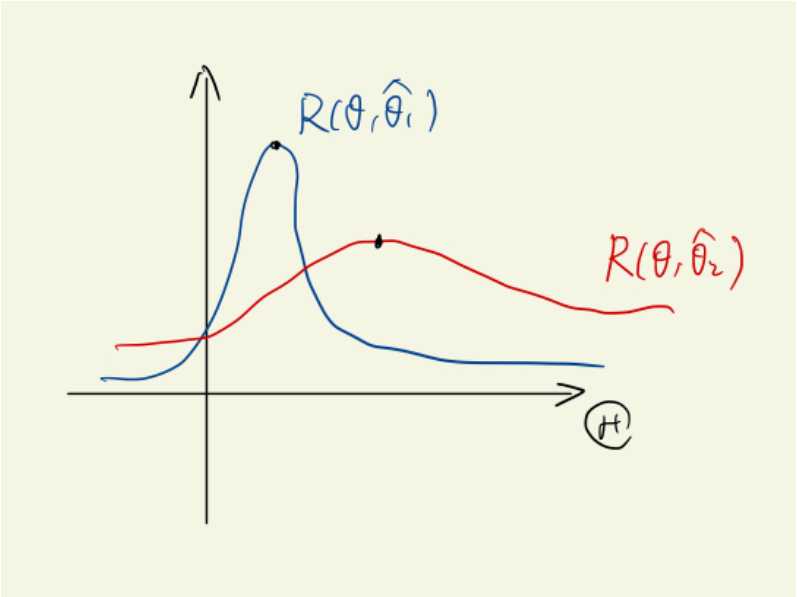
small $R \Leftrightarrow$ good $\hat{\theta}$.

$\hat{\theta}$ might be good for some θ, bad for other θ, decision theory deals with this trade-off.

Comparing Risk Functions

Q : how to compare risk function?

We need a one-number summary of the risk function:



The risk function given by $\hat{\theta}$ is a curve of θ, we usually evaluate it via a functional of curve.

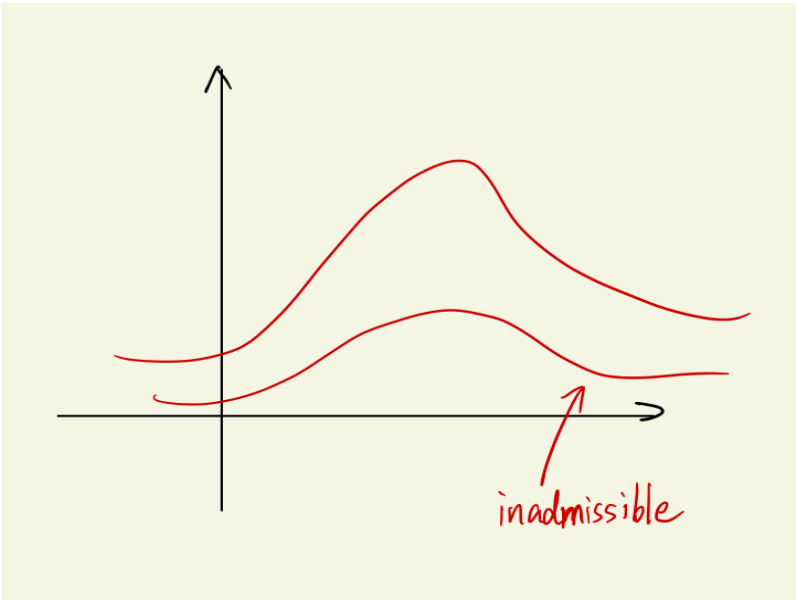
- the worse case (Minimax)
- the average case (Bayesion)

Admissibility

An estimator $\hat{\theta}$ is inadmissible if there exists another rule $\hat{\theta}'$ such that

- $R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta})$ for all θ and
- $R(\theta, \hat{\theta}') < R(\theta, \hat{\theta})$ for at least one θ.

otherwise, θ is admissible.



Stein's Paradox

Suppose that $X_1, ..., X_n \sim N(\theta, 1)$, we consider evaluate the estimator of θ with squared error loss. We know that $\hat{\theta} = \bar{X}$ is admissible. [**?**]
Now we assume $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_k)$ and the loss function is taken as $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{k} (\theta_j - \hat{\theta}_j)^2$.

Stein astounded everyone when he proved that, if $k \geq 3$, then $\hat{\boldsymbol{\theta}} = \boldsymbol{X}$ is inadmissible. It can be shown that the James-Stein estimator $\hat{\theta}^S$ has smaller risk, where $\hat{\theta}^S = (\hat{\theta}_1^S, ..., \hat{\theta}_k^S)$

$$\widehat{\theta}_i^S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i$$

and $(z)^+ = max\{z, 0\}$.

James–Stein estimator

Seeing the James–Stein estimator as an empirical Bayes method gives some intuition to this result: One assumes that $\boldsymbol{\theta}$ itself is a random variable with prior distribution $\sim N(\mathbf{0}, \boldsymbol{A})$, where $\boldsymbol{A}$ is estimated from the data itself. Estimating $\boldsymbol{A}$ only gives an advantage compared to the maximum-likelihood estimator when the dimension m is large enough; hence it does not work for $m \leq 2$. The James–Stein estimator is a member of a class of Bayesian estimators that dominate the maximum-likelihood estimator. [**?**]

Table of Contents

History and background of learning theory

Vapnik (2000) points out the interesting fact that the 1960s saw four major developments that were to have lasting influence on learning theory.

- First, Tikhonov and others developed regularization theory for the solution of ill-posed inverse problems. Regularization theory has had and continues to have tremendous impact on learning theory.
- Regularization: regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting.

History and background of learning theory

Well-posed problem means a mathematical models have the properties:

- a solution exists
- the solution is unique
- the solution's behaviour changes continuously with the initial conditions.

Problems that are not well-posed in this sense are termed ill-posed. Inverse problems are often ill-posed.

History and background of learning theory

- Second, Rosenblatt, Parzen, and Chentsov pioneered nonparametric methods for density estimation. These beginnings were crucial for a distribution free theory of non-parametric classification and regression to emerge later.

History and background of learning theory

- Third, Vapnik and Chervonenkis proved the uniform law of large numbers for indicators of sets with finite Vapnik-Chervonenkis (VC) dimension.

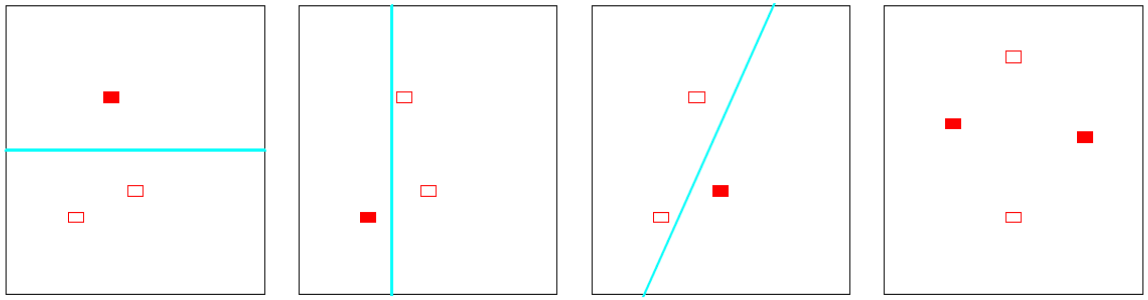


FIGURE 7.6. *The first three panels show that the class of lines in the plane can shatter three points. The last panel shows that this class cannot shatter four points, as no line will put the hollow points on one side and the solid points on the other. Hence the VC dimension of the class of straight lines in the plane is three. Note that a class of nonlinear curves could shatter four points, and hence has VC dimension greater than three.*

History and background of learning theory

- Fourth, Solomono, Kolmogorov, and Chaitin all independently discovered algorithmic complexity: the idea that the complexity of a string of 0's and 1's can be defined by the length of the shortest program that can generate that string. This later led Rissanen to propose his Minimum Description Length (MDL) principle.

History and background of learning theory

Kolmogorov complexity:
Take the following two strings (length 64) as an example.

```
01010101010101010101010101010101
10101010101010101010101010101010
```

```
11001000011000011101111011101100111
11010010000100101011110010110
```

The first string can be briefly described as "repeat 32 '01's". The second string has no obvious short description.
algorithmic complexity: the idea that the complexity of a string of 0's and 1's can be defined by the length of the shortest program that can generate that string.

History and background of learning theory

Minimum Description Length (MDL) principle:
Suppose first that the possible messages we might want to transmit are $z_1, z_2, \ldots, z_m$.
Here is an example with four possible messages and a binary coding:

Message	z_1	z_2	z_3	z_4
Code	0	10	110	111

for example we use codes 110, 10, 111,0 for z_1, z_2, z_3, z_4. How do we decide which to use? It depends on how often we will be sending each of the messages. If, for example, we will be sending z1 most often, it makes sense to use the shortest code 0 for z_1. Using this kind of strategy—shorter codes for more frequent messages—the average message length will be shorter.

Element Framework

We assume that $(X_1, Y_1), ..., (X_n, Y_n)$ the *i.i.d* samples and we want to get f_0 possesses the best prediction behavior, i.e.

$$f_0 := \underset{f}{argmin} E\left(L(Y, f(X))\right) \stackrel{\Delta}{=} \underset{f}{argmin} R(f)$$

where $L(\cdot, \cdot)$ is a loss function.

Examples

Proof that $f_0(X) = E[Y|X]$ for square error loss and f_0 is the Bayesian classification rule for zero-one loss.

Square error loss

① $E(f(x)-Y)^2 = E[E(f(x)-Y)^2]|X]$

$\Rightarrow E[(f(x)-Y)^2 | X=x] = f(x)^2 - 2f(x)\cdot E[Y|X=x] + E[Y^2|X=x]$

二次函数： $f(x) = E[Y|X=x] \Rightarrow f(X) = E[Y|X]$

$f(x) = E[Y|X] = \arg\min_f E[L(Y, f(X))]$

Examples

Proof that $f_0(X) = E[Y|X]$ for square error loss and f_0 is the Bayesian classification rule for zero-one loss.

Zero-one loss

② $Y \in \{1, \cdots k\}$

$$E[L(f(x) - Y)] = E\left[E[L(f(x) - Y)|X]\right]$$

$$\Rightarrow E[L(f(x) - Y)|X = x] = \sum_{i=1}^{k} 1(f(x) = i) \cdot P(Y = i|X = x)$$

$$f(x) = \arg\min_i P(Y = i|X = x)$$

$$\Rightarrow f(X) = \arg\min_i P(Y = i|X)$$

Element Framework

Limitations: the expectation can't be calculated since we don't know the joint distribution of (X, Y), while we can extract f from

$$R_n(f) := \frac{1}{n} \sum_{k=1}^{n} L(Y_k, f(X_k))$$

which is called the emprical risk.

In practice, we should confine f into some hypothesis spaces $\mathcal{H}$, i.e.

$$\hat{f}_n := \underset{f \in \mathcal{H}}{argmin}\ R_n(f),$$

since the f that interpolates $(X_1, Y_1), ..., (X_n, Y_n)$ may be not unique and meaningless (over-fitting).

Examples about the hypothesis spaces for regression and classification

Linear Functions This is one of the simplest functions classes. In any finite dimensional Euclidean space $\mathbb{R}^d$, define the class of real valued functions

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle \ : \ w \in \mathcal{W} \subseteq \mathbb{R}^d\} \ ,$$

where $\langle w, x \rangle = \sum_{j=1}^{d} w_j x_j$ denotes the standard *inner product*. The *weight vector* might be additionally constrained. For instance, we may require $\|w\| \leq W$ for some norm $\|\cdot\|$. The set $\mathcal{W}$ takes care of such constraints.

[?]

Examples about the hypothesis spaces for regression and classification

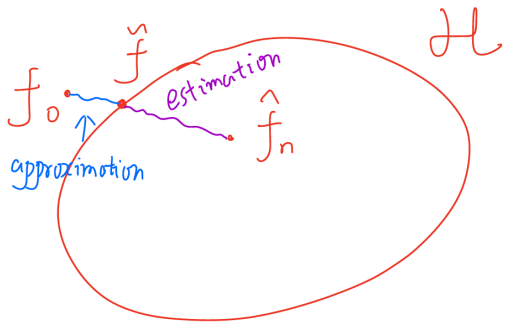
Linear Threshold Functions or Halfspaces This is class of binary valued function obtained by thresholding linear functions at zero. This gives us the class

$$\mathcal{F} = \{x \mapsto \text{sign}(\langle w, x \rangle) \ : \ w \in \mathcal{W} \subseteq \mathbb{R}^d\} \ .$$

[?]

Excess Risk

- Excess Risk: $R(\hat{f}_n) - R(f_0)$, we want to minimize the excess risk



- $R(\hat{f}_n) - R(f_0) = \underbrace{R(\hat{f}_n) - R(\hat{f})}_{\text{estimation error (random)}} + \underbrace{R(\hat{f}) - R(f_0)}_{\text{approximation error}}$,

  where $\hat{f} := \underset{f \in H}{\text{argmin}} R(f)$

When the $\mathcal{H}$ is "larger", the estimation error is increased with high probability while the approximation error is smaller. The optimal hypothesis space is preferable balancing two types of error.

Table of Contents

Why model selection?

- Are models ever true?
  "All models are wrong, but some are usefull!"—George Box
  Models are only approximations to unknown reality or truth.
- What is model selection?
  Model selection is the task of selecting a statistical model from a model class, given a set of data. For example, selecting variables for linear regression, order of AR model, number of components in a mixture model and so on.
- Why model selection?
  Extract the true information in the data. An improper choice of model or method can lead to purely noisy discoveries, disappointing predictive performances or severely misleading conclusions.

Why model selection?

- Cubic polynomial model

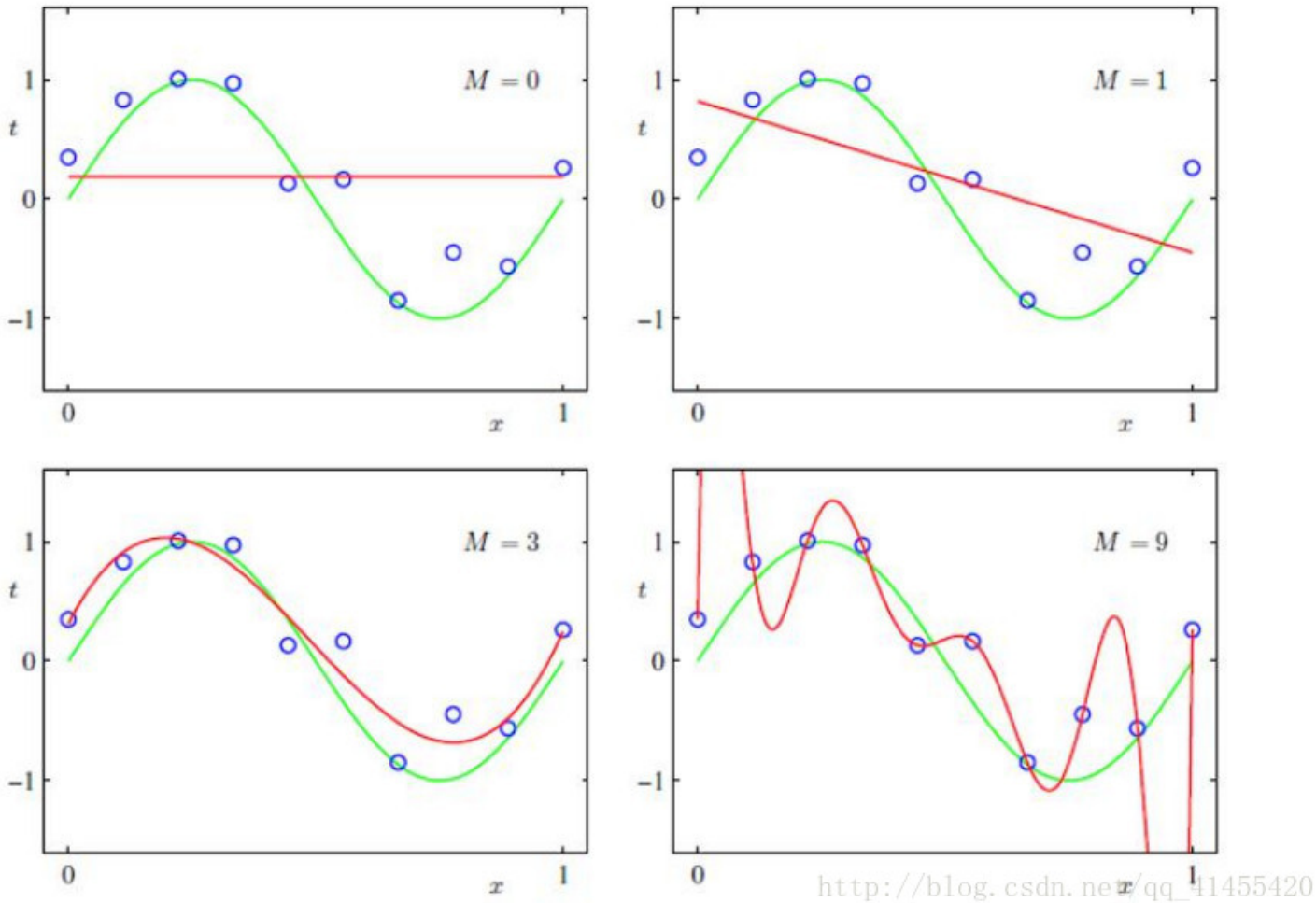


Figure: Polynomial models

Basic concepts

- Candidate model:
  $\mathcal{M}_m = \{p_{\theta_m}; \theta_m \in \mathcal{H}_m\}$, θ_m is the parameter, $\mathcal{H}_m$ is the parameter space associated with $\mathcal{M}_m$, and p_{θ_m} is the density function with parameter θ_m.
- Candidate model set:
  $\{\mathcal{M}_m\}_{m \in \mathbb{M}}$, which is a collection of models indexed by $m \in \mathbb{M}$.

Typical data analysis

Two steps:

- Fititing model:
  For each candidate model $\mathcal{M}_m$, fit all of the observed data to that model by estimating its parameter $\theta_m \in \mathcal{H}_m$.
- Selecting model:
  Once we have a set of estimated candidate models $p_{\tilde{\theta}_m}(m \in \mathbb{M})$, select the most appropriate one for either interpretation or prediction.

Model fitting

By minimizing the loss function.

- Loss function: $s(p_{\theta_m}, z_t)$.
  p_{θ_m} represent the density function and z_t is the observation. A commonly used loss function is logarithmic loss:

$$s(p, z_t) = -\log p(z_t),$$

  which is the maximum likelihood estimation (MLE).
- Parameter estimation: $\tilde{\theta}_m = \underset{\theta_m \in \mathcal{H}_m}{\arg\min} \sum_{t=1}^{n} s(p_{\theta_m}, z_t)$.

Model selection

We assumed that there is a single correct or at least, best model, and that model suffices as the sole model for making inferences from the data. Although the identity of that model is unknown, it seems to be assumed that it can be estimated—in fact, well estimated.

- p_*: the true density with respect to the true distribution.
- E_* :the true expectation with respect to the true distribution.
- Out-sample prediction loss: let $p_{\hat{\theta}_m} = \hat{p}_m$,

$$E_* \left(s \left(\hat{p}_m, Z \right) \right) = \int s \left(\hat{p}_m(z), z \right) p_*(z) dz.$$

Theoretical examinations of model selection criteria

For inference: identify the best model for the data, which hopefully provides a reliable characterization of the sources of uncertainty for scientific insight and interpretation.

- Consistency: the quasi-true model (the most parsimonious model that is closest to the true model as measured by the K-L information) is selected with probability going to 1 as n $\rightarrow \infty$.
- Asymptotically efficiency (weaker concept): demand that the loss of the selected model or method is asymptotically equivalent to the smallest among all of the candidates.

$$\frac{\min_{m \in \mathbb{M}} \mathcal{L}_m}{\mathcal{L}_{\hat{m}}} \xrightarrow{p} 1$$

as $n \rightarrow \infty$, where $\mathcal{L}_m = E_* \left(s \left(\hat{p}_m, Z \right) \right) - E_* \left(s \left(p_*, Z \right) \right)$ and $\hat{m}$ denotes the selected model.

Theoretical examinations of model selection criteria

For prediction: choose a model or method that offers top performance.

- Minimax optimal: In the learning theory framework, we have i.i.d. (x_i, Y_i) generating from $Y_i = f(x_i) + \varepsilon_i$, where $f \in \mathcal{F}$, $\mathcal{F}$ is the hypothesis space which is unknown. A selection method v is said to be minimax-rate optimal over $\mathcal{F}$ if

$$\sup_{f \in \mathcal{F}} R(f, v, n) = n^{-1} \sum_{i=1}^{n} E_Y \left\{ \hat{f}_v(x_i) - f(x_i) \right\}^2$$

converges at the same rate as the minimax risk:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^{n} E_Y \left(\hat{f}(x_i) - f(x_i) \right)^2,$$

where $\hat{f}_v$ is the least square estimators (based upon observational data) selected by v criterion among a class of candidate models.

Example

- Case 1 (Parametric framework): AR(3) model.

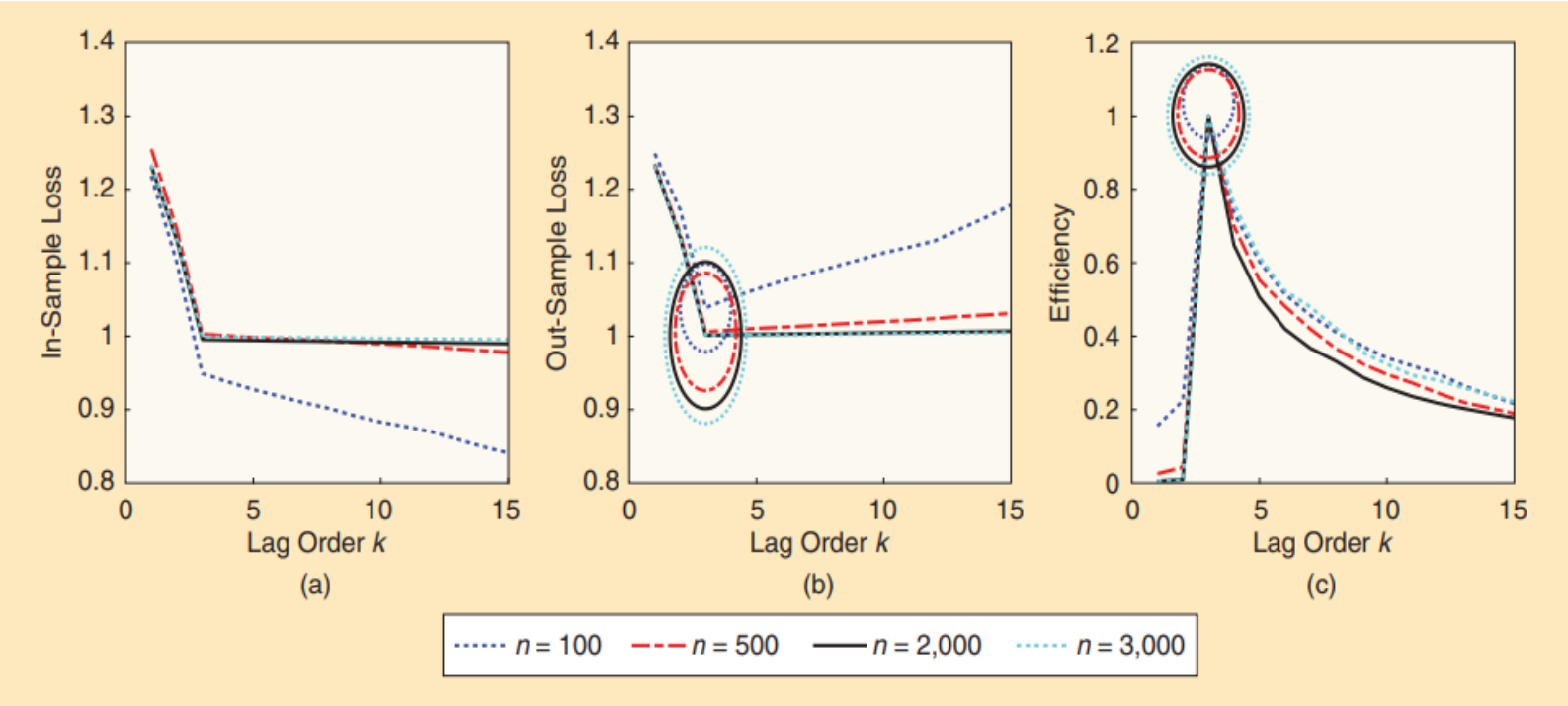


Figure: Parametric framework: AR(3) model

Example

- Case 2 (Nonparametric framework): MA(1) model.

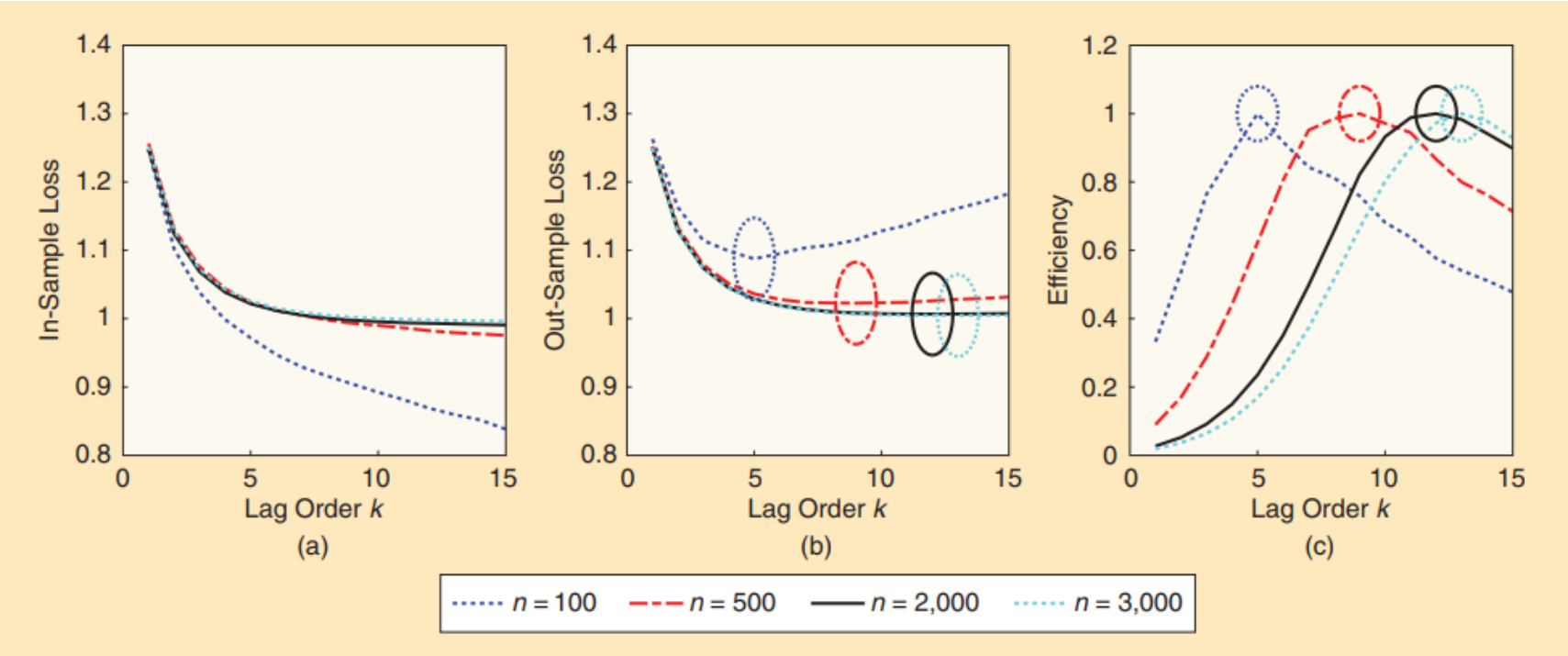


Figure: Nonparametric framework: MA(1) model

Example

- Case 3 (Practically nonparametric framework): AR(10) model.

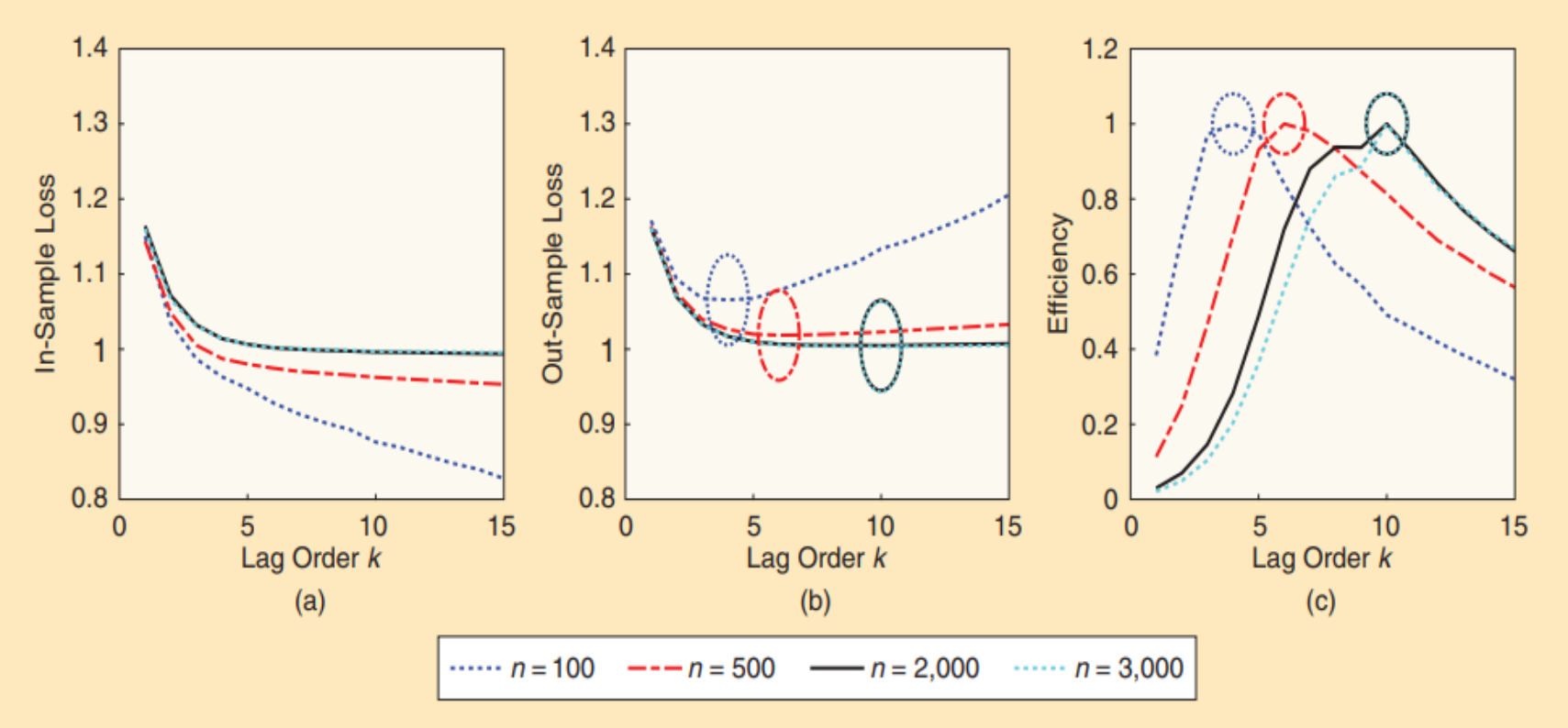


Figure: Practically Nonparametric framework: AR(10) model

The Best Model

Model selection methods:

1. Akaike information criterion (based on information theory).
2. Bayesian information criterion (based on Bayesian approaches).
3. Cross validation.

Kullback-Leibler Information

K-L Information is a measure of distance between full reality and a model. It's the information loss when model g is used to approximate f.

- K-L Information: $I(f,g) = \int f(x) \ln\left(\frac{f(x)}{g(x|\theta)}\right) dx \geq 0$ and $I(f,g) = 0$ if and only if $f(x) = g(x \mid \theta)$. We use the fact that

$$-\ln\lambda \geq 1-\lambda \quad \text{for any } \lambda > 0 \quad \text{and}$$

$$-\ln\lambda = 1-\lambda \quad \text{if and only if } \lambda = 1$$

Hence, letting $\lambda(x) = g(x \mid \theta)/f(x)$,

$$\begin{aligned} I(f,g) &= \int f(x)[-\ln\lambda(x)]dx \\ &\geq \int f(x)\left[1-\lambda(x)\right]dx \\ &= \int f(x)\left[1-\frac{g(x \mid \theta)}{f(x)}\right]dx = 0. \end{aligned}$$

Akaike Information Criterion

Noting that

$$\begin{aligned} I(f,g) &= \int f(x)\ln\left(f(x)\right)dx - \int f(x)\ln\left(g\left(x \mid \theta\right)\right)dx \\ &= E_f\left[\ln\left(f(x)\right)\right] - E_f\left[\ln\left(g\left(x \mid \theta\right)\right)\right] \\ &= C - E_f\left[\ln\left(g\left(x \mid \theta\right)\right)\right]. \end{aligned}$$

Let

$$\theta := \text{argmin}_{\theta' \in \mathcal{H}} E_f\left[\ln\left(g\left(x \mid \theta'\right)\right)\right]$$

be the optimal parameter of this model minimizing the K-L information of data, which is estimated by the MLE of θ given x: $\hat{\theta}_x$.

Akaike Information Criterion

The $E_f[\ln(g(x \mid \theta))]$ cannot be used directly in model selection because it requires knowledge of true model f. A approximation of $E_f[\ln(g(x \mid \theta))]$ is to use the Taylor expansion and $n \to \infty$

$$\ln(g(x \mid \theta)) \approx \ln\left(g(x \mid \hat{\theta}_x)\right) - \frac{1}{2}\left(\theta - \hat{\theta}_x\right)^T H_n(\hat{\theta}_x)\left(\theta - \hat{\theta}_x\right)$$

where $H_n(\hat{\theta}_x)$ is the negative Hessian matrix and would converge to Fisher information $I_n(\theta)$. Moreover,

$$E_f\left[\frac{1}{2}\left(\theta - \hat{\theta}_x\right)^T I_n(\theta_x)\left(\theta - \hat{\theta}_x\right)\right] = \frac{1}{2}\text{tr}\left(I_n(\theta)E_f\left[(\theta - \hat{\theta}_x)(\theta - \hat{\theta}_x)^T\right]\right) \approx \frac{K}{2},$$

where K is the number of parameters. Therefore, $-2\ln\left(g(x \mid \hat{\theta}_x)\right) + K$ is the asymptotic unbiased estimator for $-2E_f[\ln(g(x \mid \theta))]$.

Akaike Information Criterion

Instead, we use the out-sample prediction loss

$$-2E_x E_y \left[\ln \left(g(x \mid \hat{\theta}_y) \right) \right]$$

is proposed to select the optimal model, where y is the fictious data vector with the same size n and the same pdf as $x \sim f$ but which is independent with x.

Akaike Information Criterion

- Taylor expansion:

$$\ln g\left(x \mid \hat{\theta}_y\right) \approx \ln g\left(x \mid \hat{\theta}_x\right) - \frac{1}{2}\left(\hat{\theta}_y - \hat{\theta}_x\right)^T I_n(\theta)\left(\hat{\theta}_y - \hat{\theta}_x\right).$$

- Calculate Expectation: $E_x E_y\left[\left(\hat{\theta}_y - \hat{\theta}_x\right)^T I_n(\theta)\left(\hat{\theta}_y - \hat{\theta}_x\right)\right]$

$$\begin{aligned}
&E_x\left\{E_y\left[(\hat{\theta}_y - \hat{\theta}_x)^T I_n(\theta)(\hat{\theta}_y - \hat{\theta}_x)\right]\right\} \\
&= \operatorname{tr}\left(E_x\left\{E_y\left[(\hat{\theta}_y - \hat{\theta}_x)^T I_n(\theta)(\hat{\theta}_y - \hat{\theta}_x)\right]\right\}\right) \\
&= \operatorname{tr}\left(E_x E_y\left\{I_n(\theta)\left[(\hat{\theta}_y - \theta) - (\hat{\theta}_x - \theta)\right]\left[(\hat{\theta}_y - \theta) - (\hat{\theta}_x - \theta)\right]^T\right\}\right) \\
&\approx \operatorname{tr}\left(I_n(\theta)\left\{E_y\left[(\hat{\theta}_y - \theta)(\hat{\theta}_y - \theta)^T\right] + E_x\left[(\hat{\theta}_x - \theta)(\hat{\theta}_x - \theta)^T\right]\right\}\right) \\
&\approx \operatorname{tr}\left(I_n(\theta)(I_n^{-1}(\theta_y) + I_n^{-1}(\theta))\right) \\
&= \operatorname{tr}\left(2I_K\right) = 2K.
\end{aligned}$$

Akaike Information Criterion

- An approximately unbiased estimator of $-2E_xE_y\left[\ln(g(x \mid \hat{\theta}_y))\right]$:

$$-2\ln g\left(x \mid \hat{\theta}_x\right) + 2K,$$

  which is called the Akaike's information criterion (AIC).
- Trade off fitting and complexity.
- The first term represents the quality of model fitting and tends to decrease as more parameters are added.
- The second term is a penalty and gets larger as more parameters are added.

Example

- Simple linear model: $y = x^T\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad x \perp \epsilon$
  There is a sample data $(x_1, y_1), \ldots, (x_n, y_n)$. The log-likelihood function is:

$$l(\theta \mid (x_1, y_1) \ldots, (x_n, y_n)) = \ln\left(\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{1}{2\sigma^2}\sum(y_i - x_i^T\beta)^2\right)\right)$$
$$= -\frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - x_i^T\beta)^2 + c.$$

  The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - x_i^T\beta)^2.$$

  The maximum likelihood function is

$$l\left(\hat{\theta}_{MLE} \mid (x_1, y_1), \ldots, (x_n, y_n)\right) = -\frac{n}{2}\ln\hat{\sigma}^2 + c'.$$

  Thus AIC $= n\ln\hat{\sigma}^2 + 2K, \quad K = p + 1.$

Bayesian Informatin Criterion

Applying Bayes Theorem, the joint posterior of $\mathcal{M}_m$ and θ_m can be written as

$$P\left((\mathcal{M}_m, \theta_m) \mid x\right) \propto \pi(\mathcal{M}_m) h\left(\theta_m \mid \mathcal{M}_m\right) L\left(\theta_m \mid x\right).$$

- $\pi(\mathcal{M}_m)$ denote a discrete prior over the models $\{\mathcal{M}_m, m \in \mathbb{M}\}$.
- $h\left(\theta_m \mid \mathcal{M}_m\right)$ denotes a prior on θ_m given the model $\mathcal{M}_m$.

A Bayesian model selection rule aims to choose the model which is a posteriori most probable. The posterior probability for $\mathcal{M}_m$ is

$$P\left(\mathcal{M}_m \mid x\right) \propto \pi(\mathcal{M}_m) \int_{\mathcal{H}_m} L\left(\theta_m \mid x\right) h\left(\theta_m \mid \mathcal{M}_m\right) \mathrm{d}\theta_m.$$

Bayesian Informatin Criterion

$$\ln P(\mathcal{M}_m \mid x) \propto \ln\{\pi(\mathcal{M}_m)\} + \ln\left\{\int L(\theta_m \mid x)\, h(\theta_m \mid \mathcal{M}_m)\, \mathrm{d}\theta_m\right\}.$$

Similar to Taylor expansion in the previous section,

$$\begin{aligned} L(\theta_m \mid x) &:= \exp(l(\theta_m \mid x)) \\ &\approx L\left(\hat{\theta}_m \mid x\right) \exp\left\{-\frac{1}{2}\left(\theta_m - \hat{\theta}_m\right)^T H_n(\hat{\theta}_m)\left(\theta_m - \hat{\theta}_m\right)\right\}. \end{aligned}$$

where $H_n(\hat{\theta}_m)$ is the negative Hessian matrix. We therefore have the following approximation for the integral

$$\begin{aligned} &\int L(\theta_m \mid x)\, h(\theta_m \mid \mathcal{M}_m)\, \mathrm{d}\theta_m \\ \approx &L\left(\hat{\theta}_m \mid x\right) \int \exp\left\{-\frac{1}{2}\left(\theta_m - \hat{\theta}_m\right)^T H_n(\hat{\theta}_m)\left(\theta_m - \hat{\theta}_m\right)\right\} h(\theta_m \mid \mathcal{M}_m)\, d\theta_m. \end{aligned}$$

Bayesian Informatin Criterion

The likelihood $L\left(\hat{\theta}_m \mid x\right)$ should dominate the prior $h\left(\theta_m \mid \mathcal{M}_m\right)$ within a small neighborhood of $\hat{\theta}_m$, Outside of this neighborhood, $L\left(\hat{\theta}_m \mid x\right)$ and the exponential term should be small enough to force the corresponding integrands.
Therefore, it is defensible to simplify the justification by using the noninformative prior $h\left(\theta_m \mid \mathcal{M}_m\right) = 1$, we can evaluate the integral as

$$\int \exp\left\{-\frac{1}{2}\left(\theta_m - \hat{\theta}_m\right)^T H_n(\hat{\theta}_m)\left(\theta_m - \hat{\theta}_m\right)\right\} \mathrm{d}\theta_m = (2\pi)^{K/2}\left|H_n(\hat{\theta}_m)\right|^{-\frac{1}{2}}.$$

Bayesian Information Criterion

In most cases ,we have that

$$\begin{aligned}\ln|H_n(\hat{\theta}_m)| &= \ln\left|n\cdot\frac{1}{n}H_n(\hat{\theta}_m)\right| \\ &= K\ln n + \ln\left|\frac{1}{n}H_n(\hat{\theta}_m)\right| \\ &= K\ln n + \mathcal{O}(1).\end{aligned}$$

This lead to an approximation of $P(\mathcal{M}_m \mid x)$

$$\begin{aligned}P(\mathcal{M}_m \mid x) &\approx \ln\{\pi(\mathcal{M}_m)\} + \ln\left[L\left(\hat{\theta}_m \mid x\right)(2\pi)^{K/2}\left|H_n(\hat{\theta}_m)\right|^{-\frac{1}{2}}\right] \\ &\approx \ln L\left(\hat{\theta}_m \mid x\right) - \frac{1}{2}K\ln n.\end{aligned}$$

Bayesian Information Criterion

Bayesian Information Criterion (BIC): $-2 \ln L\left(\hat{\theta}_m \mid x\right) + K \log n$.

- Trade off fitting and complexity.
- The first term represents the quality of model fitting and tends to decrease as more parameters are added.
- The second term is a penalty and gets larger as more parameters are added. The penalty term is larger than that in AIC.

The war batween AIC and BIC

The war between AIC and BIC originates from two fundamentally different goals: one to minimize certain loss for prediction purpose and the other to select the best model for inference purpose.

Table 1. The AR order selection: The average efficiency, dimension, and PI (along with standard errors).

		AIC	BC	BIC
Case 1	Efficiency	0.78 (0.04)	0.93 (0.02)	0.99 (0.01)
	Dimension	3.95 (0.20)	3.29 (0.13)	3.01 (0.01)
	PI		0.93 (0.03)	
Case 2	Efficiency	0.77 (0.02)	0.76 (0.02)	0.56 (0.02)
	Dimension	9.34 (0.25)	9.29 (0.26)	5.39 (0.13)
	PI		0.13 (0.03)	
Case 3	Efficiency	0.71 (0.02)	0.67 (0.02)	0.55 (0.02)
	Size	6.99 (0.23)	6.61 (0.26)	4.02 (0.10)
	PI		0.35 (0.05)	

Figure: The AR order selection: The average efficiency, dimension

Properties

When the number of parameters in the true model is infinite or increases with increasing n or when the true model is not in the candidate model set, it is difficult to conceptualize consistency. There is no possibility of selecting the true model. In these circumstances one considers efficiency (or minimization) for some loss function.

- AIC:
  1. Minimax optimal + asymptotically efficient for non-parametric framework.
  2. The AIC will fail to select the true model with nonvanishing probability as n grows large, even when the true model is under consideration.
  3. In nonparametetric framework or practically parametric framework or if the true model is not in the candidate model set, the AIC is asymptotically efficient in MSE of estimation/prediction and in K-L divergence .

Properties

- BIC
  1. Consistent + asymptotically efficient for parametric framework.
  2. The BIC is efficient, however, when the true model is among the candidates.
  3. Consisitency requires some assumptions: (a) the true model is under consideration; (b) the true model's dimension (denoted 0) remains fixed as N grows; and (c) the number of parameters in the true model is finite.

Rational Choice of AIC and BIC

The choice between the AIC and BIC depends on one's purpose.

- If one's main purpose is for prediction, AIC may be better, and the model is more complex. For example, in a nonparametric framework, we are more interested in prediction.
- If one's main purpose is for inference, BIC may be better, and the model is more simple. For example, in a parametric framework, we are more interested in the structure of the model.

Validation (Hold-Out)

Training an algorithm and evaluatving its statistical performance on the same data yields an overoptimistic result. Usually, testing the output of the algorithm on new data would yield a good estimate of its performance.

- Splitting the data, for example, 70% training data and 30% testing data.
- Training from the training data and validated on the remaining data.

The training set can play the role of "new data" as long as data are i.i.d.. CV doesn't require the candidate models to be parametric, it works as long as the data are permutable and one can assess the predictive performance based on some measure.

Validation(Hold-Out)

For every model $\mathcal{M}_m$ and any $I^{(t)} \subset \{1, \ldots, n\}$ with cardinality n_t, the loss function of Hold-Out is

$$\hat{\mathcal{L}}^{\mathrm{HO}}\left(\mathcal{M}_m; z; I^{(t)}\right) = \frac{1}{n_v} \sum_{i \in I^{(v)}} s\left(p_{\hat{\theta}_{m, I^{(t)}}}, z_i\right).$$

Cross Validation

A single data split yields a validation estimate of the risk, and averaging over several splits yields a cross-validation estimate.

- Split the data into a training set of n_t and a validation set of n_v for several times.
- Each candidate model is then trained from the n_t data and validated on the remaining data.
- The model with the smallest average validation loss will be selected.

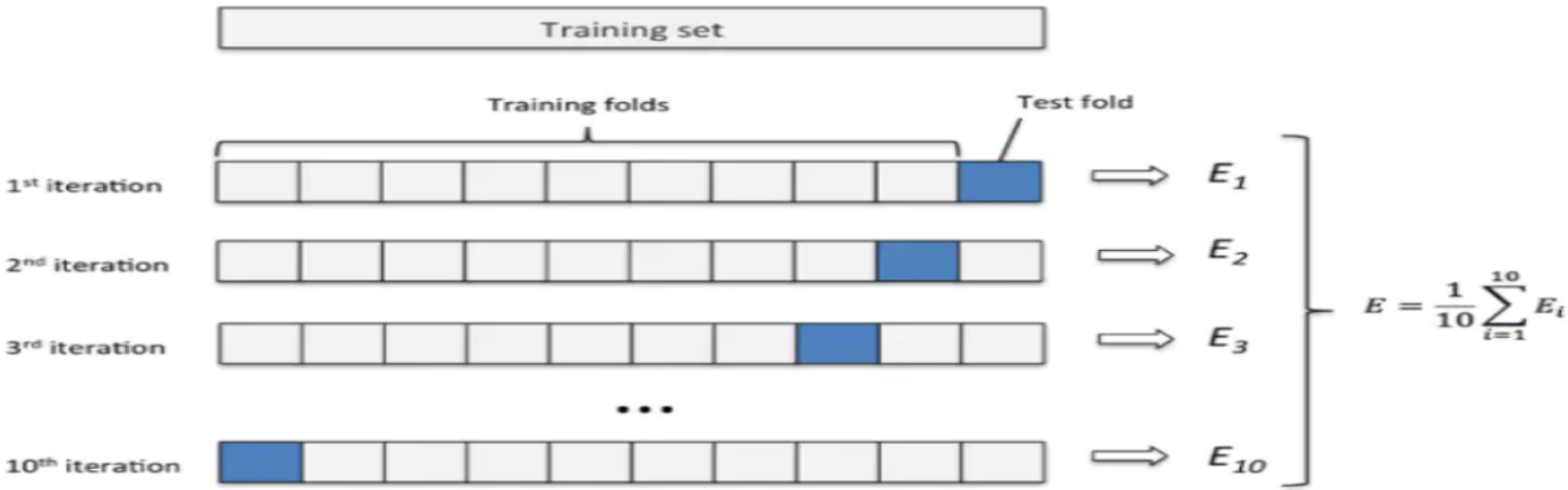


Figure: Cross Validation

Leave-one-out CV

For each candidate model $\mathcal{M}_m$:

- For the kth observation, fit the model using the remaining $n-1$ observations.
- Calculate the prediction error of the fitted model, and find the loss function

$$\hat{\mathcal{L}}^{\text{LOO}}\ (\mathcal{M}_m; z) = \frac{1}{n}\sum_{i=1}^{n} s\left(p_{\hat{\theta}_{m,-i}}, z_i\right).$$

k-fold CV

For each candidate model $\mathcal{M}_m$:

- Split data into k parts or folds

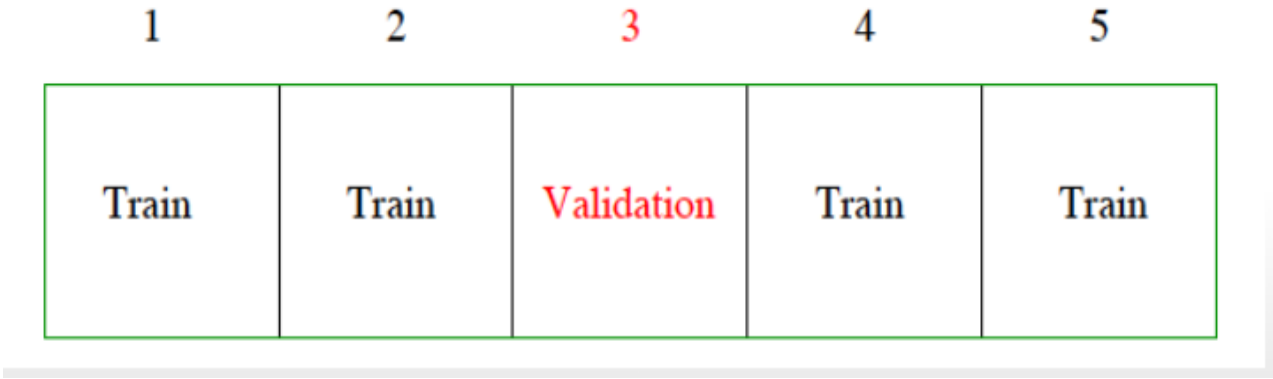


Figure: k-fold CV

- Use all but one fold to train your model/method
- Use the left out folds to make predictions
- Rotate around the roles of folds, k rounds total, calculate the loss function

$$\hat{\mathcal{L}}^{\mathrm{kF}}\left(\mathcal{M}_m; z; \{I_j\}_{1\le j\le k}\right) = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{1}{|I_j^c|}\sum_{i\in I_j^c} s\left(p_{\hat{\theta}_{m,I_j}}, z_i\right)\right].$$

Statistical properties of CV

- Bias
  The independence of training and validation samples implies that for every model $\mathcal{M}_m$ and any $I(t) \subset 1, ..., n$ with cardinality,

$$E\left[\hat{\mathcal{L}}^{\mathrm{HO}}\left(\mathcal{M}_m; z; I^{(t)}\right)\right] = E\left[s\left(p_{\hat{\theta}_{m,I^{(t)}}}, z_i\right)\right].$$

  Therefore, if $|I_j| = n_t$ for $j = 1, \ldots, k$, the expectation of the k-fold CV estimator of the risk only depends on n_t :

$$E\left[\hat{\mathcal{L}}^{\mathrm{kF}}\left(\mathcal{M}_m; z; \{I_j^{(t)}\}_{1\leq j\leq k}\right)\right] = E\left[s\left(p_{\hat{\theta}_{m,I_j}}, z_i\right)\right].$$

  The bias is the difference between the risks of $\mathcal{M}_m$ respectively trained with n_t and with n observations. The bias of CV is usually nonnegative and tends to decrease when n_t increases.

Cross Validation

Some notices:

1. The strongest argument for CV is its quasi-universality: Provided data are i.i.d..
2. When splitting the data, it is often recommended to take into account the structure of data when choosing the splits.
3. The model selection performances of CV can be less accurate, while its computational cost is higher.

Table of Contents

Summary

Some general recommendations:

1. If one needs to declare a model for inference, model selection consistency is the right concept to think about. If one's main goal is prediction, model selection instability is less of a concern, and any choice among the best performing models may give a satisfying prediction accuracy.
2. When model selection is for prediction, the minimax consideration gives more protection in the worst case.
3. When prediction is the goal, one may consider different types of models and methods and then apply cross validation to choose one for final prediction.

Theoretical examinations of model selection criteria

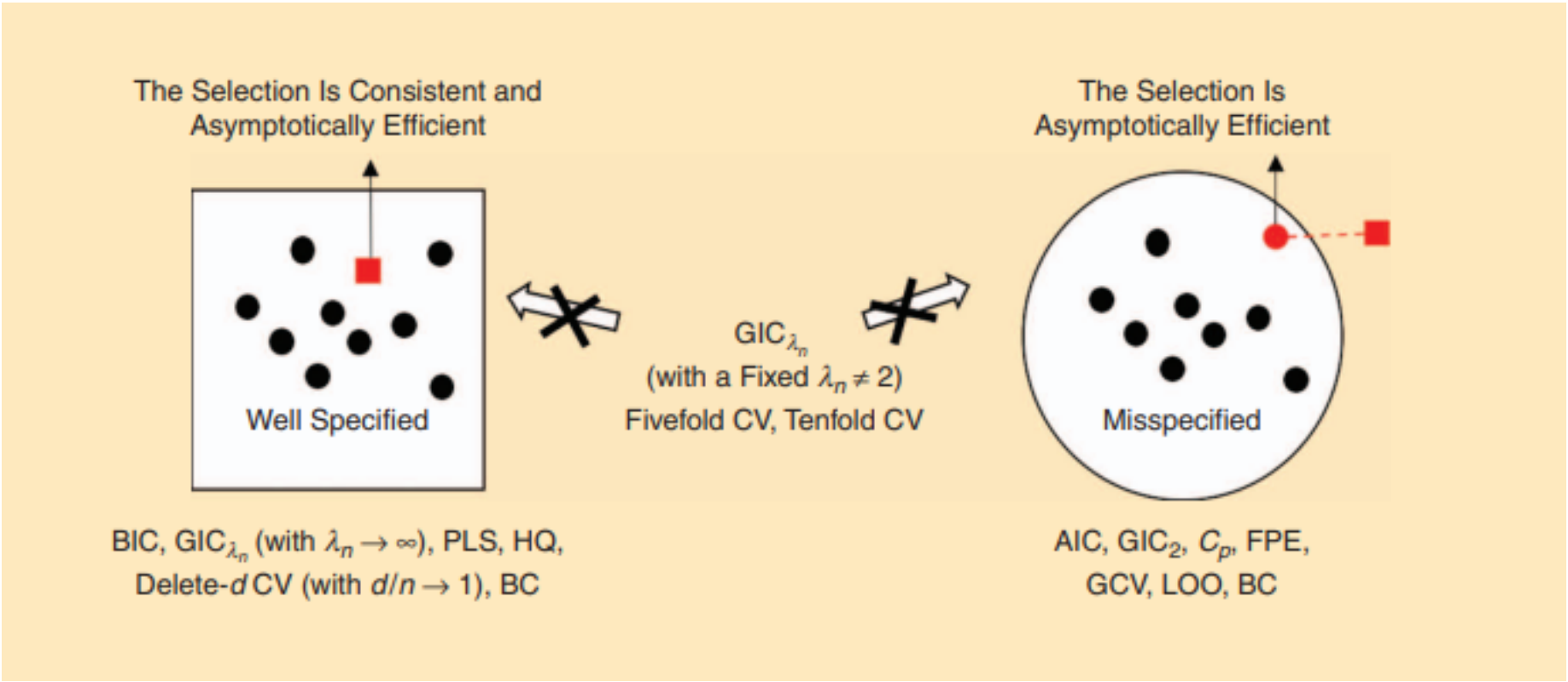


Figure: A graph illustrating a parametric setting and a nonparametric setting of consisitency and asympotically efficiency.

References I