Statistics
MDI220
2. Quadratic Risk

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

Let $\hat{\theta}$ be the estimator of some parameter θ. We introduce the quadratic risk of the estimator, and the notion of efficient estimation based on the Cramér-Rao bound.

1 Definition

Assume that $\theta \in \mathbb{R}$. The quadratic risk is the mean square error of the estimation.

The quadratic risk of the estimator $\hat{\theta}$ is $R(\theta, \hat{\theta}) = \mathrm{E}((\hat{\theta}(X) - \theta)^2)$.

Example. *The quadratic risk of the estimator* $\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^n x_i$ *for the Bernoulli model is given by:*

$$\begin{aligned}
R(\hat{\theta}, \theta) &= \mathrm{E}\left((\frac{1}{n}\sum_{i=1}^n X_i - \theta)^2\right),\\
&= \frac{1}{n^2}\mathrm{E}\left((\sum_{i=1}^n (X_i - \theta))^2\right),\\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{E}(X_i^2) - \frac{2\theta}{n^2}\sum_{i=1}^n \mathrm{E}(X_i) + \frac{\theta^2}{n},\\
&= \frac{\theta(1-\theta)}{n}.
\end{aligned}$$

2 Bias-variance decomposition

The quadratic risk can be easily derived using the bias-variance decomposition.

The quadratic risk of the estimator $\hat{\theta}$ is given by $R(\theta, \hat{\theta}) = b(\theta, \hat{\theta})^2 + \mathrm{var}(\hat{\theta}(X))$.

Example. *Consider the estimator* $\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^n x_i$ *for the Bernoulli model. Since it is unbiased, we get*

$$R(\theta, \hat{\theta}) = \mathrm{var}(\hat{\theta}(X)) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}(X_i) = \frac{\theta(1-\theta)}{n}.$$

The bias-variance decomposition is a consequence of the following equality:

$$\begin{aligned}(\hat\theta-\theta)^2 &= (\hat\theta - \mathrm{E}(\hat\theta) + \mathrm{E}(\hat\theta) - \theta)^2,\\ &= (\hat\theta - \mathrm{E}(\hat\theta))^2 + 2(\hat\theta - \mathrm{E}(\hat\theta))(\mathrm{E}(\hat\theta)-\theta) + (\mathrm{E}(\hat\theta)-\theta)^2.\end{aligned}$$

using the notation $\hat\theta \equiv \hat\theta(X)$ for convenience. Taking the expectation, we get:

$$R(\theta,\hat\theta) = \underbrace{\mathrm{E}((\hat\theta - \mathrm{E}(\hat\theta))^2}_{\text{variance}} + \underbrace{(\mathrm{E}(\hat\theta)-\theta)^2}_{\text{square bias}}.$$

3 Fisher information

The estimation can be more or less difficult depending on the statistical model. The Fisher information is a measure of the quantity of information carried by the observation on the unknown parameter θ; the higher the information, the easier the estimation. It is worth noting that the Fisher information depends on the statistical model only, not on the estimator.

We assume that the model is *regular*[1]; in particular, it is dominated. The Fisher information is the opposite of the expected *curvature* of the log-likelihood:

The Fisher information is defined by $I(\theta) = -\mathrm{E}\left(\frac{\partial^2 \log p_\theta}{\partial\theta^2}(X)\right)$.

The derivative of the log-likelihood in the parameter θ is called the *score*. It is centered:

$$\mathrm{E}\left(\frac{\partial \log p_\theta}{\partial\theta}(X)\right) = 0.$$

We have the following equivalent definition of the Fisher information:

The Fisher information is the variance of the score, $I(\theta) = \mathrm{var}\left(\frac{\partial \log p_\theta}{\partial\theta}(X)\right)$.

Example. *For the Bernoulli model with n obervations, the log-likelihood is:*

$$\log p_\theta(x) = s\log\theta + (n-s)\log(1-\theta),$$

with $s = \sum_{i=1}^n x_i$. We deduce the score:

$$\frac{\partial \log p_\theta}{\partial\theta}(X) = \frac{S}{\theta} - \frac{n-S}{1-\theta},$$

with $S = \sum_{i=1}^n X_i$, and the Fisher information,

$$I(\theta) = \left(\frac{1}{\theta(1-\theta)}\right)^2 \mathrm{var}(S) = \frac{n}{\theta(1-\theta)}$$

Let $I_n(\theta)$ be the Fisher information for n observations. Then $I_n(\theta) = nI_1(\theta)$.

[1] This includes the models of the exponential family; see the text book for the exact technical conditions.

4 Cramér-Rao bound

The Cramér-Rao bound is a lower bound on the quadratic risk. We assume that the estimator $\hat{\theta}$ is unbiased and has finite variance.

The quadratic risk of the estimator $\hat{\theta}$ satisfies $R(\theta, \hat{\theta}) \geq \frac{1}{I(\theta)}$.

Any estimator that is unbiased and reaches the Cramér-Rao bound is said to be efficient.

Example. *The estimator $\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ for the Bernoulli model is efficient since:*

$$R(\theta, \hat{\theta}) = \frac{\theta(1-\theta)}{n} = \frac{1}{I(\theta)}.$$

To prove the Cramér-Rao bound, we use the fact that the estimator is unbiased:

$$\theta = \mathrm{E}(\hat{\theta}(X)).$$

By derivation in θ we get:

$$\begin{aligned}
1 &= \frac{\partial}{\partial \theta} \int \hat{\theta}(x) p_\theta(x) \mathrm{d}\nu(x), \\
&= \int \hat{\theta}(x) \frac{\partial p_\theta}{\partial \theta}(x) \mathrm{d}\nu(x), \\
&= \int \hat{\theta}(x) \frac{\partial \log p_\theta}{\partial \theta}(x) p_\theta(x) \mathrm{d}\nu(x), \\
&= \mathrm{E}\left(\hat{\theta}(X) \frac{\partial \log p_\theta}{\partial \theta}(X)\right).
\end{aligned}$$

Using again the fact that the estimator is unbiased, we can write this equality:

$$\mathrm{cov}\left(\hat{\theta}(X), \frac{\partial \log p_\theta}{\partial \theta}(X)\right) = 1.$$

We conclude by Cauchy-Schwartz inequality:

$$\underbrace{\mathrm{var}(\hat{\theta}(X))}_{R(\theta,\hat{\theta})} \underbrace{\mathrm{var}\left(\frac{\partial \log p_\theta}{\partial \theta}(X)\right)}_{I(\theta)} \geq 1.$$

5 General case

Now consider the estimation of some function g of the parameter θ. Assume g is differentiable. Denote by $\hat{g}$ the new estimator. We define the bias as:

$$b(\theta, \hat{g}) = \mathrm{E}(\hat{g}(X)) - g(\theta).$$

The quadratic risk of the estimator $\hat{g}$ is $R(\theta, \hat{g}) = \mathrm{E}((\hat{g}(X) - g(\theta))^2)$.

We have the bias-variance decomposition:

The quadratic risk of the estimator $\hat{g}$ is given by $R(\theta, \hat{g}) = b(\theta, \hat{g})^2 + \mathrm{var}(\hat{g}(X))$.

Example. *Consider the exponential model $\mathcal{P} = \{P_\theta \sim \mathcal{E}(\theta), \theta > 0\}$ with n observations. Assume you want to estimate the mean $g(\theta) = \frac{1}{\theta}$. The estimator $\hat{g}(x) = \frac{1}{n}\sum_{i=1}^n x_i$ is unbiased:*

$$\mathrm{E}(\hat{g}(X)) = \frac{1}{\theta} = g(\theta).$$

Its quadratic risk is:

$$R(\theta, \hat{g}) = \mathrm{var}(\hat{g}(X)) = \frac{1}{n\theta^2}.$$

The Cramér-Rao bound applies to any unbiased estimator $\hat{g}$:

The quadratic risk of the estimator $\hat{g}$ satisfies $R(\theta, \hat{g}) \geq \frac{g'(\theta)^2}{I(\theta)}$.

Example. *Consider the exponential model. For $n = 1$ observation, the probability density function is $p_\theta(x) = \theta e^{-\theta x}$. We deduce the score:*

$$\frac{\partial \log p_\theta}{\partial \theta}(x) = \frac{1}{\theta} - x,$$

and the Fisher information:

$$I_1(\theta) = \frac{1}{\theta^2}.$$

For n observations, the Fisher information is

$$I(\theta) = nI_1(\theta) = \frac{n}{\theta^2}.$$

The Cramér-Rao bound is:

$$R(\theta, \hat{g}) \geq \frac{1}{n\theta^2}.$$

The estimator $\hat{g}(x) = \frac{1}{n}\sum_{i=1}^n x_i$ is efficient.

6 Vectorial case

Finally, we consider the vectorial case $\theta \in \Theta \subset \mathbb{R}^k$. Assume you want to estimate $g(\theta)$, for some differentiable function $g : \Theta \to \mathbb{R}$. The Fisher information is the opposite of the Hessian matrix of the log-likelihood:

The Fisher information is defined by $I(\theta) = -\mathrm{E}\left(\frac{\partial^2 \log p_\theta}{\partial \theta_i \partial \theta_j}(X)\right)_{i,j=1,\dots,k}$.

We have the following equivalent definition:

The Fisher information is the covariance matrix of the score: $I(\theta) = \mathrm{cov}\left(\nabla \log p_\theta(X)\right)$.

Assuming that this matrix is invertible, we get the Cramér-Rao bound for any unbiased estimator $\hat{g}$:

The quadratic risk of the estimator $\hat{g}$ satisfies $R(\theta, \hat{g}) \geq \nabla g(\theta)^T I(\theta)^{-1} \nabla g(\theta)$.