Analyzing Information Metrics and Properties of Maximum Likelihood Estimators

Søren Lund Pedersen

In this article, we will discuss observed information, expected (Fisher) information, properties of maximum likelihood estimators, Kullback-Leibler (KL) divergence, and Akaike Information Criterion (AIC). Additionally, we will delve into moment generating functions (MGFs), cumulant generating functions (CGFs), and the Delta Method to provide a thorough understanding of these statistical tools.

1 Observed Information

The observed information pertains to the curvature of the log-likelihood function evaluated at a specific data set. It is derived from the second derivative (the Hessian in higher dimensions) of the log-likelihood function with respect to the parameter θ. The intuition behind observed information is that the curvature indicates how sensitive the likelihood is around the estimated parameter value:

- A sharp peak (high curvature) suggests that even small deviations from the estimated parameter result in a significant decrease in the likelihood. This implies that the data provide a precise estimate of θ.
- A flat peak (low curvature) indicates that the likelihood does not change much with variations in θ, signifying less information in the data about the parameter.

Formally, the observed information $J_{\text{obs}}(\theta)$ is expressed as:

$$J_{\text{obs}}(\theta) = -\frac{\partial^2 \ell(\theta; x)}{\partial \theta^2},$$

where $\ell(\theta; x)$ is the log-likelihood function given the data x.

2 Fisher Information

The expected information, known as the Fisher information $I(\theta)$, is the expectation of the observed information under the assumed model. It provides a measure of the average amount of information that an observable random variable X carries about the parameter θ. Fisher information considers the variability of the observed information over different possible samples.

Formally, Fisher information is defined as:

$$I(\theta) = \mathbb{E}_\theta \left[J_{\text{obs}}(\theta) \right] = \mathbb{E}_\theta \left[-\frac{\partial^2 \ell(\theta; X)}{\partial \theta^2} \right],$$

and it can also be expressed using the variance of the score function:

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \ell(\theta; X)}{\partial \theta} \right)^2 \right].$$

Fisher information has several important implications:

- It quantifies the **"informativeness"** of the data regarding θ. Higher Fisher information means more precise estimates are possible.
- It underpins the Cramér-Rao Bound (CRB), which states that the variance of any unbiased estimator $\hat{\theta}$ of θ is bounded below by the reciprocal of the Fisher information:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

- The standard error (SE) of the estimator $\hat{\theta}$ can be approximated by:

$$SE(\hat{\theta}) \approx \frac{1}{\sqrt{I(\theta)}}.$$

- Tight confidence intervals around θ suggest high Fisher information, indicating that the data provide a lot of information about the parameter.

3 Properties of Maximum Likelihood Estimators

Maximum Likelihood Estimators (MLEs) possess several desirable properties under regularity conditions. Here, we focus on relative efficiency and consistency.

3.1 Relative Efficiency

Relative efficiency compares the performance of two estimators in terms of their variances. Specifically, it measures how the variance of a given estimator $\hat{\theta}_1$ compares to that of the MLE $\hat{\theta}_{\text{MLE}}$, which is considered the benchmark for efficiency under the CRB.

The relative efficiency (RE) is defined as:

$$RE = \frac{\text{Var}(\hat{\theta}_{\text{MLE}})}{\text{Var}(\hat{\theta}_1)}.$$

- If $RE > 1$, the MLE has a smaller variance than $\hat{\theta}_1$, making it more efficient.
- If $RE < 1$, the alternative estimator $\hat{\theta}_1$ is more efficient than the MLE for the given sample.

3.2 Consistency

An estimator $\hat{\theta}_n$ is said to be consistent if it converges in probability to the true parameter θ as the sample size n tends to infinity. Consistency ensures that with more data, the estimator becomes arbitrarily close to the true value.

There are three primary modes of convergence for sequences of random variables $\{X_n\}$:

- **Convergence in Probability**:

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} P(|X_n - X| > \epsilon) = 0.$$

  This mode of convergence implies that for large n, X_n is very likely to be close to X.

- **Convergence in Distribution**:

$$X_n \xrightarrow{d} X \implies F_{X_n}(x) \to F_X(x) \quad \text{for all } x \text{ at which } F_X \text{ is continuous.}$$

  Convergence in distribution, often used in the context of the Central Limit Theorem, is the weakest form of convergence, ensuring that the distribution of X_n approximates that of X as n grows.

- **Almost Sure Convergence**:

$$P\Big(\lim_{n \to \infty} X_n = X\Big) = 1.$$

  This is the strongest form of convergence, indicating that X_n will converge to X with probability 1 as n approaches infinity.

The relationships among these modes of convergence are as follows:

$$\text{Almost sure convergence} \implies \text{Convergence in probability} \implies \text{Convergence in distribution.}$$

However, the converses of these implications do not generally hold.

For MLEs, under regularity conditions:

- MLEs are **consistent**, meaning $\hat{\theta}_n$ converges in probability (and sometimes almost surely) to θ as $n \to \infty$.
- They are **asymptotically normal**, meaning the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ approaches a normal distribution with mean 0 and variance $1/I(\theta)$.

4 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions. Given a true distribution P and an approximating distribution Q, the KL divergence from Q to P is defined as:

$$D_{KL}(P \,\|\, Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx.$$

Key points about KL divergence:

- It is not symmetric, meaning $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$ in general.
- It measures the expected log difference between the true distribution P and the approximation Q, weighted by P.
- A smaller KL divergence indicates that Q is a better approximation of P.

In practical applications, KL divergence is used for:

- **Model selection:** Comparing how well different models approximate the true data-generating process.
- **Machine learning:** Techniques such as knowledge distillation and training of variational autoencoders (VAEs) minimize KL divergence.
- **Information theory:** Quantifying information loss when approximating one distribution by another.

5 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a metric for model selection based on the concept of information loss. It balances the model fit and complexity to prevent overfitting. The AIC is calculated as:

$$AIC = 2k - 2\ell(\hat{\theta}; x),$$

where:

- k is the number of parameters in the model.
- $\ell(\hat{\theta}; x)$ is the log-likelihood of the data given the model evaluated at the MLE $\hat{\theta}$.

The AIC can also be derived from the perspective of KL divergence:

$$AIC = 2k + n \cdot D_{KL},$$

where n is the sample size. The model with the lowest AIC is usually preferred, as it is expected to have the smallest information loss relative to the true model.

6 Moment Generating Functions (MGFs)

The Moment Generating Function (MGF) of a random variable X provides a compact way to encode all of its moments (if the MGF exists). It is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

MGFs have several important properties:

- The nth moment of X can be obtained by differentiating $M_X(t)$ n times and evaluating at $t = 0$:
$$\mathbb{E}[X^n] = M_X^{(n)}(0).$$

- If two random variables have the same MGF in a neighborhood of $t = 0$, they have the same distribution.
- MGFs simplify the process of finding the distribution of sums of independent random variables, as the MGF of the sum equals the product of their MGFs.

7 Cumulant Generating Functions (CGFs)

The Cumulant Generating Function (CGF) is the logarithm of the MGF:

$$K_X(t) = \log M_X(t).$$

The CGF generates cumulants, which are analogous to moments but with properties that make them additive for independent random variables. The first few cumulants are:

- First cumulant $\kappa_1 = \mathbb{E}[X]$ (mean).
- Second cumulant $\kappa_2 = \text{Var}(X)$ (variance).
- Third cumulant related to skewness.
- Fourth cumulant related to kurtosis.

Cumulants are particularly useful because the cumulant of a sum of independent random variables is the sum of their cumulants.

8 Delta Method

The Delta Method is a technique used in asymptotic statistics to approximate the distribution of a function of an estimator. Suppose $\hat{\theta}$ is an estimator for θ such that:

$$\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2) \quad \text{for large } n.$$

For a differentiable function $g(\cdot)$, the Delta Method states that:

$$g(\hat{\theta}) \approx \mathcal{N}\Big(g(\theta), \big[g'(\theta)\big]^2 \sigma^2\Big).$$

This result is obtained using a first-order Taylor expansion of $g(\hat{\theta})$ around θ:

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta).$$

Because $\hat{\theta} - \theta$ is approximately normal with mean 0 and variance σ^2, the linear approximation implies $g(\hat{\theta})$ is approximately normal with the stated mean and variance.

Example: If $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$, to approximate the variance of $\exp(\hat{\theta})$ using the Delta Method:

$$\text{Var}(\exp(\hat{\theta})) \approx \Big(\exp(\theta)\Big)^2 \sigma^2.$$

This approximation is particularly useful when dealing with nonlinear transformations of estimators.