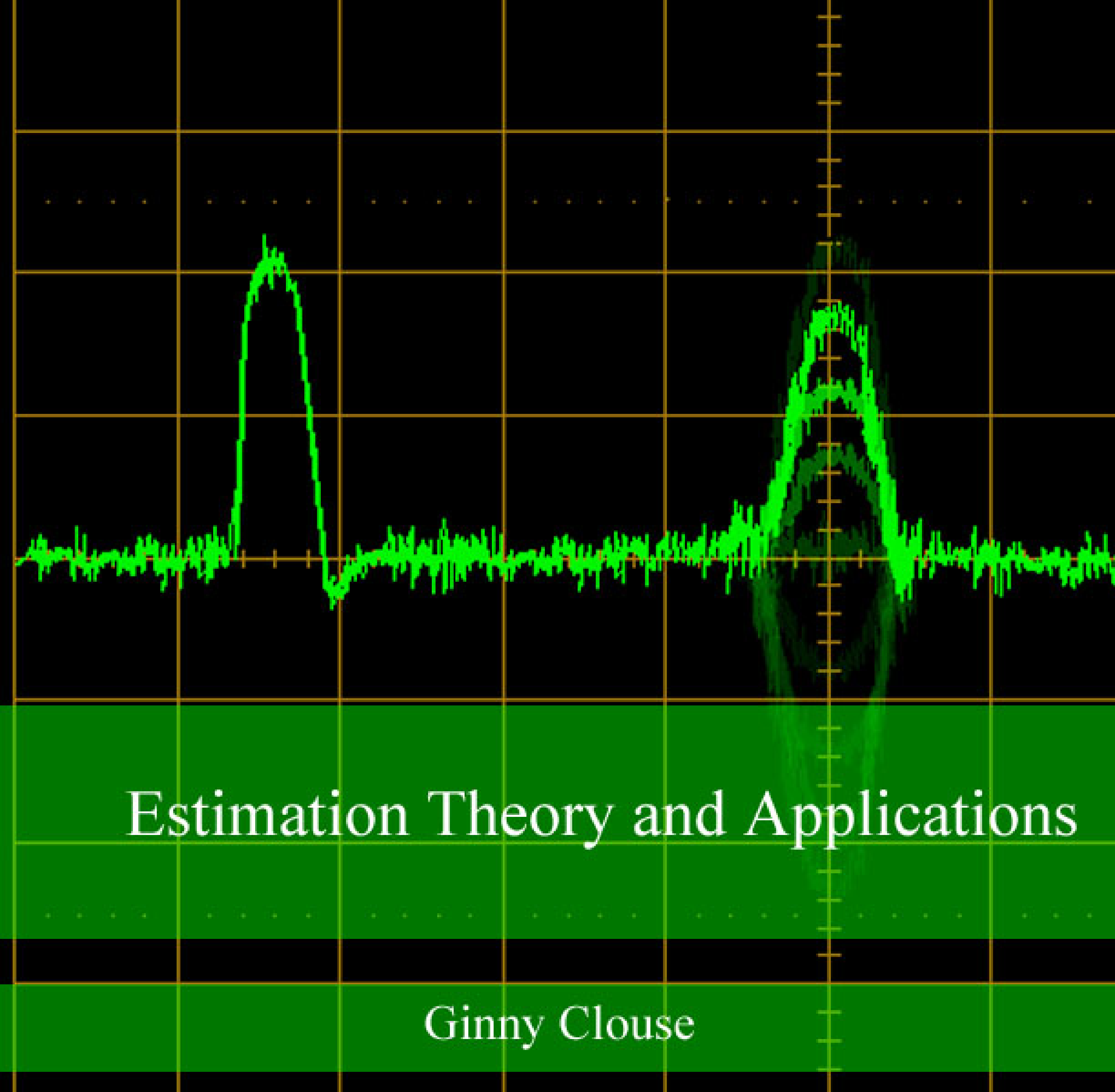


Estimation Theory and Applications

Ginny Clouse

First Edition, 2012

ISBN 978-81-323-3050-9

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Estimation Theory

Estimation theory is a branch of statistics and signal processing that deals with estimating the values of parameters based on measured/empirical data that has a random component. The parameters describe an underlying physical setting in such a way that the value of the parameters affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements.

For example, it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the unobservable parameter; the estimate is based on a small random sample of voters.

Or, for example, in radar the goal is to estimate the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses. Since the reflected pulses are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated.

In estimation theory, it is assumed the measured data is random with probability distribution dependent on the parameters of interest. For example, in electrical communication theory, the measurements which contain information regarding the parameters of interest are often associated with a noisy signal. Without randomness, or noise, the problem would be deterministic and estimation would not be needed.

Estimation process

The entire purpose of estimation theory is to arrive at an estimator, and preferably an implementable one that could actually be used. The estimator takes the measured data as input and produces an estimate of the parameters.

It is also preferable to derive an estimator that exhibits optimality. Estimator optimality usually refers to achieving minimum average error over some class of estimators, for example, a minimum variance unbiased estimator. In this case, the class is the set of unbiased estimators, and the average error measure is variance (average squared error between the value of the estimate and the parameter). However, optimal estimators do not always exist.

These are the general steps to arrive at an estimator:

- In order to arrive at a desired estimator, it is first necessary to determine a probability distribution for the measured data, and the distribution's dependence on the unknown parameters of interest. Often, the probability distribution may be derived from physical models that explicitly show how the measured data depends on the parameters to be estimated, and how the data is corrupted by random errors or noise. In other cases, the probability distribution for the measured data is simply "assumed", for example, based on familiarity with the measured data and/or for analytical convenience.
- After deciding upon a probabilistic model, it is helpful to find the limitations placed upon an estimator. This limitation, for example, can be found through the Cramér–Rao bound.
- Next, an estimator needs to be developed or applied if an already known estimator is valid for the model. The estimator needs to be tested against the limitations to determine if it is an optimal estimator (if so, then no other estimator will perform better).
- Finally, experiments or simulations can be run using the estimator to test its performance.

After arriving at an estimator, real data might show that the model used to derive the estimator is incorrect, which may require repeating these steps to find a new estimator. A non-implementable or infeasible estimator may need to be scrapped and the process started anew.

In summary, the estimator estimates the parameters of a physical model based on measured data.

Teori Estimasi

Basics

To build a model, several statistical "ingredients" need to be known. These are needed to ensure the estimator has some mathematical tractability instead of being based on "good feel".

The first is a set of statistical samples taken from a random vector (RV) of size *N*. Put into a vector,

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}.$$

Secondly, we have the corresponding *M* parameters

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix},$$

which need to be established with their probability density function (pdf) or probability mass function (pmf)

$$p(\mathbf{x}|\theta).$$

It is also possible for the parameters themselves to have a probability distribution (e.g., Bayesian statistics). It is then necessary to define the Bayesian probability

$$\pi(\theta).$$

After the model is formed, the goal is to estimate the parameters, commonly denoted $\hat{\theta}$, where the "hat" indicates the estimate.

One common estimator is the minimum mean squared error (MMSE) estimator, which utilizes the error between the estimated parameters and the actual value of the parameters

$$\mathbf{e} = \hat{\theta} - \theta$$

as the basis for optimality. This error term is then squared and minimized for the MMSE estimator.

Estimators

Commonly-used estimators and estimation methods, and topics related to them:

- Maximum likelihood estimators
- Bayes estimators
- Method of moments estimators
- Cramér–Rao bound

- Minimum mean squared error (MMSE), also known as Bayes least squared error (BLSE)
- Maximum a posteriori (MAP)
- Minimum variance unbiased estimator (MVUE)
- Best linear unbiased estimator (BLUE)
- Unbiased estimators.
- Particle filter
- Markov chain Monte Carlo (MCMC)
- Kalman filter
- Ensemble Kalman filter (EnKF)
- Wiener filter

Examples

Unknown constant in additive white Gaussian noise

Consider a received discrete signal, *x*[*n*], of *N* independent samples that consists of an unknown constant *A* with additive white Gaussian noise (AWGN) *w*[*n*] with known variance σ^2 (*i.e.*, $\mathcal{N}(0, \sigma^2)$). Since the variance is known then the only unknown parameter is *A*.

The model for the signal is then

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1$$

Two possible (of many) estimators are:

- $\hat{A}_1 = x[0]$
- $\hat{A}_2 = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$ which is the sample mean

Both of these estimators have a mean of *A*, which can be shown through taking the expected value of each estimator

$$\mathrm{E}\left[\hat{A}_1\right] = \mathrm{E}\left[x[0]\right] = A$$

and

$$\mathrm{E}\left[\hat{A}_2\right] = \mathrm{E}\left[\frac{1}{N} \sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N}\left[\sum_{n=0}^{N-1} \mathrm{E}\left[x[n]\right]\right] = \frac{1}{N}\left[NA\right] = A$$

At this point, these two estimators would appear to perform the same. However, the difference between them becomes apparent when comparing the variances.

$$\mathrm{var}\left(\hat{A}_1\right) = \mathrm{var}\left(x[0]\right) = \sigma^2$$

and

$$\mathrm{var}\left(\hat{A}_2\right) = \mathrm{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) \overset{independence}{=} \frac{1}{N^2}\left[\sum_{n=0}^{N-1} \mathrm{var}(x[n])\right] = \frac{1}{N^2}\left[N\sigma^2\right] = \frac{\sigma^2}{N}$$

It would seem that the sample mean is a better estimator since, as $N \to \infty$, the variance goes to zero.

Maximum likelihood

Continuing the example using the maximum likelihood estimator, the probability density function (pdf) of the noise for one sample *w*[*n*] is

$$p(w[n]) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} w[n]^2\right)$$

and the probability of *x*[*n*] becomes (*x*[*n*] can be thought of a $\mathcal{N}(A, \sigma^2)$)

$$p(x[n]; A) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x[n] - A)^2\right)$$

By independence, the probability of $\mathbf{x}$becomes

$$p(\mathbf{x}; A) = \prod_{n=0}^{N-1} p(x[n]; A) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right)$$

Taking the natural logarithm of the pdf

$$\ln p(\mathbf{x}; A) = -N\ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2$$

and the maximum likelihood estimator is

$$\hat{A} = \arg\max \ln p(\mathbf{x}; A)$$

Taking the first derivative of the log-likelihood function

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A)=\frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}(x[n]-A)\right]=\frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right]$$

and setting it to zero

$$0=\frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right]=\sum_{n=0}^{N-1}x[n]-NA$$

This results in the maximum likelihood estimator

$$\hat{A}=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

which is simply the sample mean. From this example, it was found that the sample mean is the maximum likelihood estimator for *N* samples of a fixed, unknown parameter corrupted by AWGN.

Cramér–Rao lower bound

To find the Cramér–Rao lower bound (CRLB) of the sample mean estimator, it is first necessary to find the Fisher information number

$$\mathcal{I}(A)=\mathrm{E}\left(\left[\frac{\partial}{\partial\theta}\ln p(\mathbf{x};A)\right]^2\right)=-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{x};A)\right]$$

and copying from above

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A)=\frac{1}{\sigma^2}\left[\sum_{n=0}^{N-1}x[n]-NA\right]$$

Taking the second derivative

$$\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x};A)=\frac{1}{\sigma^2}(-N)=\frac{-N}{\sigma^2}$$

and finding the negative expected value is trivial since it is now a deterministic constant

$$-\mathbf{E}\left[\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x};A)\right] = \frac{N}{\sigma^2}$$

Finally, putting the Fisher information into

$$\mathrm{var}\left(\hat{A}\right) \geq \frac{1}{\mathcal{I}}$$

results in

$$\mathrm{var}\left(\hat{A}\right) \geq \frac{\sigma^2}{N}$$

Comparing this to the variance of the sample mean (determined previously) shows that the sample mean is *equal to* the Cramér–Rao lower bound for all values of *N* and *A*. In other words, the sample mean is the (necessarily unique) efficient estimator, and thus also the minimum variance unbiased estimator (MVUE), in addition to being the maximum likelihood estimator.

Maximum of a uniform distribution

One of the simplest non-trivial examples of estimation is the estimation of the maximum of a uniform distribution. It is used as a hands-on classroom exercise and to illustrate basic principles of estimation theory. Further, in the case of estimation based on a single sample, it demonstrates philosophical issues and possible misunderstandings in the use of maximum likelihood estimators and likelihood functions.

Given a discrete uniform distribution $1, 2, \dots, N$ with unknown maximum, the UMVU estimator for the maximum is given by

$$\frac{k+1}{k}m - 1 = m + \frac{m}{k} - 1$$

where *m* is the sample maximum and *k* is the sample size, sampling without replacement. This problem is commonly known as the German tank problem, due to application of maximum estimation to estimates of German tank production during World War II.

The formula may be understood intuitively as:

> "The sample maximum plus the average gap between observations in the sample",

the gap being added to compensate for the negative bias of the sample maximum as an estimator for the population maximum.

This has a variance of

$$\frac{1}{k}\frac{(N-k)(N+1)}{(k+2)} \approx \frac{N^2}{k^2} \text{ for small samples } k \ll N$$

so a standard deviation of approximately N / k, the (population) average size of a gap between samples; compare $\frac{m}{k}$ above. This can be seen as a very simple case of maximum spacing estimation.

The sample maximum is the maximum likelihood estimator for the population maximum, but, as discussed above, it is biased.

Applications

Numerous fields require the use of estimation theory. Some of these fields include (but are by no means limited to):

- Interpretation of scientific experiments
- Signal processing
- Clinical trials
- Opinion polls
- Quality control
- Telecommunications
- Project management
- Software engineering
- Control theory
- Network intrusion detection system
- Orbit determination

Measured data are likely to be subject to noise or uncertainty and it is through statistical probability that optimal solutions are sought to extract as much information from the data as possible.

Chapter 2

Bayes Estimator

In estimation theory and decision theory, a **Bayes estimator** or a **Bayes action** is an estimator or decision rule that minimizes the posterior expected value of a loss function (i.e., the **posterior expected loss**). Equivalently, it maximizes the posterior expectation of a utility function. An alternative way of formulating an estimator within Bayesian statistics is Maximum a posteriori estimation.

Definition

Suppose an unknown parameter θ is known to have a prior distribution π. Let δ = δ(*x*) be an estimator of θ (based on some measurements *x*), and let *L*(θ,δ) be a loss function, such as squared error. The **Bayes risk** of δ is defined as $E_\pi\{L(\theta,\delta)\}$, where the expectation is taken over the probability distribution of θ: this defines the risk function as a function of δ. An estimator δ is said to be a *Bayes estimator* if it minimizes the Bayes risk among all estimators. Equivalently, the estimator which minimizes the posterior expected loss *E*{*L*(θ,δ) | *x*} *for each x* also minimizes the Bayes risk and therefore is a Bayes estimator.

If the prior is improper then an estimator which minimizes the posterior expected loss *for each x* is called a **generalized Bayes estimator**.

Examples

Minimum mean square error estimation

The most common risk function used for Bayesian estimation is the mean square error (MSE), also called squared error risk. The MSE is defined by

$$\mathrm{MSE} = E\left[(\widehat{\theta}(x) - \theta)^2\right],$$

where the expectation is taken over the joint distribution of θ and *x*.

Using the MSE as risk, the Bayes estimate of the unknown parameter is simply the mean of the posterior distribution,

$$\widehat{\theta}(x) = E[\theta|x] = \int \theta f(\theta|x)\, d\theta.$$

This is known as the *minimum mean square error* (MMSE) estimator. The Bayes risk, in this case, is the posterior variance.

Bayes estimators for conjugate priors

If there is no inherent reason to prefer one prior probability distribution over another, a conjugate prior is sometimes chosen for simplicity. A conjugate prior is defined as a prior distribution belonging to some parametric family, for which the resulting posterior distribution also belongs to the same family. This is an important property, since the Bayes estimator, as well as its statistical properties (variance, confidence interval, etc.), can all be derived from the posterior distribution.

Conjugate priors are especially useful for sequential estimation, where the posterior of the current measurement is used as the prior in the next measurement. In sequential estimation, unless a conjugate prior is used, the posterior distribution typically becomes more complex with each added measurement, and the Bayes estimator cannot usually be calculated without resorting to numerical methods.

Following are some examples of conjugate priors.

- If x|θ is normal, $x|\theta \sim N(\theta,\sigma^2)$, and the prior is normal, $\theta \sim N(\mu,\tau^2)$, then the posterior is also normal and the Bayes estimator under MSE is given by

$$\widehat{\theta}(x) = \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}x.$$

- If $x_1,...,x_n$ are iid Poisson random variables $x_i|\theta \sim P(\theta)$, and if the prior is Gamma distributed $\theta \sim G(a,b)$, then the posterior is also Gamma distributed, and the Bayes estimator under MSE is given by

$$\widehat{\theta}(X) = \frac{n\overline{X}+a}{n+\frac{1}{b}}.$$

- If $x_1,...,x_n$ are iid uniformly distributed $x_i|\theta{\sim}U(0,\theta)$, and if the prior is Pareto distributed $\theta{\sim}Pa(\theta_0,a)$, then the posterior is also Pareto distributed, and the Bayes estimator under MSE is given by

$$\widehat{\theta}(X) = \frac{(a+n)\max(\theta_0, x_1, ..., x_n)}{a+n-1}.$$

Alternative risk functions

Risk functions are chosen depending on how one measures the distance between the estimate and the unknown parameter. The MSE is the most common risk function in use, primarily due to its simplicity. However, alternative risk functions are also occasionally used. The following are several examples of such alternatives. We denote the posterior generalized distribution function by *F*.

- A "linear" loss function, with *a* > 0, which yields the posterior median as the Bayes' estimate:

$$L(\theta,\widehat{\theta}) = a|\theta-\widehat{\theta}|$$
$$F(\widehat{\theta}(x)|X) = \tfrac{1}{2}.$$

- Another "linear" loss function, which assigns different "weights" *a*,*b* > 0 to over or sub estimation. It yields a quantile from the posterior distribution, and is a generalization of the previous loss function:

$$L(\theta,\widehat{\theta}) = \begin{cases} a|\theta-\widehat{\theta}|, & \text{for } \theta-\widehat{\theta}\geq 0 \\ b|\theta-\widehat{\theta}|, & \text{for } \theta-\widehat{\theta}< 0 \end{cases}$$
$$F(\widehat{\theta}(x)|X) = \frac{a}{a+b}.$$

- The following loss function is trickier: it yields either the posterior mode, or a point close to it depending on the curvature and properties of the posterior distribution. Small values of the parameter *K* > 0 are recommended, in order to use the mode as an approximation (*L* > 0):

$$L(\theta,\widehat{\theta}) = \begin{cases} 0, & \text{for } |\theta-\widehat{\theta}|< K \\ L, & \text{for } |\theta-\widehat{\theta}|\geq K. \end{cases}$$

Other loss functions can be conceived, although the mean squared error is the most widely used and validated.

Generalized Bayes estimators

The prior distribution π has thus far been assumed to be a true probability distribution, in that

$$\int \pi(\theta) d\theta = 1.$$

However, occasionally this can be a restrictive requirement. For example, there is no distribution (covering the set, **R**, of all real numbers) for which every real number is equally likely. Yet, in some sense, such a "distribution" seems like a natural choice for a non-informative prior, i.e., a prior distribution which does not imply a preference for any particular value of the unknown parameter. One can still define a function π(θ) = 1, but this would not be a proper probability distribution since it has infinite mass,

$$\int \pi(\theta) d\theta = \infty.$$

Such measures π(θ), which are not probability distributions, are referred to as improper priors.

The use of an improper prior means that the Bayes risk is undefined (since the prior is not a probability distribution and we cannot take an expectation under it). As a consequence, it is no longer meaningful to speak of a Bayes estimator that minimizes the Bayes risk. Nevertheless, in many cases, one can define the posterior distribution

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta) d\theta}.$$

This is a definition, and not an application of Bayes' theorem, since Bayes' theorem can only be applied when all distributions are proper. However, it is not uncommon for the resulting "posterior" to be a valid probability distribution. In this case, the posterior expected loss

$$\int L(\theta, a)\pi(\theta|x) d\theta$$

is typically well-defined and finite. Recall that, for a proper prior, the Bayes estimator minimizes the posterior expected loss. When the prior is improper, an estimator which minimizes the posterior expected loss is referred to as a **generalized Bayes estimator**.

Example

A typical example concerns the estimation of a location parameter with a loss function of the type *L*(*a* − θ). Here θ is a location parameter, i.e., *p*(*x* | θ) = *f*(*x* − θ).

It is common to use the improper prior π(θ) = 1 in this case, especially when no other more subjective information is available. This yields

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} = \frac{f(x-\theta)}{p(x)}$$

so the posterior expected loss equals

$$E[L(a-\theta)] = \int L(a-\theta)\pi(\theta|x)d\theta = \frac{1}{p(x)}\int L(a-\theta)f(x-\theta)d\theta.$$

The generalized Bayes estimator is the value $a(x)$ which minimizes this expression for all x. This is equivalent to minimizing

$$\int L(a-\theta)f(x-\theta)d\theta \text{ for all } x. \qquad (1)$$

It can be shown that, in this case, the generalized Bayes estimator has the form $x + a_0$, for some constant a_0. To see this, let a_0 be the value minimizing (1) when $x = 0$. Then, given a different value x_1, we must minimize

$$\int L(a-\theta)f(x_1-\theta)d\theta = \int L(a-x_1-\theta')f(-\theta')d\theta'. \qquad (2)$$

This is identical to (1), except that a has been replaced by $a - x_1$. Thus, the expression minimizing is given by $a - x_1 = a_0$, so that the optimal estimator has the form

$$a(x) = a_0 + x.$$

Empirical Bayes estimators

A Bayes estimator derived through the empirical Bayes method is called an *empirical Bayes estimator*. Empirical Bayes methods enable the use of auxiliary empirical data, from observations of related parameters, in the development of a Bayes estimator. This is done under the assumption that the estimated parameters are obtained from a common prior. For example, if independent observations of different parameters are performed, then the estimation performance of a particular parameter can sometimes be improved by using data from other observations.

There are parametric and non-parametric approaches to empirical Bayes estimation. Parametric empirical Bayes is usually preferable since it is more applicable and more accurate on small amounts of data.

Example

The following is a simple example of parametric empirical Bayes estimation. Given past observations $x_1, \ldots, x_n$having conditional distribution $f(x_i \mid \theta_i)$, one is interested in estimating θ_{n+1} based on x_{n+1}. Assume that the θ_i's have a common prior π which depends on unknown parameters. For example, suppose that π is normal with unknown mean $\mu_\pi$and variance σ_π.We can then use the past observations to determine the mean and variance of π in the following way.

First, we estimate the mean $\mu_m$and variance $\sigma_m$of the marginal distribution of $x_1, \ldots, x_n$using the maximum likelihood approach:

$$\widehat{\mu}_m = \frac{1}{n}\sum x_i,$$
$$\widehat{\sigma}_m^2 = \frac{1}{n}\sum (x_i - \widehat{\mu}_m)^2.$$

Next, we use the relation

$$\mu_m = E_\pi[\mu_f(\theta)],$$
$$\sigma_m^2 = E_\pi[\sigma_f^2(\theta)] + E_\pi[\mu_f(\theta) - \mu_m],$$

where $\mu_f(\theta)$ and $\sigma_f(\theta)$ are the moments of the conditional distribution $f(x_i \mid \theta_i)$, which are assumed to be known. In particular, suppose that $\mu_f(\theta) = \theta$ and that $\sigma_f^2(\theta) = K$; we then have

$$\mu_\pi = \mu_m,$$
$$\sigma_\pi^2 = \sigma_m^2 - \sigma_f^2 = \sigma_m^2 - K.$$

Finally, we obtain the estimated moments of the prior,

$$\widehat{\mu}_\pi = \widehat{\mu}_m,$$
$$\widehat{\sigma}_\pi^2 = \widehat{\sigma}_m^2 - K.$$

For example, if $x_i \mid \theta_i$~$N(\theta_i,1)$, and if we assume a normal prior (which is a conjugate prior in this case), we conclude that $\theta_{n+1} \sim N(\widehat{\mu}_\pi, \widehat{\sigma}_\pi^2)$, from which the Bayes estimator of θ_{n+1} based on x_{n+1} can be calculated.

Properties

Admissibility

Bayes rules having finite Bayes risk are typically admissible. The following are some specific examples of admissibility theorems.

- If a Bayes rule is unique then it is admissible. For example, as stated above, under mean squared error (MSE) the Bayes rule is unique and therefore admissible.
- If θ belongs to a discrete set, then all Bayes rules are admissible.

- If θ belongs to a continuous (non-discrete set), and if the risk function R(θ,δ) is continuous in θ for every δ, then all Bayes rules are admissible.

By contrast, generalized Bayes rules often have undefined Bayes risk in the case of improper priors. These rules are often inadmissible and the verification of their admissibility can be difficult. For example, the generalized Bayes estimator of a location parameter θ based on Gaussian samples (described in the "Generalized Bayes estimator" section above) is inadmissible for $p > 2$; this is known as Stein's phenomenon.

Asymptotic efficiency

Let θ be an unknown random variable, and suppose that $x_1, x_2, \cdots$ are iid samples with density $f(x_i \mid \theta)$. Let $\delta_n = \delta_n(x_1, \ldots, x_n)$ be a sequence of Bayes estimators of θ based on an increasing number of measurements. We are interested in analyzing the asymptotic performance of this sequence of estimators, i.e., the performance of δ_n for large *n*.

To this end, it is customary to regard θ as a deterministic parameter whose true value is θ_0. Under specific conditions, for large samples (large values of *n*), the posterior density of θ is approximately normal. In other words, for large *n*, the effect of the prior probability on the posterior is negligible. Moreover, if δ is the Bayes estimator under MSE risk, then it is asymptotically unbiased and it converges in distribution to the normal distribution:

$$\sqrt{n}(\delta_n - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right),$$

where $I(\theta_0)$ is the fisher information of θ_0. It follows that the Bayes estimator δ_n under MSE is asymptotically efficient.

Another estimator which is asymptotically normal and efficient is the maximum likelihood estimator (MLE). The relations between the maximum likelihood and Bayes estimators can be shown in the following simple example.

Consider the estimator of θ based on binomial sample *x*~b(θ,*n*) where θ denotes the probability for success. Assuming θ is distributed according to the conjugate prior, which in this case is the Beta distribution B(*a*,*b*), the posterior distribution is known to be B(a+x,b+n-x). Thus, the Bayes estimator under MSE is

$$\delta_n(x) = E[\theta|x] = \frac{a+x}{a+b+n}.$$

The MLE in this case is x/n and so we get,

$$\delta_n(x) = \frac{a+b}{a+b+n}E[\theta] + \frac{n}{a+b+n}\delta_{MLE}.$$

The last equation implies that, for $n \to \infty$, the Bayes estimator (in the described problem) is close to the MLE. On the other hand, when n is small, the prior information is still relevant to the decision problem and affects the estimate.

Practical example

The Internet Movie Database uses a formula for calculating the Top Rated 250 Titles which gives a true Bayesian estimate:

$$W = \frac{Rv + Cm}{v + m}$$

where:

W= weighted rating
R= average for the movie as a number from 0 to 10 (mean) = (Rating)
v= number of votes for the movie = (votes)
m= minimum votes required to be listed in the Top 250 (currently 3000)
C= the mean vote across the whole report (currently 6.9)

for the Top 250, only votes from regular voters are considered.

Chapter 3

Cramer–Rao Bound

In estimation theory and statistics, the **Cramér–Rao bound (CRB)** or **Cramér–Rao lower bound (CRLB)**, named in honor of Harald Cramér and Calyampudi Radhakrishna Rao who were among the first to derive it, expresses a lower bound on the variance of estimators of a deterministic parameter. The bound is also known as the **Cramér–Rao inequality** or the **information inequality**.

In its simplest form, the bound states that the variance of any unbiased estimator is at least as high as the inverse of the Fisher information. An unbiased estimator which achieves this lower bound is said to be efficient. Such a solution achieves the lowest possible mean squared error among all unbiased methods, and is therefore the minimum variance unbiased (MVU) estimator. However, in some cases, no unbiased technique exists which achieves the bound. This may occur even when an MVU estimator exists.

The Cramér–Rao bound can also be used to bound the variance of *biased* estimators of given bias. In some cases, a biased approach can result in both a variance and a mean squared error that are *below* the unbiased Cramér–Rao lower bound.

Statement

The Cramér–Rao bound is stated in this section for several increasingly general cases, beginning with the case in which the parameter is a scalar and its estimator is unbiased. All versions of the bound require certain regularity conditions, which hold for most well-behaved distributions.

Scalar unbiased case

Suppose θ is an unknown deterministic parameter which is to be estimated from measurements *x*, distributed according to some probability density function $f(x;\theta)$. The variance of any *unbiased* estimator $\hat{\theta}$of θ is then bounded by the inverse of the Fisher information *I*(θ):

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the Fisher information *I*(θ) is defined by

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \ell(x;\theta)}{\partial \theta}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(x;\theta)}{\partial \theta^2}\right]$$

and $\ell(x;\theta) = \log f(x;\theta)$is the natural logarithm of the likelihood function and E denotes the expected value.

The efficiency of an unbiased estimator $\hat{\theta}$measures how close this estimator's variance comes to this lower bound; estimator efficiency is defined as

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\mathrm{var}(\hat{\theta})}$$

or the minimum possible variance for an unbiased estimator divided by its actual variance. The Cramér–Rao lower bound thus gives

$$e(\hat{\theta}) \leq 1.$$

General scalar case

A more general form of the bound can be obtained by considering an unbiased estimator *T*(*X*) of a function ψ(θ) of the parameter θ. Here, unbiasedness is understood as stating that *E*{*T*(*X*)} = ψ(θ). In this case, the bound is given by

$$\mathrm{var}(T) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}$$

where ψ'(θ) is the derivative of ψ(θ), and *I*(θ) is the Fisher information defined above.

Apart from being a bound on estimators of functions of the parameter, this approach can be used to derive a bound on the variance of biased estimators with a given bias, as follows. Consider an estimator $\hat{\theta}$with bias $b(\theta) = E\{\hat{\theta}\} - \theta$, and let ψ(θ) = *b*(θ) + θ. By the result above, any unbiased estimator whose expectation is ψ(θ) has variance greater than or equal to $(\psi'(\theta))^2 / I(\theta)$. Thus, any estimator $\hat{\theta}$whose bias is given by a function $b(\theta)$ satisfies

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{[1 + b'(\theta)]^2}{I(\theta)}.$$

Clearly, the unbiased version of the bound is a special case of this result, with $b(\theta) = 0$.

Of course, it's trivial to have a small variance − an "estimator" that is constant has a variance of zero. But from the above equation we find that the mean squared error of a biased estimator is bounded by

$$\mathrm{E}\left((\hat{\theta} - \theta)^2\right) \geq \frac{[1 + b'(\theta)]^2}{I(\theta)} + b(\theta)^2,$$

and this can be less than the unbiased Cramér-Rao bound $1/I(\theta)$.

Multivariate case

Extending the Cramér–Rao bound to multiple parameters, define a parameter column vector

$$\boldsymbol{\theta} = \left[\theta_1, \theta_2, \dots, \theta_d\right]^T \in \mathbb{R}^d$$

with probability density function $f(x; \boldsymbol{\theta})$which satisfies the two regularity conditions below.

The Fisher information matrix is a $d \times d$matrix with element $I_{m,k}$ defined as

$$I_{m,k} = \mathrm{E}\left[\frac{d}{d\theta_m} \log f\left(x; \boldsymbol{\theta}\right) \frac{d}{d\theta_k} \log f\left(x; \boldsymbol{\theta}\right)\right].$$

Let $\boldsymbol{T}(X)$be an estimator of any vector function of parameters, $\boldsymbol{T}(X) = (T_1(X), \ldots, T_n(X))^T$, and denote its expectation vector $\mathrm{E}[\boldsymbol{T}(X)]$by $\boldsymbol{\psi}(\boldsymbol{\theta})$. The Cramér–Rao bound then states that the covariance matrix of $\boldsymbol{T}(X)$satisfies

$$\mathrm{cov}_{\boldsymbol{\theta}}\left(\boldsymbol{T}(X)\right) \geq \frac{\partial \boldsymbol{\psi}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} [I\left(\boldsymbol{\theta}\right)]^{-1} \left(\frac{\partial \boldsymbol{\psi}\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right)^T$$

where

- The matrix inequality $A \geq B$ is understood to mean that the matrix $A - B$ is positive semidefinite, and
- $\partial \boldsymbol{\psi}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a matrix whose *ij*th element is given by $\partial \psi_i(\boldsymbol{\theta})/\partial \theta_j$.

If $\boldsymbol{T}(X)$ is an unbiased estimator of $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\psi}\left(\boldsymbol{\theta}\right) = \boldsymbol{\theta}$), then the Cramér–Rao bound reduces to

$$\mathrm{cov}_{\boldsymbol{\theta}}\left(\boldsymbol{T}(X)\right) \geq I\left(\boldsymbol{\theta}\right)^{-1}.$$

Regularity conditions

The bound relies on two weak regularity conditions on the probability density function, *f*(*x*;θ), and the estimator *T*(*X*):

- The Fisher information is always defined; equivalently, for all *x* such that *f*(*x*;θ) > 0,

  $$\frac{\partial}{\partial\theta} \ln f(x;\theta)$$

  exists, and is finite.

- The operations of integration with respect to *x* and differentiation with respect to θ can be interchanged in the expectation of *T*; that is,

  $$\frac{\partial}{\partial\theta}\left[\int T(x) f(x;\theta)\,dx\right] = \int T(x)\left[\frac{\partial}{\partial\theta} f(x;\theta)\right]\,dx$$

  whenever the right-hand side is finite.
  This condition can often be confirmed by using the fact that integration and differentiation can be swapped when either of the following cases hold:

  1. The function *f*(*x*;θ) has bounded support in *x*, and the bounds do not depend on θ;
  2. The function *f*(*x*;θ) has infinite support, is continuously differentiable, and the integral converges uniformly for all θ.

Simplified form of the Fisher information

Suppose, in addition, that the operations of integration and differentiation can be swapped for the second derivative of *f*(*x*;θ) as well, i.e.,

$$\frac{\partial^2}{\partial\theta^2}\left[\int T(x) f(x;\theta)\,dx\right] = \int T(x)\left[\frac{\partial^2}{\partial\theta^2} f(x;\theta)\right]\,dx.$$

In this case, it can be shown that the Fisher information equals

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

The Cramèr–Rao bound can then be written as

$$\mathrm{var}\left(\widehat{\theta}\right) \geq \frac{1}{I(\theta)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]}.$$

In some cases, this formula gives a more convenient technique for evaluating the bound.

Single-parameter proof

The following is a proof of the general scalar case of the Cramér–Rao bound, which was described above; namely, that if the expectation of *T* is denoted by ψ(θ), then, for all θ,

$$\mathrm{var}(t(X)) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}.$$

Let *X* be a random variable with probability density function *f*(*x*;θ). Here *T* = *t*(*X*) is a statistic, which is used as an estimator for ψ(θ). If *V* is the score, i.e.

$$V = \frac{\partial}{\partial\theta}\ln f(X;\theta)$$

then the expectation of *V*, written E(*V*), is zero. If we consider the covariance cov(*V*,*T*) of *V* and *T*, we have cov(*V*,*T*) = E(*VT*), because E(*V*) = 0. Expanding this expression we have

$$\mathrm{cov}(V,T) = \mathrm{E}\left(T\cdot\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)$$

This may be expanded using the chain rule

$$\frac{\partial}{\partial\theta}\ln Q = \frac{1}{Q}\frac{\partial Q}{\partial\theta}$$

and the definition of expectation gives, after cancelling *f*(*x*;θ),

$$\mathrm{E}\left(T\cdot\frac{\partial}{\partial\theta}\ln f(X;\theta)\right) = \int t(x)\left[\frac{\partial}{\partial\theta}f(x;\theta)\right]\,dx = \frac{\partial}{\partial\theta}\left[\int t(x)f(x;\theta)\,dx\right] = \psi'(\theta)$$

because the integration and differentiation operations commute (second condition).

The Cauchy–Schwarz inequality shows that

$$\sqrt{\operatorname{var}(T)\operatorname{var}(V)} \geq |\operatorname{cov}(V,T)| = |\psi'(\theta)|$$

therefore

$$\operatorname{var} T \geq \frac{[\psi'(\theta)]^2}{\operatorname{var}(V)} = \frac{[\psi'(\theta)]^2}{I(\theta)} = \left[\frac{\partial}{\partial\theta}\mathrm{E}(T)\right]^2 \frac{1}{I(\theta)}$$

which proves the proposition.

Examples

Multivariate normal distribution

For the case of a *d*-variate normal distribution

$$\boldsymbol{x} \sim N_d\left(\boldsymbol{\mu}\left(\boldsymbol{\theta}\right), \boldsymbol{C}\left(\boldsymbol{\theta}\right)\right)$$

the Fisher information matrix has elements

$$I_{m,k} = \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_m} \boldsymbol{C}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_k} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_m}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_k}\right)$$

where "tr" is the trace.

For example, let *w*[*n*] be a sample of *N* independent observations) with unknown mean θ and known variance σ^2

$$w[n] \sim \mathbb{N}_N\left(\theta\boldsymbol{1}, \sigma^2\boldsymbol{I}\right).$$

Then the Fisher information is a scalar given by

$$I(\theta) = \left(\frac{\partial\boldsymbol{\mu}(\theta)}{\partial\theta}\right)^T \boldsymbol{C}^{-1}\left(\frac{\partial\boldsymbol{\mu}(\theta)}{\partial\theta}\right) = \sum_{i=1}^{N}\frac{1}{\sigma^2} = \frac{N}{\sigma^2},$$

and so the Cramér–Rao bound is

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{\sigma^2}{N}.$$

Normal variance with known mean

Suppose *X* is a normally distributed random variable with known mean μ and unknown variance σ^2. Consider the following statistic:

$$T = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}.$$

Then *T* is unbiased for σ^2, as $E(T) = \sigma^2$. What is the variance of *T*?

$$\operatorname{var}(T) = \frac{\operatorname{var}(X-\mu)^2}{n} = \frac{1}{n}\left[E\left\{(X-\mu)^4\right\} - \left(E\left\{(X-\mu)^2\right\}\right)^2\right]$$

(the second equality follows directly from the definition of variance). The first term is the fourth moment about the mean and has value $3(\sigma^2)^2$; the second is the square of the variance, or $(\sigma^2)^2$. Thus

$$\operatorname{var}(T) = \frac{2(\sigma^2)^2}{n}.$$

Now, what is the Fisher information in the sample? Recall that the score *V* is defined as

$$V = \frac{\partial}{\partial \sigma^2} \log L(\sigma^2, X)$$

where *L* is the likelihood function. Thus in this case,

$$V = \frac{\partial}{\partial \sigma^2} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X-\mu)^2/2\sigma^2}\right] = \frac{(X-\mu)^2}{2(\sigma^2)^2} - \frac{1}{2\sigma^2}$$

where the second equality is from elementary calculus. Thus, the information in a single observation is just minus the expectation of the derivative of *V*, or

$$I = -E\left(\frac{\partial V}{\partial \sigma^2}\right) = -E\left(-\frac{(X-\mu)^2}{(\sigma^2)^3} + \frac{1}{2(\sigma^2)^2}\right) = \frac{\sigma^2}{(\sigma^2)^3} - \frac{1}{2(\sigma^2)^2} = \frac{1}{2(\sigma^2)^2}.$$

Thus the information in a sample of *n* independent observations is just *n* times this, or $\frac{n}{2(\sigma^2)^2}.$

The Cramer Rao bound states that

$$\mathrm{var}(T) \geq \frac{1}{I}.$$

In this case, the inequality is saturated (equality is achieved), showing that the estimator is efficient.

However, we can achieve a lower mean squared error using a biased estimator. The estimator

$$T = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n+2}.$$

obviously has a smaller variance, which is in fact

$$\mathrm{var}(T) = \frac{2n(\sigma^2)^2}{(n+2)^2}.$$

Its bias is

$$\left(1 - \frac{n}{n+2}\right) \sigma^2 = \frac{2\sigma^2}{n+2}$$

so its mean squared error is

$$\mathrm{MSE}(T) = \left(\frac{2n}{(n+2)^2} + \frac{4}{(n+2)^2}\right) (\sigma^2)^2 = \frac{2(\sigma^2)^2}{n+2}$$

which is clearly less than the Cramér-Rao bound found above.

When the mean is not known, the minimum mean squared error estimate of the variance of a sample from Gaussian distribution is achieved by dividing by *n*+1, rather than *n*-1 or *n*+2.

Chapter 4

Extended Kalman Filter

In estimation theory, the **extended Kalman filter** (EKF) is the nonlinear version of the Kalman filter which linearizes about the current mean and covariance. The EKF has been considered the *de facto* standard in the theory of nonlinear state estimation, navigation systems and GPS.

Formulation

In the extended Kalman filter, the state transition and observation models need not be linear functions of the state but may instead be differentiable functions.

$$\begin{aligned} \mathbf{x}_k &= f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}) + \mathbf{w}_{k-1} \\ \mathbf{z}_k &= h(\mathbf{x}_k) + \mathbf{v}_k \end{aligned}$$

Where $\mathbf{w}_k$ and $\mathbf{v}_k$ are the process and observation noises which are both assumed to be zero mean multivariate Gaussian noises with covariance $\mathbf{Q}_k$ and $\mathbf{R}_k$ respectively.

The function *f* can be used to compute the predicted state from the previous estimate and similarly the function *h* can be used to compute the predicted measurement from the predicted state. However, *f* and *h* cannot be applied to the covariance directly. Instead a matrix of partial derivatives (the Jacobian) is computed.

At each timestep the Jacobian is evaluated with current predicted states. These matrices can be used in the Kalman filter equations. This process essentially linearizes the non-linear function around the current estimate.

Predict and update equations

Predict

Predicted state $\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_{k-1})$

Predicted estimate covariance $\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1}\mathbf{P}_{k-1|k-1}\mathbf{F}_{k-1}^{\top} + \mathbf{Q}_{k-1}$

Update

Innovation or measurement residual $\tilde{\mathbf{y}}_k = \mathbf{z}_k - h(\hat{\mathbf{x}}_{k|k-1})$

Innovation (or residual) covariance $\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^{\top} + \mathbf{R}_k$

Near-Optimal Kalman gain $\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^{\top}\mathbf{S}_k^{-1}$

Updated state estimate $\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\tilde{\mathbf{y}}_k$

Updated estimate covariance $\mathbf{P}_{k|k} = (I - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1}$

where the state transition and observation matrices are defined to be the following Jacobians

$$\mathbf{F}_{k-1} = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k-1|k-1},\mathbf{u}_{k-1}}$$

$$\mathbf{H}_k = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k|k-1}}$$

Continuous-time extended Kalman filter

Model

$$\begin{aligned}\dot{\mathbf{x}}(t) &= f\big(\mathbf{x}(t), \mathbf{u}(t)\big) + \mathbf{w}(t), & \mathbf{w}(t) &\sim N\big(\mathbf{0}, \mathbf{Q}(t)\big)\\ \mathbf{z}(t) &= h\big(\mathbf{x}(t)\big) + \mathbf{v}(t), & \mathbf{v}(t) &\sim N\big(\mathbf{0}, \mathbf{R}(t)\big)\end{aligned}$$

Initialize

$$\hat{\mathbf{x}}(t_0) = E\big[\mathbf{x}(t_0)\big],\ \mathbf{P}(t_0) = Var\big[\mathbf{x}(t_0)\big]$$

Predict-Update

$$\begin{aligned}
\dot{\hat{\mathbf{x}}}(t) &= f\big(\hat{\mathbf{x}}(t),\mathbf{u}(t)\big)+\mathbf{K}(t)\Big(\mathbf{z}(t)-h\big(\hat{\mathbf{x}}(t)\big)\Big)\\
\dot{\mathbf{P}}(t) &= \mathbf{F}(t)\mathbf{P}(t)+\mathbf{P}(t)\mathbf{F}(t)^{\top}-\mathbf{K}(t)\mathbf{H}(t)\mathbf{P}(t)+\mathbf{Q}(t)\\
\mathbf{K}(t) &= \mathbf{P}(t)\mathbf{H}(t)^{\top}\mathbf{R}(t)^{-1}\\
\mathbf{F}(t) &= \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t),\mathbf{u}(t)}\\
\mathbf{H}(t) &= \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t)}
\end{aligned}$$

Unlike discrete-time extended Kalman filter, the prediction and update steps are coupled in continuous-time extended Kalman filter.

Continuous-discrete extended Kalman filter

Most physical systems are represented as continuous-time models while discrete-time measurements are frequently taken for state estimation via a digital processor. Therefore, the system model and measurement model are given by

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= f\big(\mathbf{x}(t),\mathbf{u}(t)\big)+\mathbf{w}(t), &\mathbf{w}(t) &\sim N\big(\mathbf{0},\mathbf{Q}(t)\big)\\
\mathbf{z}_k &= h(\mathbf{x}_k)+\mathbf{v}_k, &\mathbf{v}_k &\sim N(\mathbf{0},\mathbf{R}_k)
\end{aligned}$$

where $\mathbf{x}_k = \mathbf{x}(t_k)$.

Initialize

$$\hat{\mathbf{x}}_{0|0} = E\big[\mathbf{x}(t_0)\big], \mathbf{P}_{0|0} = Var\big[\mathbf{x}(t_0)\big]$$

Predict

$$\begin{cases}\dot{\hat{\mathbf{x}}}(t) = f\big(\hat{\mathbf{x}}(t),\mathbf{u}(t)\big)\\ \dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t)+\mathbf{P}(t)\mathbf{F}(t)^{\top}+\mathbf{Q}(t)\end{cases}\text{, with }\begin{cases}\hat{\mathbf{x}}(t_{k-1}) = \hat{\mathbf{x}}_{k-1|k-1}\\ \mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1|k-1}\end{cases}$$

$$\Rightarrow\begin{cases}\hat{\mathbf{x}}_{k|k-1} = \hat{\mathbf{x}}(t_k)\\ \mathbf{P}_{k|k-1} = \mathbf{P}(t_k)\end{cases}$$

where

$$\mathbf{F}(t) = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}(t),\mathbf{u}(t)}$$

Update

$$\begin{aligned}
&\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^\top\big(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\top + \mathbf{R}_k\big)^{-1}\\
&\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\big(\mathbf{z}_k - h(\hat{\mathbf{x}}_{k|k-1})\big)\\
&\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1}
\end{aligned}$$

where

$$\mathbf{H}_k = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{k|k-1}}$$

The update equations are identical to those of discrete-time extended Kalman filter.

Disadvantages of the extended Kalman filter

Unlike its linear counterpart, the extended Kalman filter in general is *not* an optimal estimator (of course it is optimal if the measurement and the state transition model are both linear, as in that case the extended Kalman filter is identical to the regular one). In addition, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge, owing to its linearization. Another problem with the extended Kalman filter is that the estimated covariance matrix tends to underestimate the true covariance matrix and therefore risks becoming inconsistent in the statistical sense without the addition of "stabilising noise".

Having stated this, the extended Kalman filter can give reasonable performance, and is arguably the *de facto* standard in navigation systems and GPS.

Unscented Kalman filters

An improvement to the extended Kalman filter led to the development of the Unscented Kalman filter (UKF), also a nonlinear filter. In the UKF, the probability density is approximated by the nonlinear transformation of a random variable, which returns much more accurate results than the first-order Taylor expansion of the nonlinear functions in the EKF. The approximation utilizes a set of sample points, which guarantees accuracy with the posterior mean and covariance to the second order for any nonlinearity. The UKF tends to be more robust and more accurate than the EKF in its estimation of error.

"The extended Kalman filter (EKF) is probably the most widely used estimation algorithm for nonlinear systems. However, more than 35 years of experience in the estimation community has shown that is difficult to implement, difficult to tune, and only reliable for systems that are almost linear on the time scale of the updates. Many of these difficulties arise from its use of linearization."

Invariant extended Kalman filter

The invariant extended Kalman filter (IEKF) is a modified version of the EKF for nonlinear systems possessing symmetries (or *invariances*). It combines the advantages of both the EKF and the recently introduced symmetry-preserving filters. Indeed, instead of using a linear correction term based on a linear output error, it uses a geometrically adapted correction term based on an invariant output error; in the same way the gain matrix is not updated from a linear state error, but from an invariant state error. The main benefit is that the gain and covariance equations converge to constant values on a much bigger set of trajectories than equilibrium points as it is the case for the EKF, which results in a better convergence of the estimation.

Chapter 5

Fisher Information

In mathematical statistics and information theory, the **Fisher information** (sometimes simply called **information**) is the variance of the score. In Bayesian statistics, the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior (according to the Bernstein–von Mises Theorem, which was anticipated by Laplace for exponential families). The role of the Fisher information in the asymptotic theory of maximum-likelihood estimation was emphasized by the statistician R.A. Fisher (following some initial results by F. Y. Edgeworth). The Fisher information is also used in the calculation of the Jeffreys prior, which is used in Bayesian statistics.

The Fisher-information matrix is used to calculate the covariance matrices associated with maximum-likelihood estimates. It can also be used in the formulation of test statistics, such as the Wald test.

History

The Fisher information was discussed by several early statisticians, notably F. Y. Edgeworth. For example, Savage says: "In it [Fisher information], he [Fisher] was to some extent anticipated (Edgeworth 1908–9 esp. 502, 507–8, 662, 677–8, 82–5 and references he [Edgeworth] cites including Pearson and Filon 1898 [. . .])." There are a number of early historical sources and a number of reviews of this early work.

Definition

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends. The probability function for X, which is also the likelihood function for θ, is a function $f(X; \theta)$; it is the probability mass (or probability density) of the random variable X conditional on the value of θ. The partial derivative with respect to θ of the log of the likelihood function is called the score. Under certain regularity conditions, it can be shown that the first moment of the score is 0. The second moment is called the Fisher information:

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left.\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right|\theta\right],$$

where, for any given value of θ, the expression E[…|*θ*] denotes the conditional expectation over values for *X* with respect to the probability function *f*(*x*; *θ*) given *θ*. Note that $0 \leq \mathcal{I}(\theta) < \infty$. A random variable carrying high Fisher information implies that the absolute value of the score is often high.The Fisher information is not a function of a particular observation, as the random variable *X* has been averaged out.

Since the expectation of the score is zero, the Fisher information is also the variance of the score.

If ln *f*(*x*; *θ*) is twice differentiable with respect to *θ*, and under certain regularity conditions, then the Fisher information may also be written as

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\left.\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right|\theta\right].$$

Thus, the Fisher information is the negative of the expectation of the second derivative with respect to *θ* of the log of *f*. Information may be seen to be a measure of the "curvature" of the support curve near the maximum likelihood estimate of θ. A "blunt" support curve (one with a shallow maximum) would have a low negative expected second derivative, and thus low information; while a sharp one would have a high negative expected second derivative and thus high information.

Information is additive, in that the information yielded by two independent experiments is the sum of the information from each experiment separately:

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

This result follows from the elementary fact that if random variables are independent, the variance of their sum is the sum of their variances. Hence the information in a random sample of size *n* is *n* times that in a sample of size 1 (if observations are independent).

The information provided by a sufficient statistic is the same as that of the sample *X*. This may be seen by using Neyman's factorization criterion for a sufficient statistic. If *T*(*X*) is sufficient for θ, then

$$f(X;\theta) = g(T(X),\theta)h(X)$$

for some functions *g* and *h*. The equality of information then follows from the following fact:

$$\frac{\partial}{\partial\theta}\ln\left[f(X;\theta)\right] = \frac{\partial}{\partial\theta}\ln\left[g(T(X);\theta)\right]$$

which follows from the definition of Fisher information, and the independence of *h*(*X*) from θ. More generally, if *T* = *t*(*X*) is a statistic, then

$$\mathcal{I}_T(\theta) \leq \mathcal{I}_X(\theta)$$

with equality if and only if *T* is a sufficient statistic.

Informal derivation of the Cramér–Rao bound

The Cramér–Rao bound states that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. Van Trees (1968) and Frieden (2004) provide the following method of deriving the Cramér–Rao bound, a result which describes use of the Fisher information, informally:

Consider an unbiased estimator $\hat\theta(X)$. Mathematically, we write

$$\mathrm{E}\left[\hat\theta(X) - \theta\middle|\theta\right] = \int\left[\hat\theta(X) - \theta\right]\cdot f(X;\theta)\,dx = 0.$$

The likelihood function *f*(*X*; *θ*) describes the probability that we observe a given sample *x given* a known value of *θ*. If *f* is sharply peaked with respect to changes in θ, it is easy to intuit the "correct" value of θ given the data, and hence the data contains a lot of information about the parameter. If the likelihood *f* is flat and spread-out, then it would take many, many samples of *X* to estimate the actual "true" value of *θ*. Therefore, we would intuit that the data contain much less information about the parameter.

Now, we differentiate the unbiased-ness condition above to get

$$\frac{\partial}{\partial\theta}\int\left[\hat\theta(X) - \theta\right]\cdot f(X;\theta)\,dx = \int\left(\hat\theta - \theta\right)\frac{\partial f}{\partial\theta}\,dx - \int f\,dx = 0.$$

We now make use of two facts. The first is that the likelihood *f* is just the probability of the data given the parameter. Since it is a probability, it must be normalized, implying that

$$\int f\,dx = 1.$$

Second, we know from basic calculus that

$$\frac{\partial f}{\partial \theta} = f \, \frac{\partial \ln f}{\partial \theta}.$$

Using these two facts in the above let us write

$$\int \left(\hat{\theta} - \theta\right) f \, \frac{\partial \ln f}{\partial \theta} \, dx = 1.$$

Factoring the integrand gives

$$\int \left(\left(\hat{\theta} - \theta\right) \sqrt{f}\right) \left(\sqrt{f} \, \frac{\partial \ln f}{\partial \theta}\right) \, dx = 1.$$

If we square the equation, the Cauchy–Schwarz inequality lets us write

$$\left[\int \left(\hat{\theta} - \theta\right)^2 f \, dx\right] \cdot \left[\int \left(\frac{\partial \ln f}{\partial \theta}\right)^2 f \, dx\right] \geq 1.$$

The right-most factor is defined to be the Fisher Information

$$\mathcal{I}\left(\theta\right) = \int \left(\frac{\partial \ln f}{\partial \theta}\right)^2 f \, dx.$$

The left-most factor is the expected mean-squared error of the estimator *θ*, since

$$\mathrm{E}\left[\left(\hat{\theta}\left(X\right) - \theta\right)^2 \middle| \theta\right] = \int \left(\hat{\theta} - \theta\right)^2 f \, dx.$$

Notice that the inequality tells us that, fundamentally,

$$\mathrm{Var}\left[\hat{\theta}\right] \, \geq \, \frac{1}{\mathcal{I}\left(\theta\right)}.$$

In other words, the precision to which we can estimate *θ* is fundamentally limited by the Fisher Information of the likelihood function.

Single-parameter Bernoulli experiment

A Bernoulli trial is a random variable with two possible outcomes, "success" and "failure", with "success" having a probability of θ. The outcome can be thought of as determined by a coin toss, with the probability of obtaining a "head" being θ and the probability of obtaining a "tail" being $1 - \theta$.

The Fisher information contained in n independent Bernoulli trials may be calculated as follows. In the following, A represents the number of successes, B the number of failures, and $n = A + B$ is the total number of trials.

$$\begin{aligned}
\mathcal{I}(\theta) &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(A;\theta))\middle|\theta\right] \qquad (1)\\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln\left(\theta^A(1-\theta)^B\frac{(A+B)!}{A!B!}\right)\middle|\theta\right] \qquad (2)\\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\left(A\ln(\theta)+B\ln(1-\theta)\right)\middle|\theta\right] \qquad (3)\\
&= -\mathrm{E}\left[\frac{\partial}{\partial\theta}\left(\frac{A}{\theta}-\frac{B}{1-\theta}\right)\middle|\theta\right]\text{(on differentiating } \ln x\text{)} \; (4)\\
&= +\mathrm{E}\left[\frac{A}{\theta^2}+\frac{B}{(1-\theta)^2}\middle|\theta\right] \qquad (5)\\
&= \frac{n\theta}{\theta^2}+\frac{n(1-\theta)}{(1-\theta)^2}\text{(as the expected value of } A \text{ given } \theta \text{ is } n\theta\text{, etc.)} \; (6)\\
&= \frac{n}{\theta(1-\theta)} \qquad (7)
\end{aligned}$$

(1) defines Fisher information. (2) invokes the fact that the information in a sufficient statistic is the same as that of the sample itself. (3) expands the log term and drops a constant. (4) and (5) differentiate with respect to θ. (6) replaces A and B with their expectations. (7) is algebra.

The end result, namely,

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)},$$

is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, as expected.

Matrix form

When there are *N* parameters, so that θ is a *N*x1 vector $\theta = \begin{bmatrix}\theta_1, \theta_2, \dots, \theta_N\end{bmatrix}^{\mathrm{T}}$, then the Fisher information takes the form of an *N*x*N* matrix, the Fisher Information Matrix (FIM), with typical element:

$$\left(\mathcal{I}\left(\theta\right)\right)_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\ln f(X;\theta)\right)\middle|\theta\right].$$

The FIM is a *N*x*N* positive semidefinite symmetric matrix, defining a Riemannian metric on the *N*-dimensional parameter space, thus connecting Fisher information to differential geometry. In that context, this metric is known as the Fisher information metric, and the topic is called information geometry.

Under certain regularity conditions, the Fisher Information Matrix may also be written as:

$$\left(\mathcal{I}\left(\theta\right)\right)_{i,j} = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\middle|\theta\right]\,.$$

Orthogonal parameters

We say that two parameters θ_i and θ_j are orthogonal if the element of the *i*th row and *j*th column of the Fisher information matrix is zero. Orthogonal parameters are easy to deal with in the sense that their maximum likelihood estimates are independent and can be calculated separately. When dealing with research problems, it is very common for the researcher to invest some time searching for an orthogonal parametrization of the densities involved in the problem.

Multivariate normal distribution

The FIM for a *N*-variate multivariate normal distribution has a special form. Let $\mu(\theta) = \begin{bmatrix}\mu_1(\theta), \mu_2(\theta), \dots, \mu_N(\theta)\end{bmatrix}^{\mathrm{T}}$, and let Σ(*θ*) be the covariance matrix. Then the typical element $\mathcal{I}_{m,n}$, $0 \le m, n < M$, of the FIM for $X \sim N(\mu(\theta), \Sigma(\theta))$ is:

$$\mathcal{I}_{m,n} = \frac{\partial\mu^{\mathrm{T}}}{\partial\theta_m}\Sigma^{-1}\frac{\partial\mu}{\partial\theta_n} + \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_m}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_n}\right),$$

where $(..)^{\mathrm{T}}$ denotes the transpose of a vector, tr(..) denotes the trace of a square matrix, and:

- $\frac{\partial\mu}{\partial\theta_m} = \begin{bmatrix}\frac{\partial\mu_1}{\partial\theta_m} & \frac{\partial\mu_2}{\partial\theta_m} & \cdots & \frac{\partial\mu_N}{\partial\theta_m}\end{bmatrix}^{\mathrm{T}};$

- $$\frac{\partial \Sigma}{\partial \theta_m} = \begin{bmatrix} \frac{\partial \Sigma_{1,1}}{\partial \theta_m} & \frac{\partial \Sigma_{1,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{1,N}}{\partial \theta_m} \\ \frac{\partial \Sigma_{2,1}}{\partial \theta_m} & \frac{\partial \Sigma_{2,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{2,N}}{\partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \Sigma_{N,1}}{\partial \theta_m} & \frac{\partial \Sigma_{N,2}}{\partial \theta_m} & \cdots & \frac{\partial \Sigma_{N,N}}{\partial \theta_m} \end{bmatrix}.$$

Note that a special, but very common, case is the one where Σ(θ) = Σ, a constant. Then

$$\mathcal{I}_{m,n} = \frac{\partial \mu^{\mathrm{T}}}{\partial \theta_m} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_n}.$$

In this case the Fisher information matrix may be identified with the coefficient matrix of the normal equations of least squares estimation theory.

Properties

Reparametrization

The Fisher information depends on the parametrization of the problem. If θ and η are two scalar parametrizations of an estimation problem, and θ is a continuously differentiable function of η, then

$$\mathcal{I}_\eta(\eta) = \mathcal{I}_\theta(\theta(\eta)) \left(\frac{\mathrm{d}\theta}{\mathrm{d}\eta} \right)^2$$

where $\mathcal{I}_\eta$and $\mathcal{I}_\theta$are the Fisher information measures of η and θ, respectively.

In the vector case, suppose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$are k-vectors which parametrize an estimation problem, and suppose that $\boldsymbol{\theta}$ is a continuously differentiable function of $\boldsymbol{\eta}$, then,

$$\mathcal{I}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \boldsymbol{J}^{\mathrm{T}} \mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{\eta})) \boldsymbol{J}$$

where the (i, j)th element of the k × k Jacobian matrix $\boldsymbol{J}$is defined by

$$J_{ij} = \frac{\partial \theta_i}{\partial \eta_j},$$

and where $\boldsymbol{J}^{\mathrm{T}}$is the matrix transpose of $\boldsymbol{J}$.

In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrization.

Applications

Optimal design of experiments

Fisher information is widely used in optimal experimental design. Because of the reciprocity of estimator-variance and Fisher information, ***minimizing* the *variance*** corresponds to ***maximizing* the *information***.

When the linear (or linearized) statistical model has several parameters, the mean of the parameter-estimator is a vector and its variance is a matrix. The inverse matrix of the variance-matrix is called the "information matrix". Because the variance of the estimator of a parameter vector is a matrix, the problem of "minimizing the variance" is complicated. Using statistical theory, statisticians compress the information-matrix using real-valued summary statistics; being real-valued functions, these "information criteria" can be maximized.

Traditionally, statisticians have evaluated estimators and designs by considering some summary statistic of the covariance matrix (of a mean-unbiased estimator), usually with positive real values (like the determinant or matrix trace). Working with positive real-numbers brings several advantages: If the estimator of a single parameter has a positive variance, then the variance and the Fisher information are both positive real numbers; hence they are members of the convex cone of nonnegative real numbers (whose nonzero members have reciprocals in this same cone). For several parameters, the covariance-matrices and information-matrices are elements of the convex cone of nonnegative-definite symmetric matrices in a partially ordered vector space, under the Loewner (Löwner) order. This cone is closed under matrix-matrix addition, under matrix-inversion, and under the multiplication of positive real-numbers and matrices. An exposition of matrix theory and the Loewner-order appears in Pukelsheim.

The traditional optimality-criteria are the information-matrix's invariants; algebraically, the traditional optimality-criteria are functionals of the eigenvalues of the (Fisher) information matrix.

Jeffreys prior in Bayesian statistics

In Bayesian statistics, the Fisher information is used to calculate the Jeffreys prior, which is a standard, non-informative prior for continuous distribution parameters.

Relation to KL-divergence

Fisher information is the curvature of the Kullback–Leibler information.

Distinction from the Hessian of the entropy

In general, the Fisher Information

$$\mathcal{I}(\theta) = \int f(X;\theta)\left(-\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)dx = \int f(X;\theta)\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2 dx$$

is not the same as the negative of the second derivative of the entropy

$$-\frac{\partial^2 H}{\partial\theta^2} = \frac{\partial^2}{\partial\theta^2}\int f(X;\theta)\ln f(X;\theta)\,dx\;.$$

For instance, with $f(X;\theta) = \frac{e^{-(x-\theta)^2/2}}{\sqrt{2\pi}}$, the entropy *H* is independent of the distribution mean *θ*. Thus, in this case, the second derivative of the entropy is zero. However, for the Fisher information, we have $\mathcal{I}(\theta) = 1\;.$

Chapter 6

Generalized Method of Moments

In econometrics, **generalized method of moments** (**GMM**) is a generic method for estimating parameters in statistical models. Usually it is applied in the context of semiparametric models, where the parameter of interest is finite-dimensional, whereas the full shape of the distribution function of the data may not be known, and therefore the maximum likelihood estimation is not applicable.

The method requires that a certain number of *moment conditions* were specified for the model. These moment conditions are functions of the model parameters and the data, such that their expectation is zero at the true values of the parameters. The GMM method then minimizes a certain norm of the sample averages of the moment conditions.

The GMM estimators are known to be consistent, asymptotically normal, and efficient in the class of all estimators that don't use any extra information aside from that contained in the moment conditions.

GMM was developed by Lars Peter Hansen in 1982 as a generalization of the method of moments.

Description

Suppose the available data consists of T of iid observations $\{Y_t\}_{t=1,\dots,T}$, where each observation Y_t is an n-dimensional multivariate random variable. The data comes from a certain statistical model, defined up to an unknown parameter $\theta \in \Theta$. The goal of the estimation problem is to find the "true" value of this parameter, θ_0, or at least a reasonably close estimate.

In order to apply GMM there should exist a vector-valued function $g(Y,\theta)$ such that

$$m(\theta_0) \equiv \mathrm{E}[\,g(Y_t,\theta_0)\,] = 0,$$

where E denotes expectation, and Y_t is a generic observation, which are all assumed to be iid. Moreover, function $m(\theta)$ must not be equal to zero for $\theta \neq \theta_0$, or otherwise parameter θ will not be identified.

The basic idea behind GMM is to replace the theoretical expected value E[·] with its empirical analog — sample average:

$$\hat{m}(\theta) = \hat{\mathrm{E}}\big[\,g(Y_t,\theta)\,\big] \equiv \frac{1}{T}\sum_{t=1}^{T} g(Y_t,\theta)$$

and then to minimize the norm of this expression with respect to θ.

By the law of large numbers, $\hat{m}(\theta) \approx \mathrm{E}[g(Y_t,\theta)] = m(\theta)$for large values of T, and thus we expect that $\hat{m}(\theta_0) \approx m(\theta_0) = 0$. The generalized method of moments looks for a number $\hat{\theta}$which would make $\hat{m}(\hat{\theta})$as close to zero as possible. Mathematically, this is equivalent to minimizing a certain norm of $\hat{m}(\theta)$(norm of m, denoted as $||m||$, measures the distance between m and zero). The properties of the resulting estimator will depend on the particular choice of the norm function, and therefore the theory of GMM considers an entire family of norms, defined as

$$\|\hat{m}(\theta)\|^2_W = \hat{m}(\theta)'\,W\hat{m}(\theta),$$

where W is a positive-definite weighting matrix, and m' denotes transposition. In practice, the weighting matrix W is computed based on the available data set, which will be denoted as $\hat{W}$. Thus, the GMM estimator can be written as

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta\in\Theta} \bigg(\frac{1}{T}\sum_{t=1}^{T} g(Y_t,\theta)\bigg)' \hat{W} \bigg(\frac{1}{T}\sum_{t=1}^{T} g(Y_t,\theta)\bigg)$$

Under suitable conditions this estimator is consistent, asymptotically normal, and with right choice of weighting matrix $\hat{W}$asymptotically efficient.

Properties

Consistency

Consistency is the most important property of an estimator. It means that having sufficient number of observations, the estimator will get arbitrarily close to the true value of parameter:

$$\hat{\theta} \xrightarrow{p} \theta_0 \ \text{ as } T \to \infty$$

Necessary and sufficient conditions for a GMM estimator to be consistent are as follows:

1. $\hat{W}_T \xrightarrow{p} W,$where *W* is a positive semi-definite matrix,
2. $W\mathrm{E}[\,g(Y_t,\theta)\,] = 0$ only for $\theta = \theta_0,$
3. $\theta_0 \in \Theta,$which is compact,
4. $g(Y,\theta)$ is continuous at each *θ* with probability one,
5. $\mathrm{E}[\,\sup_{\theta\in\Theta}\|g(Y,\theta)\|\,] < \infty.$

The second condition here (so-called **Global identification** condition) is often particularly hard to verify. There exist simpler necessary but not sufficient conditions, which may be used to detect non-identification problem:

- **Order condition**. The dimension of moment function *m(θ)* should be at least as large as the dimension of parameter vector *θ*.
- **Local identification**. If *g(Y,θ)* is continuously differentiable in a neighborhood of θ_0, then matrix $W\mathrm{E}[\nabla_\theta g(Y_t,\theta_0)]$must have full column rank.

In practice applied econometricians often simply *assume* that global identification holds, without actually proving it.

Asymptotic normality

Asymptotic normality is a useful property, as it allows us to construct confidence bands for the estimator, and conduct different tests. Before we can make a statement about the asymptotic distribution of the GMM estimator, we need to define two auxiliary matrices:

$$G = \mathrm{E}[\,\nabla_\theta\, g(Y_t,\theta_0)\,], \qquad \Omega = \mathrm{E}[\,g(Y_t,\theta_0)g(Y_t,\theta_0)'\,]$$

Then under conditions 1–6 listed below, the GMM estimator will be asymptotically normal with limiting distribution

$$\sqrt{T}\big(\hat\theta - \theta_0\big) \xrightarrow{d} \mathcal{N}\big[0,(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}\big]$$

Conditions:

1. $\hat\theta$is consistent,
2. $\theta_0$lies in the interior of set $\Theta,$
3. $g(Y,\theta)$is continuously differentiable in some neighborhood *N* of θ_0 with probability one,
4. $\mathrm{E}[\,\|g(Y_t,\theta)\|^2\,] < \infty,$
5. $\mathrm{E}[\,\sup_{\theta\in N}\|\nabla_\theta g(Y_t,\theta)\|\,] < \infty,$
6. matrix *G'WG* is nonsingular.

Efficiency

So far we have said nothing about the choice of matrix *W*, except that it must be positive semi-definite. In fact any such matrix will produce a consistent and asymptotically normal GMM estimator, the only difference will be in the asymptotic variance of that estimator. It can be shown that taking

$$W \propto \Omega^{-1}$$

will result in the most efficient estimator in the class of all asymptotically normal estimators. Efficiency in this case means that such an estimator will have the smallest possible variance (we say that matrix *A* is smaller than matrix *B* if *B–A* is positive semi-definite).

In this case the formula for the asymptotic distribution of the GMM estimator simplifies to

$$\sqrt{T}\big(\hat{\theta}-\theta_0\big) \xrightarrow{d} \mathcal{N}\big[0,(G'\,\Omega^{-1}G)^{-1}\big]$$

The proof that such a choice of weighting matrix is indeed optimal is quite elegant, and is often adopted with slight modifications when establishing efficiency of other estimators. As a rule of thumb, a weighting matrix is optimal whenever it makes the “sandwich formula” for variance collapse into a simpler expression.

Proof. We will consider the difference between asymptotic variance with arbitrary *W* and asymptotic variance with $W = \Omega^{-1}$. If we can factor this difference into a symmetric product of the form *CC′* for some matrix *C*, then it will guarantee that this difference is nonnegative-definite, and thus $W = \Omega^{-1}$ will be optimal by definition.

$$\begin{aligned} V(W)-V(\Omega^{-1}) &= (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}-(G'\Omega^{-1}G)^{-1} \\ &= (G'WG)^{-1}\Big(G'W\Omega WG - G'WG(G'\Omega^{-1}G)^{-1}G'WG\Big)(G'WG)^{-1} \\ &= (G'WG)^{-1}G'W\Omega^{1/2}\Big(I-\Omega^{-1/2}G(G'\Omega^{-1}G)^{-1}G'\Omega^{-1/2}\Big)\Omega^{1/2}WG(G'WG)^{-1} \\ &= A(I-B)A', \end{aligned}$$

where we introduced matrices *A* and *B* in order to slightly simplify notation; *I* is an identity matrix. We can see that matrix *B* here is symmetric and idempotent: $B^2 = B$. This means *I–B* is symmetric and idempotent as well: $I - B = (I - B)(I - B)'$. Thus we can continue to factor the previous expression as

$$= A(I-B)(I-B)'A' = \Big(A(I-B)\Big)\Big(A(I-B)\Big)' \geq 0$$

Implementation

One difficulty with implementing the outlined method is that we cannot take $W = \Omega^{-1}$ because, by the definition of matrix Ω, we need to know the value of θ_0 in order to compute this matrix, and θ_0 is precisely the quantity we don't know and are trying to estimate in the first place.

Several approaches exist to deal with this issue, the first one being the most popular:

- **Two-step feasible GMM**:
  - *Step 1*: Take $W = I$ (the identity matrix), and compute preliminary GMM estimate $\hat{\theta}_{(1)}$. This estimator is consistent for θ_0, although not efficient.
  - *Step 2*: Take

$$\hat{W} = \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(1)}) g(Y_t, \hat{\theta}_{(1)})'\right)^{-1},$$

where we have plugged in our first-step preliminary estimate $\hat{\theta}_{(1)}$. This matrix converges in probability to Ω^{-1} and therefore if we compute $\hat{\theta}$with this weighting matrix, the estimator will be asymptotically efficient.

- **Iterated GMM**. Essentially the same procedure as 2-step GMM, except that the matrix $\hat{W}_T$is recalculated several times. That is, the estimate obtained in step 2 is used to calculate the weighting matrix for step 3, and so on. Such estimator, denoted $\hat{\theta}_{(i)}$, is equivalent to solving the following system of equations:

$$\left(\frac{1}{T}\sum_{t=1}^{T} \frac{\partial g}{\partial \theta'}(Y_t, \hat{\theta}_{(i)})\right)' \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(i)}) g(Y_t, \hat{\theta}_{(i)})'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \hat{\theta}_{(i)})\right) = 0$$

Asymptotically no improvement can be achieved through such iterations, although certain Monte-Carlo experiments suggest that finite-sample properties of this estimator are slightly better.

- **Continuously Updating GMM** (CUGMM, or CUE). Estimates $\hat{\theta}$simultaneously with estimating the weighting matrix W:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta)\right)' \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta) g(Y_t, \theta)'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \theta)\right)$$

In Monte-Carlo experiments this method demonstrated a better performance than

> the traditional two-step GMM: the estimator has smaller median bias (although fatter tails), and the J-test for overidentifying restrictions in many cases was more reliable.

Another important issue is implementation of minimization procedure is that the function is supposed to search through (possibly high-dimensional) parameter space *Θ* and find the value of *θ* which minimizes the objective function. No generic recommendation for such procedure exists, it is a subject of its own field, numerical optimization.

J-test

When the number of moment conditions is greater than the dimension of the parameter vector *θ*, the model is said to be *over-identified*. Over-identification allows us to check whether the model's moment conditions match the data well or not.

Conceptually we can check whether $\hat{m}(\hat{\theta})$is sufficiently close to zero to suggest that the model fits the data well. The GMM method has then replaced the problem of solving the equation $\hat{m}(\theta) = 0$, which chooses θ to match the restrictions exactly, by a minimization calculation. The minimization can always be conducted even when no θ_0 exists such that $m(\theta_0) = 0$. This is what J-test does. The J-test is also called a *test for over-identifying restrictions*.

Formally we consider two hypotheses:

- $H_0 :\ m(\theta_0) = 0$ (the null hypothesis that the model is "valid"), and
- $H_1 :\ m(\theta) \neq 0,\ \forall\theta \in \Theta$ (the alternative hypothesis that model is "invalid"; the data do not come close to meeting the restrictions)

Under hypothesis H_0, the following so-called J-statistic is asymptotically *chi-squared* with *k–l* degrees of freedom. Define *J* to be:

$$J \equiv T \cdot \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \hat{\theta})\right)' \hat{W}_T \left(\frac{1}{T}\sum_{t=1}^{T} g(Y_t, \hat{\theta})\right) \xrightarrow{d} \chi^2_{k-\ell}$$ under H_0,

where $\hat{\theta}$is the GMM estimator of the parameter θ_0, *k* is the number of moment conditions (dimension of vector *g*), and *l* is the number of estimated parameters (dimension of vector *θ*). Matrix $\hat{W}_T$must converge in probability to Ω^{-1}, the efficient weighting matrix (note that previously we only required that *W* be proportional to Ω^{-1} for estimator to be efficient; however in order to conduct the J-test *W* must be exactly equal to Ω^{-1}, not simply proportional).

Under the alternative hypothesis H_1, the J-statistic is asymptotically unbounded:

$J \xrightarrow{p} \infty$ under H_1

To conduct the test we compute the value of J from the data. It is a nonnegative number. We compare it with (say) the 0.95 quantile of the $\chi^2_{k-\ell}$distribution:

- H_0 is *rejected* at 95% confidence level if $J > q_{0.95}^{\chi^2_{k-\ell}}$
- H_0 cannot be rejected at 95% confidence level if $J < q_{0.95}^{\chi^2_{k-\ell}}$

Scope

Many other popular estimation techniques can be cast in terms of GMM optimization:

- Ordinary Least Squares (OLS) is equivalent to GMM with moment conditions:

$$\mathrm{E}[\, x_t(y_t - x_t'\beta)\,] = 0$$

- Generalized Least Squares (GLS)

$$\mathrm{E}[\, x_t(y_t - x_t'\beta)/\sigma^2(x_t)\,] = 0$$

- Instrumental variables regression (IV)

$$\mathrm{E}[\, z_t(y_t - x_t'\beta)\,] = 0$$

- Non-linear Least Squares (NLLS):

$$\mathrm{E}[\, \nabla_\beta\, g(x_t,\beta) \cdot (y_t - g(x_t,\beta))\,] = 0$$

- Maximum likelihood estimation (MLE):

$$\mathrm{E}[\, \nabla_\theta \ln f(x_t,\theta)\,] = 0$$

Chapter 7

Estimator

In statistics, an **estimator** is a rule for calculating an estimate of a given quantity based on observed data: thus the rule and its result (the estimate) are distinguished. Here we discusses estimators that are point estimators; that is, they yield single-valued results, although this includes the possibility of single vector-valued results and results that can be expressed as a single function. This is in contrast to an interval estimator, where the result would be a range of plausible values (or vectors or functions).

Statistical theory is concerned with the properties of estimators; that is, with defining properties that can be used to compare different estimators (different rules for creating estimates) for the same quantity, based on the same data. Such properties can be used to determine the best rules to use under given circumstances. However, in robust statistics, statistical theory goes on to consider the balance between having good properties, if tightly defined assumptions hold, and having less good properties that hold under wider conditions.

Background

An "estimator" or "point estimate" is a statistic (that is, a measurable function of the data) that is used to infer the value of an unknown parameter in a statistical model. The parameter being estimated is sometimes called the *estimand*. It can be either finite-dimensional (in parametric and semi-parametric models), or infinite-dimensional (semi-nonparametric and non-parametric models). If the parameter is denoted θ then the estimator is typically written by adding a “hat” over the symbol: $\hat{\theta}$. Being a function of the data, the estimator is itself a random variable; a particular realization of this random variable is called the "estimate". Sometimes the words “estimator” and “estimate” are used interchangeably.

The definition places virtually no restrictions on which functions of the data can be called the “estimators”. The attractiveness of different estimators can be judged by looking at their properties, such as unbiasedness, mean square error, consistency, asymptotic distribution, etc.. The construction and comparison of estimators are the subjects of the estimation theory. In the context of decision theory, an estimator is a type of decision rule, and its performance may be evaluated through the use of loss functions.

When the word “estimator” is used without a qualifier, it usually refers to point estimation. The estimate in this case is a single point in the parameter space. Other types of estimators also exist: interval estimators, where the estimates are subsets of the parameter space.

The problem of density estimation arises in two applications. Firstly, in estimating the probability density functions of random variables and secondly in estimating the spectral density function of a time series. In these problems the estimates are functions that can be thought of as point estimates in an infinite dimensional space, and there are corresponding interval estimation problems.

Definition

Suppose there is a fixed *parameter* θthat needs to be estimated. Then an "estimator" is a function that maps the sample space to a set of *sample estimates*. An estimator of $\theta$is usually denoted by the symbol $\widehat{\theta}$. It is often convenient to express the theory using the algebra of random variables: thus if *X* is used to denote a random variable corresponding to the observed data, the estimator (itself treated as a random variable) is symbolised as a function of that random variable, $\widehat{\theta}(X)$. The estimate for a particular observed dataset (i.e. for *X*=*x*) is then $\widehat{\theta}(x)$, which is a fixed value. Often an abbreviated notation is used in which $\widehat{\theta}$ is interpreted directly as a random variable, but this can cause confusion.

Quantified properties

The following definitions and attributes apply:

Error

For a given sample x, the "error" of the estimator $\widehat{\theta}$ is defined as

$$e(x) = \widehat{\theta}(x) - \theta,$$

where $\theta$is the parameter being estimated. Note that the error, *e*, depends not only on the estimator (the estimation formula or procedure), but on the sample.

Mean squared error

The *mean squared error* of $\widehat{\theta}$ is defined as the expected value (probability-weighted average, over all samples) of the squared errors; that is,

$$\operatorname{MSE}(\widehat{\theta}) = \operatorname{E}[(\widehat{\theta}(X) - \theta)^2].$$

It is used to indicate how far, on average, the collection of estimates are from the single parameter being estimated. Consider the following analogy. Suppose the parameter is the bull's-eye of a target, the estimator is the process of shooting arrows at the target, and the individual arrows are estimates (samples). Then high MSE means the average distance of the arrows from the bull's-eye is high, and low MSE means the average distance from the bull's-eye is low. The arrows may or may not be clustered. For example, even if all arrows hit the same point, yet grossly miss the target, the MSE is still relatively large. Note, however, that if the MSE is relatively low, then the arrows are likely more highly clustered (than highly dispersed).

Sampling deviation

For a given sample x, the *sampling deviation* of the estimator $\widehat{\theta}$ is defined as

$$d(x) = \widehat{\theta}(x) - \operatorname{E}(\widehat{\theta}(X)) = \widehat{\theta}(x) - \operatorname{E}(\widehat{\theta}),$$

where $\operatorname{E}(\widehat{\theta}(X))$ is the expected value of the estimator. Note that the sampling deviation, *d*, depends not only on the estimator, but on the sample.

Variance

The *variance* of $\widehat{\theta}$ is simply the expected value of the squared sampling deviations; that is, $\operatorname{var}(\widehat{\theta}) = \operatorname{E}[(\widehat{\theta} - \operatorname{E}(\widehat{\theta}))^2]$. It is used to indicate how far, on average, the collection of estimates are from the *expected value* of the estimates. Note the difference between MSE and variance. If the parameter is the bull's-eye of a target, and the arrows are estimates, then a relatively high variance means the arrows are dispersed, and a relatively low variance means the arrows are clustered. Some things to note: even if the variance is low, the cluster of arrows may still be far off-target, and even if the variance is high, the diffuse collection of arrows may still be unbiased. Finally, note that even if all arrows grossly miss the target, if they nevertheless all hit the same point, the variance is zero.

Bias

The *bias* of $\widehat{\theta}$ is defined as $B(\widehat{\theta}) = \operatorname{E}(\widehat{\theta}) - \theta$. It is the distance between the average of the collection of estimates, and the single parameter being estimated. It also is the expected value of the error, since $\operatorname{E}(\widehat{\theta}) - \theta = \operatorname{E}(\widehat{\theta} - \theta)$. If the parameter is the bull's-eye of a target, and the arrows are estimates, then a relatively high absolute value for the bias means the average position of the arrows is off-target, and a relatively low absolute bias means the average position of the arrows is on target. They may be dispersed, or may be clustered. The relationship between bias and variance is analogous to the relationship between accuracy and precision.

Unbiased

The estimator $\widehat{\theta}$ is an *unbiased estimator* of θif and only if $B(\widehat{\theta}) = 0$. Note that bias is a property of the estimator, not of the estimate. Often, people refer to a "biased estimate" or an "unbiased estimate," but they really are talking about an "estimate from a biased estimator," or an "estimate from an unbiased estimator." Also, people often confuse the "error" of a single estimate with the "bias" of an estimator. Just because the error for one estimate is large, does not mean the estimator is biased. In fact, even if all estimates have astronomical absolute values for their errors, if the expected value of the error is zero, the estimator is unbiased. Also, just because an estimator is biased, does not preclude the error of an estimate from being zero (we may have gotten lucky). The ideal situation, of course, is to have an unbiased estimator with low variance, and also try to limit the number of samples where the error is extreme (that is, have few outliers). Yet unbiasedness is not essential. Often, if just a little bias is permitted, then an estimator can be found with lower MSE and/or fewer outlier sample estimates.

An alternative to the version of "unbiased" above, is "median-unbiased", where the median of the distribution of estimates agrees with the true value; thus, in the long run half the estimates will be too low and half too high. While this applies immediately only to scalar-valued estimators, it can be extended to any measure of central tendency of a distribution.

Relationships

- The MSE, variance, and bias, are related: $\operatorname{MSE}(\widehat{\theta}) = \operatorname{var}(\widehat{\theta}) + (B(\widehat{\theta}))^2,$ i.e. mean squared error = variance + square of bias. In particular, for an unbiased estimator, the variance equals the MSE.
- The standard deviation of an estimator of θ (the square root of the variance), or an estimate of the standard deviation of an estimator of θ, is called the *standard error* of θ.

Behavioural properties

Consistency

A consistent sequence of estimators is a sequence of estimators that converge in probability to the quantity being estimated as the index (usually the sample size) grows without bound. In other words, increasing the sample size increases the probability of the estimator being close to the population parameter.

Mathematically, a sequence of estimators $\{t_n; n \geq 0\}$ is a consistent estimator for parameter θ if and only if, for all $\epsilon > 0$, no matter how small, we have

$$\lim_{n\to\infty} \Pr\left\{|t_n - \theta| < \epsilon\right\} = 1.$$

The consistency defined above may be called weak consistency. The sequence is *strongly consistent*, if it converges almost surely to the true value.

An estimator that converges to a *multiple* of a parameter can be made into a consistent estimator by multiplying the estimator by a scale factor, namely the true value divided by the asymptotic value of the estimator. This occurs frequently in estimation of scale parameters by measures of statistical dispersion.

Asymptotic normality

An asymptotically normal estimator is a consistent estimator whose distribution around the true parameter θ approaches a normal distribution with standard deviation shrinking in proportion to $1/\sqrt{n}$ as the sample size n grows. Using $\xrightarrow{D}$ to denote convergence in distribution, t_n is asymptotically normal if

$$\sqrt{n}(t_n - \theta) \xrightarrow{D} N(0, V),$$

for some V, which is called the *asymptotic variance* of the estimator.

The central limit theorem implies asymptotic normality of the sample mean $\bar{x}$ as an estimator of the true mean. More generally, maximum likelihood estimators are asymptotically normal under fairly weak regularity conditions section of the maximum likelihood article. However, not all estimators are asymptotically normal, the simplest examples being case where the true value of a parameter lies in the boundary of the allowable parameter region.

Efficiency

Two naturally desirable properties of estimators are for them to be unbiased and have minimal mean squared error (MSE). These cannot in general both be satisfied simultaneously: a biased estimator may have lower mean squared error (MSE) than any unbiased estimator: despite having bias, the estimator variance may be sufficiently smaller than that of any unbiased estimator, and it may be preferable to use, despite the bias.

Among unbiased estimators, there often exists one with the lowest variance, called the minimum variance unbiased estimator (MVUE). In some cases an unbiased efficient estimator exists, which, in addition to having the lowest variance among unbiased estimators, satisfies the Cramér–Rao bound, which is an absolute lower bound on variance for statistics of a variable.

Chapter 8

Invariant Estimator

In statistics, the concept of being an **invariant estimator** is a criterion that can be used to compare the properties of different estimators for the same quantity. It is a way of formalising the idea that an estimator should have certain intuitively appealing qualities. Strictly speaking, "invariant" would mean that the estimates themselves are unchanged when both the measurements and the parameters are transformed in a compatible way, but the meaning has been extended to allow the estimates to change in appropriate ways with such transformations. The term **equivariant estimator** is used in formal mathematical contexts that include a precise description of the relation of the way the estimator changes in response to changes to the dataset and parameterisation: this corresponds to the use of "equivariance" in more general mathematics.

General setting

Background

In statistical inference, there are several approaches to estimation theory that can be used to decide immediately what estimators should be used according to those approaches. For example, ideas from Bayesian inference would lead directly to Bayesian estimators. Similarly, the theory of classical statistical inference can sometimes lead to strong conclusions about what estimator should be used. However, the usefulness of these theories depends on having a fully prescribed statistical model and may also depend on having a relevant loss function to determine the estimator. Thus a Bayesian analysis might be undertaken, leading to a posterior distribution for relevant parameters, but the use of a specific utility or loss function may be unclear. Ideas of invariance can then be applied to the task of summarising the posterior distribution. In other cases, statistical analyses are undertaken without a fully defined statistical model or the classical theory of statistical inference cannot be readily applied because the family of models being considered are not amenable to such treatment. In addition to these cases where general theory does not prescribe an estimator, the concept of invariance of an estimator can be applied when seeking estimators of alternative forms, either for the sake of simplicity of application of the estimator or so that the estimator is robust.

The concept of invariance is sometimes used on its own as a way of choosing between estimators, but this is not necessarily definitive. For example, a requirement of invariance may be incompatible with the requirement that the estimator be mean-unbiased; on the other hand, the criterion of median-unbiasedness is defined in terms of the estimator's sampling distribution and so is invariant under many transformations.

One use of the concept of invariance is where a class or family of estimators is proposed and a particular formulation must be selected amongst these. One procedure is to impose relevant invariance properties and then to find the formulation within this class that has the best properties, leading to what is called the optimal invariant estimator.

Some classes of invariant estimators

There are several types of transformations that are usefully considered when dealing with invariant estimators. Each gives rise to a class of estimators which are invariant to those particular types of transformation.

- Shift invariance: Notionally, estimates of a location parameter should be invariant to simple shifts of the data values. If all data values are increased by a given amount, the estimate should change by the same amount. When considering estimation using a weighted average, this invariance requirement immediately implies that the weights should sum to one. While the same result is often derived from a requirement for unbiasedness, the use of "invariance" does not require that a mean value exists and makes no use of any probability distribution at all.
- Scale invariance: Note that this is a topic not directly covered in scale invariance.
- Parameter-transformation invariance: Here, the transformation applies to the parameters alone. The concept here is that essentially the same inference should be made from data and a model involving a parameter θ as would be made from the same data if the model used a parameter φ, where φ is a one-to-one transformation of θ, $\varphi=h(\theta)$. According to this type of invariance, results from transformation-invariant estimators should also be related by $\varphi=h(\theta)$. Maximum likelihood estimators have this property.
- Permutation invariance: Where a set of data values can be represented by a statistical model that they are outcomes from independent and identically distributed random variables, it is reasonable to impose the requirement that any estimator of any property of the common distribution should be permutation-invariant: specifically that the estimator, considered as a function of the set of data-values, should not change if items of data are swapped within the dataset.

The combination of permutation invariance and location invariance for estimating a location parameter from an independent and identically distributed dataset using a weighted average implies that the weights should be identical and sum to one. Of course, estimators other than a weighted average may be preferable.

Optimal invariant estimators

Under this setting, we are given a set of measurements x which contains information about an unknown parameter θ. The measurements x are modelled as a vector random variable having a probability density function $f(x \mid \theta)$ which depends on a parameter vector θ.

The problem is to estimate θ given x. The estimate, denoted by a, is a function of the measurements and belongs to a set A. The quality of the result is defined by a loss function $L = L(a,\theta)$ which determines a risk function $R = R(a,\theta) = E[L(a,\theta) \mid \theta]$. The sets of possible values of x, θ, and a are denoted by X, Θ, and A, respectively.

Mathematical setting

Definition

An invariant estimator is an estimator which obeys the following two rules:

1. Principle of Rational Invariance: The action taken in a decision problem should not depend on transformation on the measurement used
2. Invariance Principle: If two decision problems have the same formal structure (in terms of X, Θ, $f(x \mid \theta)$ and L), then the same decision rule should be used in each problem.

To define an invariant or equivariant estimator formally, some definitions related to groups of transformations are needed first. Let X denote the set of possible data-samples. A group of transformations of X, to be denoted by G, is a set of (measurable) 1:1 and onto transformations of X into itself, which satisfies the following conditions:

1. If $g_1 \in G$ and $g_2 \in G$ then $g_1 g_2 \in G$
2. If $g \in G$ then $g^{-1} \in G$, where $g^{-1}(g(x)) = x$.(That is, each transformation has an inverse within the group.)
3. $e \in G$(i.e there is an identity transformation $e(x) = x$)

Datasets x_1 and x_2 in X are equivalent if $x_1 = g(x_2)$ for some $g \in G$. All the equivalent points form an equivalence class. Such an equivalence class is called an orbit (in X). The x_0 orbit, $X(x_0)$, is the set $X(x_0) = \{g(x_0) : g \in G\}$. If X consists of a single orbit then g is said to be transitive.

A family of densities *F* is said to be invariant under the group *G* if, for every $g \in G$and $\theta \in \Theta$there exists a unique $\theta^* \in \Theta$such that *Y* = *g*(*x*) has density *f*(*y* | θ *). θ * will be denoted $\bar{g}(\theta)$.

If *F* is invariant under the group *G* then the loss function *L*(θ,*a*) is said to be invariant under *G* if for every $g \in G$and $a \in A$there exists an $a^* \in A$ such that $L(\theta, a) = L(\bar{g}(\theta), a^*)$for all $\theta \in \Theta$. The transformed value *a* * will be denoted by $\tilde{g}(a)$.

In the above, $\bar{G} = \{\bar{g} : g \in G\}$is a group of transformations from Θ to itself and $\tilde{G} = \{\tilde{g} : g \in G\}$is a group of transformations from *A* to itself.

An estimation problem is invariant(equivariant) under *G* if there exist three groups $G, \bar{G}, \tilde{G}$as defined above.

For an estimation problem that is invariant under *G*, estimator δ(*x*) is an invariant estimator under *G* if, for all $x \in X$and $g \in G$,

$$\delta(g(x)) = \tilde{g}(\delta(x)).$$

Properties

1. The risk function of an invariant estimator, δ, is constant on orbits of Θ. Equivalently $R(\theta, \delta) = R(\bar{g}(\theta), \delta)$for all $\theta \in \Theta$and $\bar{g} \in \bar{G}$.
2. The risk function of an invariant estimator with transitive $\bar{g}$is constant.

For a given problem, the invariant estimator with the lowest risk is termed the "best invariant estimator". Best invariant estimator cannot always be achieved. A special case for which it can be achieved is the case when $\bar{g}$is transitive.

Example: Location parameter

Suppose θ is a location parameter if the density of *X* is of the form *f*(*x* − θ). For $\Theta = A = \mathbb{R}^1$and *L* = *L*(*a* − θ), the problem is invariant under $g = \bar{g} = \tilde{g} = \{g_c : g_c(x) = x + c, c \in \mathbb{R}\}$. The invariant estimator in this case must satisfy

$$\delta(x + c) = \delta(x) + c, \text{ for all } c \in \mathbb{R},$$

thus it is of the form δ(*x*) = *x* + *K* ($K \in \mathbb{R}$). $\bar{g}$is transitive on Θ so the risk does not vary with θ: that is, $R(\theta, \delta) = R(0, \delta) = \mathrm{E}[L(X + K)|\theta = 0]$. The best invariant estimator is the one that brings the risk *R*(θ,δ) to minimum.

In the case that L is the squared error $\delta(x) = x - \mathrm{E}[X|\theta = 0].$

Pitman estimator

The estimation problem is that $X = (X_1, \dots, X_n)$has density $f(x_1 - \theta, \dots, x_n - \theta)$, where *θ* is a parameter to be estimated, and where the loss function is *L*(| *a* − θ |). This problem is invariant with the following (additive) transformation groups:

$$\begin{aligned} G &= \{g_c : g_c(x) = (x_1 + c, \dots, x_n + c), c \in \mathbb{R}^1\}, \\ \bar{G} &= \{g_c : g_c(\theta) = \theta + c, c \in \mathbb{R}^1\}, \\ \tilde{G} &= \{g_c : g_c(a) = a + c, c \in \mathbb{R}^1\}. \end{aligned}$$

The best invariant estimator δ(*x*) is the one that minimizes

$$\frac{\int_{-\infty}^{\infty} L(\delta(x) - \theta) f(x_1 - \theta, \dots, x_n - \theta) d\theta}{\int_{-\infty}^{\infty} f(x_1 - \theta, \dots, x_n - \theta) d\theta},$$

and this is Pitman's estimator (1939).

For the squared error loss case, the result is

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \theta f(x_1 - \theta, \dots, x_n - \theta) d\theta}{\int_{-\infty}^{\infty} f(x_1 - \theta, \dots, x_n - \theta) d\theta}.$$

If $x \sim N(\theta 1_n, I)$(i.e. a multivariate normal distribution with independent, unit-variance components) then

$$\delta_{pitman} = \delta_{ML} = \frac{\sum x_i}{n}.$$

If $x \sim C(\theta 1_n, I\sigma^2)$(independent components having a Cauchy distribution with scale parameter *σ*) then $\delta_{pitman} \neq \delta_{ML}$,. However the result is

$$\delta_{pitman} = \sum_{k=1}^{n} x_k \left[\frac{Re\{w_k\}}{\sum_{m=1}^{n} Re\{w_k\}} \right], \qquad n > 1,$$

with

$$w_k = \prod_{j \neq k} \left[\frac{1}{(x_k - x_j)^2 + 4\sigma^2} \right] \left[1 - \frac{2\sigma}{(x_k - x_j)} i \right].$$

Chapter 9

James–Stein Estimator & Kaplan–Meier Estimator

James–Stein Estimator

The **James–Stein estimator** is a nonlinear estimator which can be shown to dominate, or outperform, the "ordinary" (least squares) technique. As such, it is the best-known example of Stein's phenomenon.

An earlier version of the estimator was developed by Charles Stein in 1956, and is sometimes referred to as **Stein's estimator**. The result was improved by Willard James and Charles Stein in 1961.

Setting

Suppose $\boldsymbol{\theta}$ is an unknown parameter vector of length *m*, and let **y** be a vector of observations of $\boldsymbol{\theta}$ of length *m*, such that the observations are normally distributed:

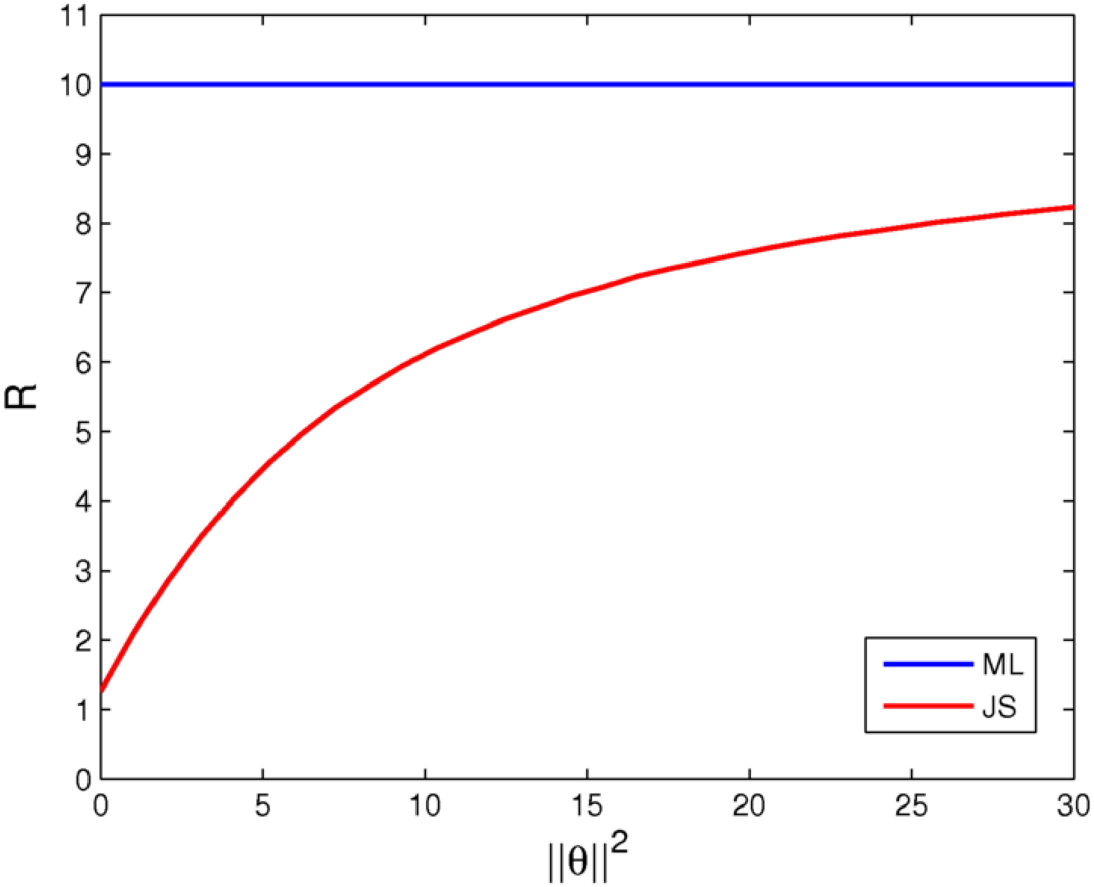
$$\mathbf{y} \sim N(\boldsymbol{\theta}, \sigma^2 I).$$

We are interested in obtaining an estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$, based on the observations **y**.

This is an everyday situation in which a set of parameters is measured, and the measurements are corrupted by independent Gaussian noise. Since the noise has zero mean, it is very reasonable to use the measurements themselves as an estimate of the parameters. This is the approach of the least squares estimator, which is $\widehat{\boldsymbol{\theta}}_{LS} = \overline{\mathbf{y}}$, where $\overline{\mathbf{y}}$ is the average of the observations. If there is only a single observation *y*, then $\widehat{\boldsymbol{\theta}}_{LS} = y$.

As a result, there was considerable shock and disbelief when Stein demonstrated that, in terms of mean squared error $E\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\}$, this approach is suboptimal. The result became known as Stein's phenomenon.

The James–Stein estimator



MSE (R) of least squares estimator (ML) vs. James–Stein estimator (JS). The James–Stein estimator gives its best estimate when the norm of the actual parameter vector θ is near zero.

If σ is known, the James–Stein estimator is given by

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}.$$

James and Stein showed that the above estimator dominates $\widehat{\boldsymbol{\theta}}_{LS}$ for any $m \geq 3$, meaning that the James–Stein estimator always achieves lower MSE than the least squares estimator.

Notice that if $(m-2)\sigma^2 < \|\overline{\mathbf{y}}\|^2$ then this estimator simply takes the natural estimator $\overline{\mathbf{y}}$ and shrinks it towards the origin **0**. In fact this is not the only direction of shrinkage that works. Let ***ν*** be an arbitrary fixed vector of length *m*. Then there exists a James–Stein estimator that shrinks toward ***ν***, namely

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\overline{\mathbf{y}} - \boldsymbol{\nu}\|^2}\right)(\overline{\mathbf{y}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

It is interesting to note that the James–Stein estimator dominates the usual estimator for any ***ν***. A natural question to ask is whether the improvement over the usual estimator is independent of the choice of ***ν***. The answer is no. The improvement is small if $\|\boldsymbol{\theta} - \boldsymbol{\nu}\|$ is large. Thus to get a very great improvement some knowledge of the location of ***θ*** is necessary. Of course this is the quantity we are trying to estimate so we don't have this knowledge a priori. But we may have some guess as to what the mean vector is. This can be considered a disadvantage of the estimator because the choice is not objective, it depends on the beliefs of the researcher.

Stein has shown that, for $m \leq 2$, the least squares estimator is admissible, meaning that no estimator dominates it.

Interpretation

A consequence of the above discussion is the following counterintuitive result: When three or more unrelated parameters are measured, their total MSE can be reduced by using a combined estimator such as the James–Stein estimator; whereas when each parameter is estimated separately, the least squares (LS) estimator is admissible. This quirk has caused some to sarcastically ask whether, in order to estimate the speed of light, one should jointly estimate tea consumption in Taiwan and hog weight in Montana. The response is that the James–Stein estimator always improves upon the *total* MSE, i.e., the sum of the expected errors of each component. Therefore, the total MSE in measuring light speed, tea consumption and hog weight would improve by using the James–Stein estimator. However, any particular component (such as the speed of light) would improve for some parameter values, and deteriorate for others. Thus, although the James–Stein estimator dominates the LS estimator when three or more parameters are estimated, any single component does not dominate the respective component of the LS estimator.

The conclusion from this hypothetical example is that measurements should be combined if one is interested in minimizing their total MSE. For example, in a telecommunication setting, it is reasonable to combine channel tap measurements in a channel estimation scenario, as the goal is to minimize the total channel estimation error. Conversely, it is probably not reasonable to combine channel estimates of different users, since no user would want their channel estimate to deteriorate in order to improve the average network performance.

Improvements

The basic James–Stein estimator has the peculiar property that for small values of $\|\overline{\mathbf{y}}-\boldsymbol{\nu}\|$, the multiplier on $\overline{\mathbf{y}}-\boldsymbol{\nu}$ is actually negative. This can be easily remedied by replacing this multiplier by zero when it is negative. The resulting estimator is called the *positive-part James–Stein estimator* and is given by

$$\widehat{\boldsymbol{\theta}}_{JS+}=\left(1-\frac{(m-2)\sigma^2}{\|\overline{\mathbf{y}}-\boldsymbol{\nu}\|^2}\right)^+(\overline{\mathbf{y}}-\boldsymbol{\nu})+\boldsymbol{\nu}.$$

This estimator has a smaller risk than the basic James–Stein estimator. It follows that the basic James–Stein estimator is itself inadmissible.

It turns out, however, that the positive-part estimator is also inadmissible. This follows from a general result which requires admissible estimators to be smooth.

Extensions

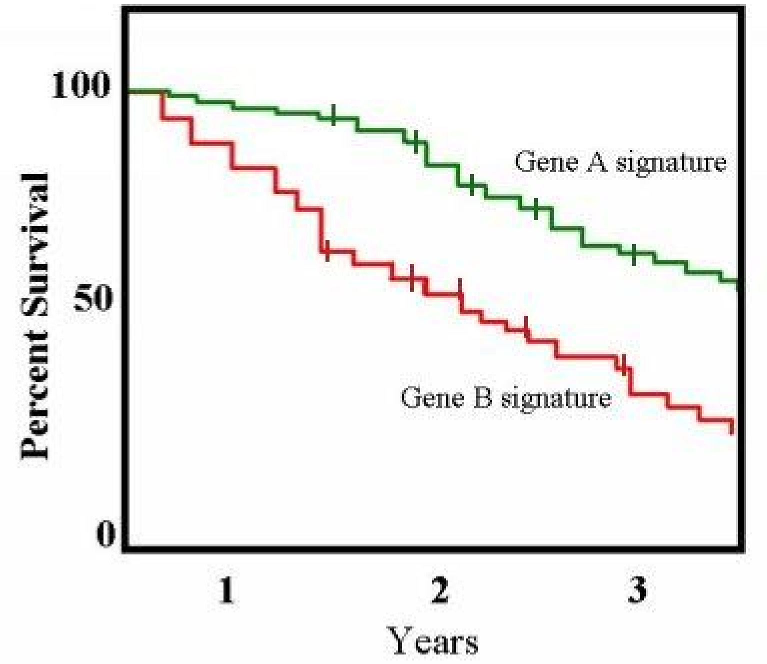
The James–Stein estimator may seem at first sight to be a result of some peculiarity of the problem setting. In fact, the estimator exemplifies a very wide-ranging effect, namely, the fact that the "ordinary" or least squares estimator is often inadmissible for simultaneous estimation of several parameters. This effect has been called Stein's phenomenon, and has been demonstrated for several different problem settings, some of which are briefly outlined below.

- James and Stein demonstrated that the estimator presented above can still be used when the variance σ^2 is unknown, by replacing it with the standard estimator of the variance, $\widehat{\sigma}^2=\frac{1}{n}\sum(y_i-\overline{y})^2$. The dominance result still holds under the same condition, namely, *m* > 2.
- Bock extended the work of James and Stein to the case of a general measurement covariance matrix, i.e., where measurements may be statistically dependent and may have differing variances. A similar dominating estimator can be constructed, with a suitably generalized dominance condition. This can be used to construct a linear regression technique which outperforms the standard application of the LS estimator.
- Stein's result was substantially extended by Lawrence D. Brown to a wide class of distributions and loss functions. However, his theorem is an existence result only, in that explicit dominating estimators were not actually exhibited. It is quite difficult to obtain explicit estimators improving upon the usual estimator without specific restrictions on the underlying distributions.

Kaplan–Meier Estimator

The **Kaplan–Meier estimator** (named after Edward L. Kaplan and Paul Meier), also known as the **product limit estimator**, estimates the survival function from life-time data. In medical research, it might be used to measure the fraction of patients living for a certain amount of time after treatment. An economist might measure the length of time people remain unemployed after a job loss. An engineer might measure the time until failure of machine parts. An ecologist may use it to estimate how long fleshy fruits remain on plants before they are removed by frugivores.

A plot of the Kaplan–Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant.



An example of a Kaplan–Meier plot for two conditions associated with patient survival

An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly *right-censoring*, which occurs if a patient withdraws from a study, i.e. is lost from the sample before the final outcome is observed. On the plot, small vertical tick-marks indicate losses, where a patient's survival time has been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is equivalent to the empirical distribution.

In medical statistics, a typical application might involve grouping patients into categories, for instance, those with Gene A profile and those with Gene B profile. In the graph, patients with Gene B die much more quickly than those with gene A. After two years about 80% of the Gene A patients still survive, but less than half of patients with Gene B.

Formulation

Let $S(t)$ be the probability that an item from a given population will have a lifetime exceeding t. For a sample from this population of size N let the observed times until death of N sample members be

$$t_1 \le t_2 \le t_3 \le \cdots \le t_N.$$

Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i, the number of deaths at time t_i.

Note that the intervals between each time typically will not be uniform. For example, a small data set might begin with 10 cases, have a death at Day 3, a loss (censored case) at Day 9, and another death at Day 11. Then we have ($t_1 = 3$, $t_2 = 11$), ($n_1 = 10$, $n_2 = 8$), and ($d_1 = 1$, $d_2 = 1$).

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of $S(t)$. It is a product of the form

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}.$$

When there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are "at risk" of an (observed) death.

There is an alternative definition that is sometimes used, namely

$$\hat{S}(t) = \prod_{t_i \le t} \frac{n_i - d_i}{n_i}.$$

The two definitions differ only at the observed event times. The latter definition is right-continuous whereas the former definition is left-continuous.

Let T be the random variable that measures the time of failure and let $F(t)$ be its cumulative distribution function. Note that

$$S(t) = P[T > t] = 1 - P[T \le t] = 1 - F(t).$$

Consequently, the right-continuous definition of $\hat{S}(t)$may be preferred in order to make the estimate compatible with a right-continuous estimate of *F*(*t*).

Statistical considerations

The Kaplan–Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common such estimators is Greenwood's formula:

$$\widehat{\mathrm{Var}}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}.$$

In some cases, one may wish to compare different Kaplan–Meier curves. This may be done by several methods including:

- The log rank test
- The Cox proportional hazards test

Chapter 10

Likelihood Function

In statistics, a **likelihood function** (often simply the **likelihood**) is a function of the parameters of a statistical model, defined as follows: the *likelihood* of a set of parameter values given some observed outcomes is equal to the *probability* of those observed outcomes given those parameter values. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics.

In non-technical parlance, "likelihood" is usually a synonym for "probability" but in statistical usage, a clear technical distinction is made. One may ask "If I were to flip a fair coin 100 times, what is the *probability* of it landing heads-up every time?" or "Given that I have flipped a coin 100 times and it has landed heads-up 100 times, what is the *likelihood* that the coin is fair?" but it would be improper to switch "likelihood" and "probability" in the two sentences.

Mathematically, writing X for the set of observed data and Θ for the set of parameter values, the expression $P(X \mid \Theta)$ (read "the probability of X given Θ") can be interpreted as the expression $L(\Theta \mid X)$, the likelihood of Θ given X. The interpretation of $L(\Theta \mid X)$ as a function of Θ is especially obvious when X is fixed and Θ is allowed to vary.

Generally, $L(\Theta \mid X)$ is permitted to be any positive multiple of $P(X \mid \Theta)$. More precisely, then, a likelihood function is any representative from an equivalence class of functions,

$$L(\Theta \mid X) \in \{\alpha \, P(X \mid \Theta) : \alpha > 0\} \ ,$$

where the constant of proportionality $\alpha > 0$ is not permitted to depend upon Θ. In particular, the numerical value $L(\Theta \mid X)$ alone is immaterial; all that matters are likelihood ratios, such as those of the form

$$\frac{L(\theta_2|X)}{L(\theta_1|X)} = \frac{\alpha P(X|\theta_2)}{\alpha P(X|\theta_1)} = \frac{P(X|\theta_2)}{P(X|\theta_1)},$$

that are invariant with respect to the constant of proportionality *α*.

A. W. F. Edwards defined **support** to be the natural logarithm of the likelihood ratio, and the **support function** as the natural logarithm of the likelihood function (the same as the log-likelihood; see below). However, there is potential for confusion with the mathematical meaning of 'support', and this terminology is not widely used outside Edwards' main applied field of phylogenetics.

Log-likelihood

For many applications involving likelihood functions, it is more convenient to work in terms of the natural logarithm of the likelihood function, called the **log-likelihood**, than in terms of the likelihood function itself. Because the logarithm is a monotonically increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques. Finding the maximum of a function often involves taking the derivative of a function and solving for the parameter being maximized, and this is often easier when the function being maximized is a log-likelihood rather than the original likelihood function.

For example, some likelihood functions are for the parameters that explain a collection of statistically independent observations. In such a situation, the likelihood function factors into a product of individual likelihood functions. The logarithm of this product is a sum of individual logarithms, and the derivative of a sum of terms is often easier to compute than the derivative of a product. In addition, several common distributions have likelihood functions that contain products of factors involving exponentiation. The logarithm of such a function is a sum of products, again easier to differentiate than the original function.

As an example, consider the gamma distribution, whose likelihood function is

$$L(\alpha, \beta|x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

and suppose we wish to find the maximum likelihood estimate of β for a single observed value *x*. This function looks rather daunting. Its logarithm, however, is much simpler to work with:

$$\log L(\alpha, \beta|x) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x \,.$$

The partial derivative with respect to β is simply

$$\frac{\partial \log L(\alpha,\beta|x)}{\partial\beta} = \frac{\alpha}{\beta} - x$$

If there are a number of samples $x_1,\dots,x_n$, then the joint log-likelihood will be the sum of individual log-likelihoods, and the derivative of this sum will be the sum of individual derivatives:

$$\frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i$$

Setting this equal to zero and solving for β yields

$$\beta^* = \frac{\alpha}{\bar{x}}$$

where β^* represents the maximum-likelihood estimate and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of the observations.

Likelihood function of a parameterized model

Among many applications, we consider here one of broad theoretical and practical importance. Given a parameterized family of probability density functions (or probability mass functions in the case of discrete distributions)

$$x \mapsto f(x \mid \theta),$$

where θ is the parameter, the **likelihood function** is

$$\theta \mapsto f(x \mid \theta),$$

written

$$L(\theta \mid x) = f(x \mid \theta),$$

where x is the observed outcome of an experiment. In other words, when $f(x \mid \theta)$ is viewed as a function of x with θ fixed, it is a probability density function, and when viewed as a function of θ with x fixed, it is a likelihood function.

Note: This is *not* the same as the probability that those parameters are the right ones, given the observed sample. Attempting to interpret the likelihood of a hypothesis given observed evidence as the probability of the hypothesis is a common error, with potentially disastrous real-world consequences in medicine, engineering or jurisprudence.

From a geometric standpoint, if we consider $f(x, \theta)$ as a function of two variables then the family of probability distributions can be viewed as level curves parallel to the x-axis, while the family of likelihood functions are the orthogonal level curves parallel to the θ-axis.

Likelihoods for continuous distributions

The use of the probability density instead of a probability in specifying the likelihood function above may be justified in a simple way. Suppose that, instead of an exact observation, x, the observation is the value in a short interval (x_{j-1}, x_j), with length Δ_j, where the subscripts refer to a predefined set of intervals. Then the probability of getting this observation (of being in interval j) is approximately

$$L_{\text{approx}}(\theta \mid x \text{ in interval } j) = \Delta_j f(x_* \mid \theta),$$

where x_* can be any point in interval j. Then, recalling that the likelihood function is defined up to a multiplicative constant, it is just as valid to say that the likelihood function is approximately

$$L_{\text{approx}}(\theta \mid x \text{ in interval } j) = f(x_* \mid \theta),$$

and then, on considering the lengths of the intervals to decrease to zero,

$$L(\theta \mid x) = f(x \mid \theta).$$

Likelihoods for mixed continuous — discrete distributions

The above can be extended in a simple way to allow consideration of distributions which contain both discrete and continuous components. Suppose that the distribution consists of a number of discrete probability masses $p_k(\theta)$ and a density $f(x \mid \theta)$, where the sum of all the p's added to the integral of f is always one. Assuming that it is possible to distinguish an observation corresponding to one of the discrete probability masses from one which corresponds to the density component, the likelihood function for an observation from the continuous component can be dealt with as above by setting the interval length short enough to exclude any of the discrete masses. For an observation from the discrete component, the probability can either be written down directly or treated within the above context by saying that the probability of getting an observation in an interval that does contain a discrete component (of being in interval j which contains discrete component k) is approximately

$$L_{\text{approx}}(\theta \mid x \text{ in interval } j \text{ containing discrete mass } k) = p_k(\theta) + \Delta_j f(x_* \mid \theta),$$

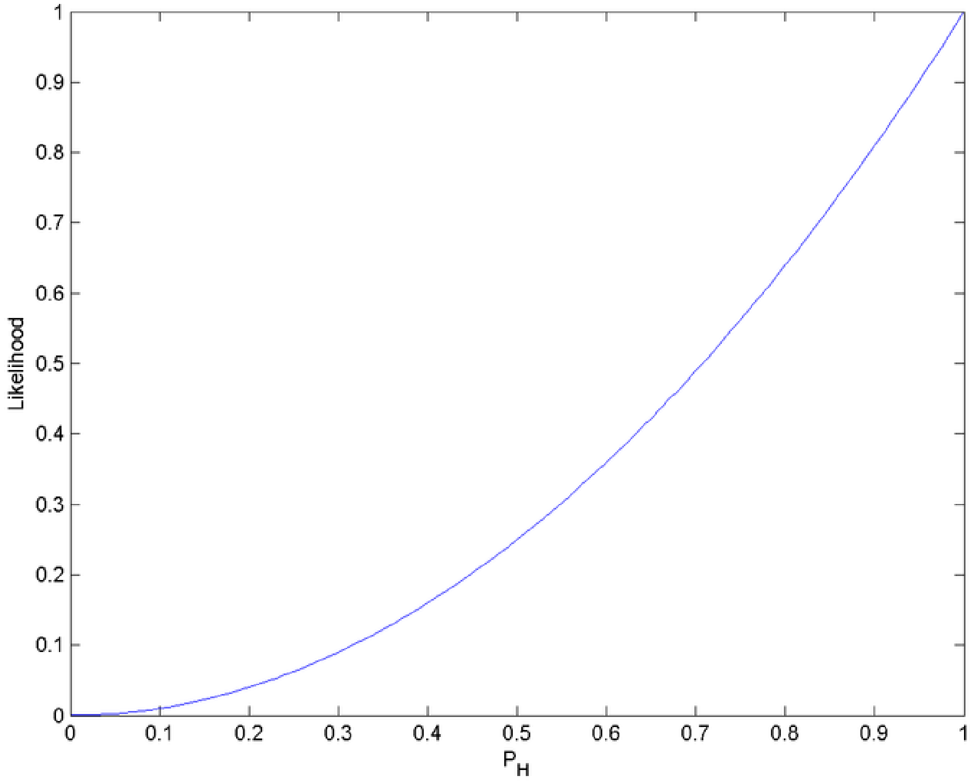
where x_* can be any point in interval j. Then, on considering the lengths of the intervals to decrease to zero, the likelihood function for a observation from the discrete component is

$$L(\theta \mid x) = p_k(\theta),$$

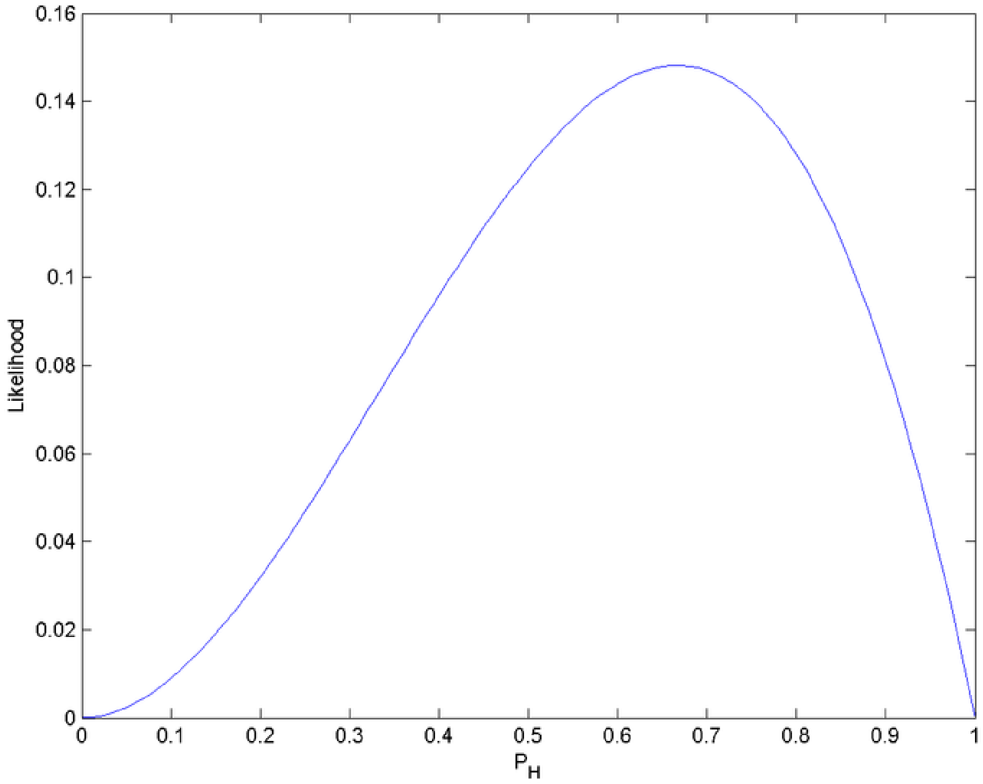
where k is the index of the discrete probability mass corresponding to observation x.

The fact that the likelihood function can be defined in a way that includes contributions that are not commensurate (the density and the probability mass) arises from the way in which the likelihood function is defined up to a constant of proportionality, where this "constant" can change with the observation x, but not with the parameter θ.

Example 1



The likelihood function for estimating the probability of a coin landing heads-up without prior knowledge after observing HH



The likelihood function for estimating the probability of a coin landing heads-up without prior knowledge after observing HHT

Let p_{H} be the probability that a certain coin lands heads up (H) when tossed. So, the probability of getting two heads in two tosses (HH) is p_{H}^2. If $p_{\mathrm{H}} = 0.5$, then the probability of seeing two heads is 0.25.

In symbols, we can say the above as:

$$P(\mathrm{HH} \mid p_{\mathrm{H}} = 0.5) = 0.25.$$

Another way of saying this is to reverse it and say that "the likelihood that $p_{\mathrm{H}} = 0.5$, given the observation HH, is 0.25"; that is:

$$L(p_{\mathrm{H}} = 0.5 \mid \mathrm{HH}) = P(\mathrm{HH} \mid p_{\mathrm{H}} = 0.5) = 0.25.$$

But this is not the same as saying that the *probability* that $p_{\mathrm{H}} = 0.5$, given the observation HH, is 0.25.

Notice that the likelihood that $p_{\mathrm{H}} = 1$, given the observation HH, is 1. But it is clearly not true that the *probability* that $p_{\mathrm{H}} = 1$, given the observation HH, is 1. Two heads in a row hardly proves that the coin *always* comes up heads. In fact, two heads in a row is possible for any $p_H > 0$.

The likelihood function is not a probability density function. Notice that the integral of a likelihood function is not in general 1. In this example, the integral of the likelihood over the interval [0, 1] in p_H is 1/3, demonstrating that the likelihood function cannot be interpreted as a probability density function for p_H.

Example 2

Consider a jar containing N lottery tickets numbered from 1 through N. If you pick a ticket randomly then you get positive integer n, with probability $1/N$ if $n \le N$ and with probability zero if $n > N$. This can be written

$$P(n|N) = \frac{[n \le N]}{N}$$

where the Iverson bracket $[n \le N]$ is 1 when $n \le N$ and 0 otherwise. When considered a function of n for fixed N this is the probability distribution, but when considered a function of N for fixed n this is a likelihood function. The maximum likelihood estimate for N is $N_0 = n$ (by contrast, the unbiased estimate is $2n - 1$).

This likelihood function is not a probability distribution, because the total

$$\sum_{N=1}^{\infty} P(n|N) = \sum_{N} \frac{[N \ge n]}{N} = \sum_{N=n}^{\infty} \frac{1}{N}$$

is a divergent series.

Suppose, however, that you pick *two* tickets rather than *one*.

The probability of the outcome $\{n_1, n_2\}$, where $n_1 < n_2$, is

$$P(\{n_1, n_2\}|N) = \frac{[n_2 \le N]}{\binom{N}{2}}.$$

When considered a function of N for fixed n_2, this is a likelihood function. The maximum likelihood estimate for N is $N_0 = n_2$.

This time the total

$$\sum_{N=1}^{\infty} P(\{n_1, n_2\}|N) = \sum_{N} \frac{[N \ge n_2]}{\binom{N}{2}} = \frac{2}{n_2 - 1}$$

is a convergent series, and so this likelihood function can be normalized into a probability distribution.

If you pick 3 or more tickets, the likelihood function has a well defined mean value, which is larger than the maximum likelihood estimate. If you pick 4 or more tickets, the likelihood function has a well defined standard deviation too.

Likelihoods that eliminate nuisance parameters

In many cases, the likelihood is a function of more than one parameter but interest focuses on the estimation of only one, or at most a few of them, with the others being considered as nuisance parameters. Several alternative approaches have been developed to eliminate such nuisance parameters so that a likelihood can be written as a function of only the parameter (or parameters) of interest; the main approaches being marginal, conditional and profile likelihoods.

These approaches are useful because standard likelihood methods can become unreliable or fail entirely when there are many nuisance parameters or when the nuisance parameters are high-dimensional. This is particularly true when the nuisance parameters can be considered to be "missing data"; they represent a non-negligible fraction of the number of observations and this fraction does not decrease when the sample size increases. Often these approaches can be used to derive closed-form formulae for statistical tests when direct use of maximum likelihood requires iterative numerical methods. These approaches find application in some specialized topics such as sequential analysis.

Conditional likelihood

Sometimes it is possible to find a sufficient statistic for the nuisance parameters, and conditioning on this statistic results in a likelihood which does not depend on the nuisance parameters.

One example occurs in 2×2 tables, where conditioning on all four marginal totals leads to a conditional likelihood based on the non-central hypergeometric distribution. This form of conditioning is also the basis for Fisher's exact test.

Marginal likelihood

Sometimes we can remove the nuisance parameters by considering a likelihood based on only part of the information in the data, for example by using the set of ranks rather than the numerical values. Another example occurs in linear mixed models, where considering a likelihood for the residuals only after fitting the fixed effects leads to residual maximum likelihood estimation of the variance components.

Profile likelihood

It is often possible to write some parameters as functions of other parameters, thereby reducing the number of independent parameters. (The function is the parameter value which maximizes the likelihood given the value of the other parameters.) This procedure is called concentration of the parameters and results in the concentrated likelihood function, also occasionally known as the maximized likelihood function, but most often called the profile likelihood function.

For example, consider a regression analysis model with normally distributed errors. The most likely value of the error variance is the variance of the residuals. The residuals depend on all other parameters. Hence the variance parameter can be written as a function of the other parameters.

Unlike conditional and marginal likelihoods, profile likelihood methods can always be used, even when the profile likelihood cannot be written down explicitly. However, the profile likelihood is not a true likelihood, as it is not based directly on a probability distribution, and this leads to some less satisfactory properties. Attempts have been made to improve this, resulting in modified profile likelihood.

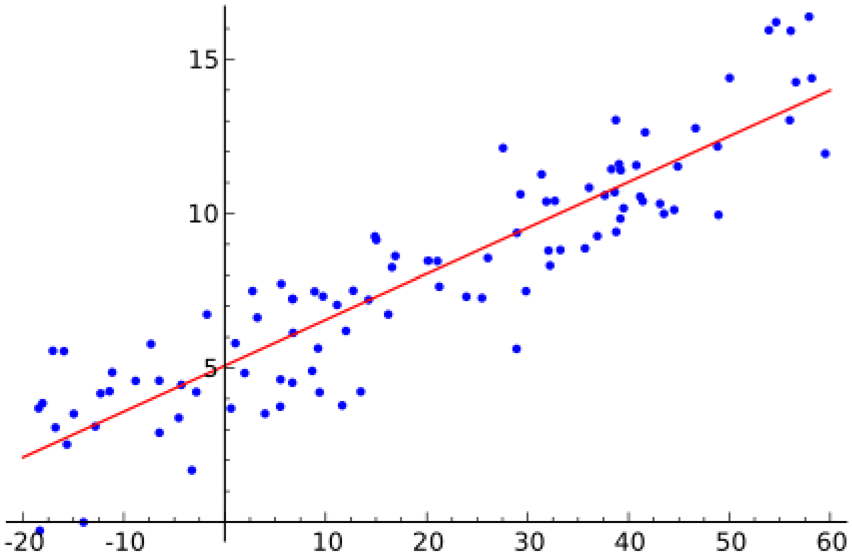
The idea of profile likelihood can also be used to compute confidence intervals that often have better small-sample properties than those based on asymptotic standard errors calculated from the full likelihood. In the case of parameter estimation in partially observed systems, the profile likelihood can be also used for identifiability analysis.

Partial likelihood

A partial likelihood is a factor component of the likelihood function that isolates the parameters of interest. It is a key component of the proportional hazards model.

Chapter 11

Linear Regression



Example of simple linear regression, which has one independent variable.

In statistics, **linear regression** is an approach to modeling the relationship between a scalar variable *y* and one or more variables denoted *X*. In linear regression, data are modeled using linear functions, and unknown model parameters are estimated from the data. Such models are called *linear models*. Most commonly, linear regression refers to a model in which the conditional mean of *y* given the value of *X* is an affine function of *X*. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of *y* given *X* is expressed as a linear function of *X*. Like all forms of regression analysis, *linear regression* focuses on the conditional probability distribution of *y* given *X*, rather than on the joint probability distribution of *y* and *X*, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
- Given a variable *y* and a number of variables X_1, ..., X_p that may be related to *y*, then linear regression analysis can be applied to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*, thus once one of them is known, the others are no longer informative.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the “lack of fit” in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, while the terms “least squares” and *linear model* are closely linked, they are not synonymous.

Introduction to linear regression

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the *p*-vector of regressors x_i is linear. This relationship is modeled through a so-called “disturbance term” ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where ′ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors x_i and β.

Often these *n* equations are stacked together and written in vector form as

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Some remarks on terminology and general use:

- $y_i$is called the *regressand*, *endogenous variable*, *response variable*, *measured variable*, or *dependent variable*. The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables. Alternatively, there may be an operational reason to model one of the variables in terms of the others, in which case there need be no presumption of causality.
- $x_i$are called *regressors*, *exogenous variables*, *explanatory variables*, *covariates*, *input variables*, *predictor variables*, or *independent variables*. The matrix *X* is sometimes called the design matrix.
    - Usually a constant is included as one of the regressors. For example we can take $x_{i1} = 1$ for $i = 1, ..., n$. The corresponding element of β is called the *intercept*. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.
    - Sometimes one of the regressors can be a non-linear function of another regressor or of the data, as in polynomial regression and segmented regression. The model remains linear as long as it is linear in the parameter vector β.
    - The regressors x_i may be viewed either as random variables, which we simply observe, or they can be considered as predetermined fixed values which we can choose. Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.
- $\beta$is a *p*-dimensional *parameter vector*. Its elements are also called *effects*, or *regression coefficients*. Statistical estimation and inference in linear regression focuses on β.
- $\varepsilon_i$is called the *error term*, *disturbance term*, or *noise*. This variable captures all other factors which influence the dependent variable y_i other than the regressors x_i. The relationship between the error term and the regressors, for example whether they are correlated, is a crucial step in formulating a linear regression model, as it will determine the method to use for estimation.

Example. Consider a situation where a small ball is being tossed up in the air and then we measure its heights of ascent h_i at various moments in time t_i. Physics tells us that, ignoring the drag, the relationship can be modeled as

$$h_i = \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i,$$

where β_1 determines the initial velocity of the ball, β_2 is proportional to the standard gravity, and ε_i is due to measurement errors. Linear regression can be used to estimate the values of β_1 and β_2 from the measured data. This model is non-linear in the time variable, but it is linear in the parameters β_1 and β_2; if we take regressors $x_i = (x_{i1}, x_{i2}) = (t_i, t_i^2)$, the model takes on the standard form

$$h_i = x_i'\beta + \varepsilon_i.$$

Assumptions

Two key assumptions are common to all estimation methods used in linear regression analysis:

- The design matrix X must have full column rank p. For this property to hold, we must have $n > p$, where n is the sample size (this is a necessary but not a sufficient condition). If this condition fails this is called the multicollinearity in the regressors. In this case the parameter vector β will be not identifiable — at most we will be able to narrow down its value to some linear subspace of $\mathbf{R}^p$. Methods for fitting linear models with multicollinearity have been developed, but require additional assumptions such as “effect sparsity” — that a large fraction of the effects are exactly zero.

  A simpler statement of this is that there must be enough data available compared to the number of parameters to be estimated. If there is too little data, then you end up with a system of equations with no unique solution.

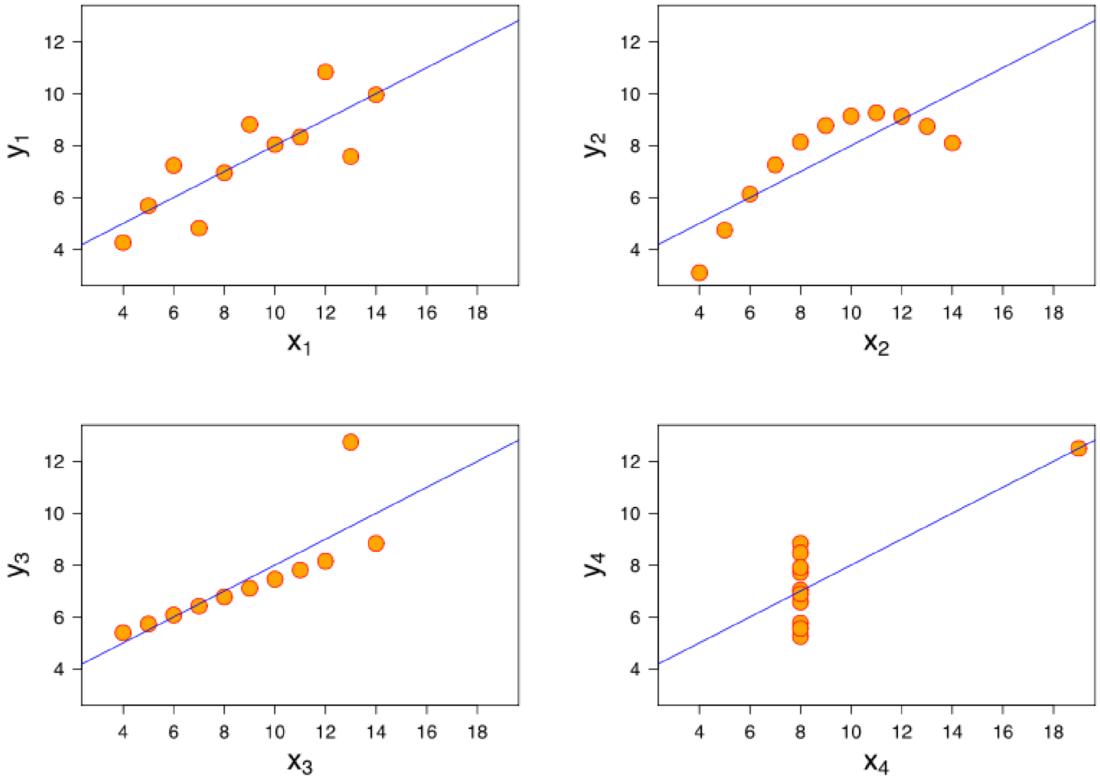
- The regressors x_i are assumed to be error-free, that is they are not contaminated with measurement errors. Although not realistic in many settings, dropping this assumption leads to significantly more difficult errors-in-variables models.

Beyond these two assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

- Some estimation methods are based on a lack of correlation, among the n observations $(y_i, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n$. Statistical independence of the observations is not needed, although it can be exploited if it is known to hold.
- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.

- The variances of the error terms may be equal across the *n* units (termed *homoscedasticity*) or not (termed *heteroscedasticity*). Some linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present.
- The arrangement, or probability distribution of the predictor variables *x* has a major influence on the precision of estimates of β. Sampling and design of experiments are highly-developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of β.

Interpretation



The sets in the Anscombe's quartet have the same linear regression line but are themselves very different.

A fitted linear regression model can be used to identify the relationship between a single predictor variable x_j and the response variable *y* when all the other predictor variables in the model are "held fixed". Specifically, the interpretation of β_j is the expected change in *y* for a one-unit change in x_j when the other covariates are held fixed. This is sometimes called the *unique effect* of x_j on *y*. In contrast, the *marginal effect* of x_j on *y* can be assessed using a correlation coefficient or simple linear regression model relating x_j to *y*.

Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes (such as dummy variables, or the intercept term), while others cannot be held fixed (recall the example from the introduction: it would be impossible to "hold t_i fixed" and at the same time change the value of t_i^2).

It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in x_j, so that once that variable is in the model, there is no contribution of x_j to the variation in y. Conversely, the unique effect of x_j can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by x_j. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to x_j, thereby strengthening the apparent relationship with x_j.

The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study.

The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design.

Estimation methods

Numerous procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency.

Some of the more common estimation techniques for linear regression are summarized below.

- **Ordinary least squares** (OLS) is the simplest and thus very common estimator. It is conceptually simple and computationally straightforward. OLS estimates are commonly used to analyze both experimental and observational data.

The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter *β*:

$$\hat{\beta} = (X'X)^{-1}X'y = \big(\tfrac{1}{n}{\textstyle\sum} x_i x_i'\big)^{-1}\big(\tfrac{1}{n}{\textstyle\sum} x_i y_i\big)$$

The estimator is unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors

$$\mathrm{E}[\,x_i\varepsilon_i\,] = 0.$$

It is also efficient under the assumption that the errors have finite variance and are homoscedastic, meaning that $\mathrm{E}[\varepsilon_i^2|x_i]$ does not depend on *i*. The condition that the

errors are uncorrelated with the regressors will generally be satisfied in an experiment, but in the case of observational data, it is difficult to exclude the possibility of an omitted covariate *z* that is related to both the observed covariates and the response variable. The existence of such a covariate will generally lead to a correlation between the regressors and the response variable, and hence to an inconsistent estimator of *β*. The condition of homoscedasticity can fail with either experimental or observational data. If the goal is either inference or predictive modeling, the performance of OLS estimates can be poor if multicollinearity is present, unless the sample size is large.
In simple linear regression, where there is only one regressor (with a constant), the OLS coefficient estimates have a simple form that is closely related to the correlation coefficient between the covariate and the response.

- **Generalized least squares** (GLS) is an extension of the OLS method, that allows efficient estimation of *β* when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data. To handle heteroscedasticity when the error terms are uncorrelated with each other, GLS minimizes a weighted analogue to the sum of squared residuals from OLS regression, where the weight for the *i*th case is inversely proportional to var(ε_i). This special case of GLS is called “weighted least squares”. The GLS solution to estimation problem is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y,$$

where Ω is the covariance matrix of the errors. GLS can be viewed as applying a

linear transformation to the data so that the assumptions of OLS are met for the transformed data. For GLS to be applied, the covariance structure of the errors must be known up to a multiplicative constant.

- **Iteratively reweighted least squares** (IRLS) is used when heteroscedasticity, or correlations, or both are present among the error terms of the model, but where little is known about the covariance structure of the errors independently of the data. In the first iteration, OLS, or GLS with a provisional covariance structure is carried out, and the residuals are obtained from the fit. Based on the residuals, an improved estimate of the covariance structure of the errors can usually be obtained. A subsequent GLS iteration is then performed using this estimate of the error structure to define the weights. The process can be iterated to convergence, but in many cases, only one iteration is sufficient to achieve an efficient estimate of β.
- **Instrumental variables** regression (IV) can be performed when the regressors are correlated with the errors. In this case, we need the existence of some auxiliary *instrumental variables* $\mathbf{z}_i$ such that $E[z_i\varepsilon_i] = 0$. If Z is the matrix of instruments, then the estimator can be given in closed form as

$$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

- **Optimal instruments** regression is an extension of classical IV regression to the situation where $E[\varepsilon_i|z_i] = 0$.
- **Least absolute deviation** (LAD) regression is a robust estimation technique in that it is less sensitive to the presence of outliers than OLS (but is less efficient than OLS when no outliers are present). It is equivalent to maximum likelihood estimation under a Laplace distribution model for ε.
- **Quantile regression** focuses on the conditional quantiles of y given X rather than the conditional mean of y given X. Linear quantile regression models a particular conditional quantile, often the conditional median, as a linear function $\beta'x$ of the predictors.
- **Maximum likelihood estimation** can be performed when the distribution of the error terms is known to belong to a certain parametric family f_θ of probability distributions. When f_θ is a normal distribution with mean zero and variance θ, the resulting estimate is identical to the OLS estimate. GLS estimates are maximum likelihood estimates when ε follows a multivariate normal distribution with a known covariance matrix.
- **Adaptive estimation**. If we assume that error terms are independent from the regressors $\varepsilon_i \perp \mathbf{x}_i$, the optimal estimator is the 2-step MLE, where the first step is used to non-parametrically estimate the distribution of the error term.
- Mixed models are widely used to analyze linear regression relationships involving dependent data when the dependencies have a known structure. Common applications of mixed models include analysis of data involving repeated measurements, such as longitudinal data, or data obtained from cluster sampling. They are generally fit as parametric models, using maximum likelihood or Bayesian estimation. In the case where the errors are modeled as normal random variables, there is a close connection between mixed models and generalized least squares. Fixed effects estimation is an alternative approach to analyzing this type of data.

- Principal component regression (PCR) is used when the number of predictor variables is large, or when strong correlations exist among the predictor variables. This two-stage procedure first reduces the predictor variables using principal component analysis then uses the reduced variables in an OLS regression fit. While it often works well in practice, there is no general theoretical reason that the most informative linear function of the predictor variables should lie among the dominant principal components of the multivariate distribution of the predictor variables. The partial least squares regression is the extension of the PCR method which does not suffer from the mentioned deficiency.
- Total least squares (TLS) is an approach to least squares estimation of the linear regression model that treats the covariates and response variable in a more geometrically symmetric manner than OLS. It is one approach to handling the "errors in variables" problem, and is sometimes used when the covariates are assumed to be error-free..
- Ridge regression, and other forms of penalized estimation such as the Lasso, deliberately introduce bias into the estimation of β in order to reduce the variability of the estimate. The resulting estimators generally have lower mean squared error than the OLS estimates, particularly when multicollinearity is present. They are generally used when the goal is to predict the value of the response variable y for values of the predictors x that have not yet been observed. These methods are not as commonly used when the goal is inference, since it is difficult to account for the bias.
- *Least angle regression* is an estimation procedure for linear regression models that was developed to handle high-dimensional covariate vectors, potentially with more covariates than observations.
- Other robust estimation techniques, including the α-trimmed mean approach, and L-, M-, S-, and R-estimators have been introduced.

Further discussion

In statistics, the problem of **numerical methods for linear least squares** is an important one because linear regression models are one of the most important types of model, both as formal statistical models and for exploration of data sets. The majority of statistical computer packages contain facilities for regression analysis that make use of linear least squares computations. Hence it is appropriate that considerable effort has been devoted to the task of ensuring that these computations are undertaken efficiently and with due regard to numerical precision.

Individual statistical analyses are seldom undertaken in isolation, but rather are part of a sequence of investigatory steps. Some of the topics involved in considering numerical methods for linear least squares relate to this point. Thus important topics can be

- Computations where a number of similar, and often nested, models are considered for the same data set. That is, where models with the same dependent variable but different sets of independent variables are to be considered, for essentially the same set of data points.

- Computations for analyses that occur in a sequence, as the number of data points increases.
- Special considerations for very extensive data sets.

Fitting of linear models by least squares often, but not always, arises in the context of statistical analysis. It can therefore be important that considerations of computational efficiency for such problems extend to all of the auxiliary quantities required for such analyses, and are not restricted to the formal solution of the linear least squares problem.

Matrix calculations, like any others, are affected by rounding errors. An early summary of these effects, regarding the choice of computational methods for matrix inversion, was provided by Wilkinson.

Extensions

- General linear model considers the situation when the response variable y is not a scalar but a vector. Conditional linearity of $E(y|x) = Bx$ is still assumed, with a matrix B replacing the vector β of the classical linear regression model. Multivariate analogues of OLS and GLS have been developed.
- Generalized linear models are a framework for modeling a response variable y in the form $g(\beta'x) + \varepsilon$, where g is an arbitrary *link function*. Single index models allow some degree of nonlinearity in the relationship between x and y, while preserving the central role of the linear predictor $\beta'x$ as in the classical linear regression model. Under certain conditions, simply applying OLS to data from a single-index model will consistently estimate β up to a proportionality constant .
- Hierarchical linear models (or *multilevel regression*) organizes the data into a hierarchy of regressions, for example where A is regressed on B, and B is regressed on C. It is often used where the data have a natural hierarchical structure such as in educational statistics, where students are nested in classrooms, classrooms are nested in schools, and schools are nested in some administrative grouping such as a school district. The response variable might be a measure of student achievement such as a test score, and different covariates would be collected at the classroom, school, and school district levels.
- Errors-in-variables models (or “measurement error models”) extend the traditional linear regression model to allow the predictor variables X to be observed with error. This error causes standard estimators of β to become biased. Generally, the form of bias is an attenuation, meaning that the effects are biased toward zero.
- In Dempster–Shafer theory, or a linear belief function in particular, a linear regression model may be represented as a partially swept matrix, which can be combined with similar matrices representing observations and other assumed normal distributions and state equations. The combination of swept or unswept matrices provides an alternative method for estimating linear regression models.

Applications of linear regression

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. It ranks as one of the most important tools used in these disciplines.

Trend line

A **trend line** represents a trend, the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set (say GDP, oil prices or stock prices) have increased or decreased over the period of time. A trend line could simply be drawn by eye through a set of data points, but more properly their position and slope is calculated using statistical techniques like linear regression. Trend lines typically are straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

Trend lines are sometimes used in business analytics to show changes in data over time. This has the advantage of being simple. Trend lines are often used to argue that a particular action or event (such as training, or an advertising campaign) caused observed changes at a point in time. This is a simple technique, and does not require a control group, experimental design, or a sophisticated analysis technique. However, it suffers from a lack of scientific validity in cases where other potential changes can affect the data.

Epidemiology

Early evidence relating tobacco smoking to mortality and morbidity came from observational studies employing regression analysis. In order to reduce spurious correlations when analyzing observational data, researchers usually include several variables in their regression models in addition to the variable of primary interest. For example, suppose we have a regression model in which cigarette smoking is the independent variable of interest, and the dependent variable is lifespan measured in years. Researchers might include socio-economic status as an additional independent variable, to ensure that any observed effect of smoking on lifespan is not due to some effect of education or income. However, it is never possible to include all possible confounding variables in an empirical analysis. For example, a hypothetical gene might increase mortality and also cause people to smoke more. For this reason, randomized controlled trials are often able to generate more compelling evidence of causal relationships than can be obtained using regression analyses of observational data. When controlled experiments are not feasible, variants of regression analysis such as instrumental variables regression may be used to attempt to estimate causal relationships from observational data.

Finance

The capital asset pricing model uses linear regression as well as the concept of Beta for analyzing and quantifying the systematic risk of an investment. This comes directly from the Beta coefficient of the linear regression model that relates the return on the investment to the return on all risky assets.

Economics

Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.

Environmental science

Linear regression finds application in a wide range of environmental science applications. In Canada, the Environmental Effects Monitoring Program uses statistical analyses on fish and benthic surveys to measure the effects of pulp mill or metal mine effluent on the aquatic ecosystem.

Chapter 12

Location Estimation in Sensor Networks

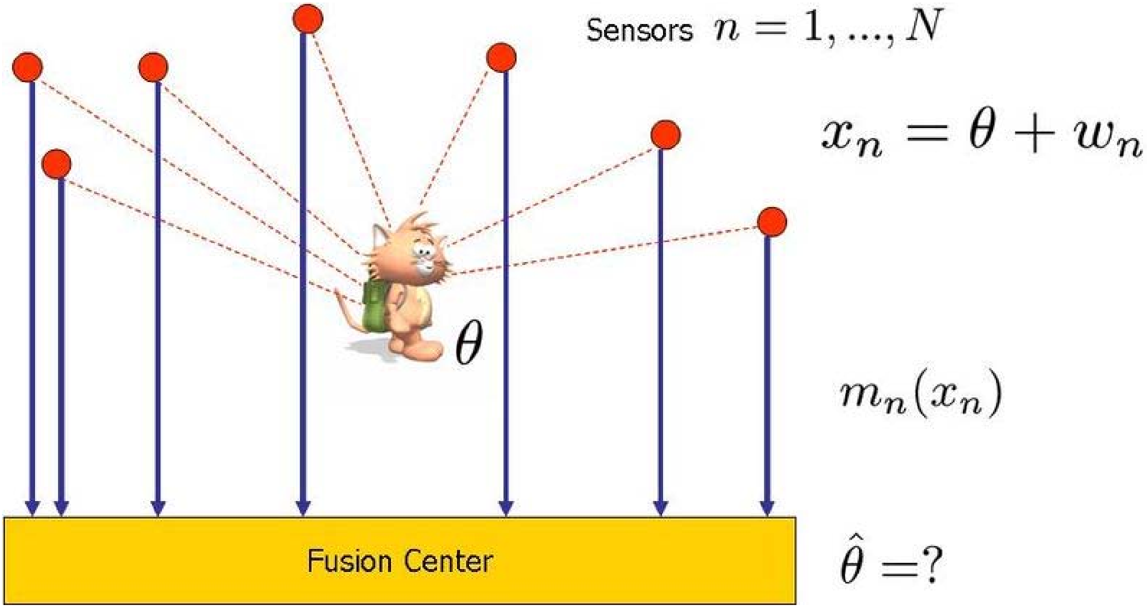
Location estimation in **wireless sensor networks** is the problem of estimating the location of an object from a set of noisy measurements, when the measurements are acquired in a distributed manner by a set of sensors.

Motivation

Many civilian and military applications require monitoring that can identify objects in a specific area, such as monitoring the front entrance of a private house by a single camera. Monitored areas that are large relative to objects of interest often require multiple sensors (e.g., infra-red detectors) at multiple locations. A centralized observer or computer application monitors the sensors. The communication to Power and bandwidth requirements call for efficient design of the sensor, transmission, and processing.

The *CodeBlue system* of Harvard university is an example where a vast number of sensors distributed among hospital facilities allow staff to locate a patient in distress. In addition, the sensor array enables online recording of medical information while allowing the patient to move around. Military applications (e.g. locating an intruder into a secured area) are also good candidates for setting a wireless sensor network.

Setting



Let θ denote the position of interest. A set of *N* sensors acquire measurements $x_n = \theta + w_n$ contaminated by an additive noise w_n owing some known or unknown probability density function (PDF). The sensors transmit measurements to a central processor. The *n*th sensor encodes x_n by a function $m_n(x_n)$. The application processing the data applies a pre-defined estimation rule $\hat{\theta} = f(m_1(x_1), \cdot, m_N(x_N))$. The set of message functions $m_n,\ 1 \leq n \leq N$ and the fusion rule $f(m_1(x_1), \cdot, m_N(x_N))$ are designed to minimize estimation error. For example: minimizing the mean squared error (MSE), $\mathbb{E}\|\theta - \hat{\theta}\|^2$.

Ideally, sensors transmit their measurements x_n exactly to the processing center, that is $m_n(x_n) = x_n$. In this settings, the maximum likelihood estimator (MLE) $\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n$ is an unbiased estimator whose MSE is $\mathbb{E}\|\theta - \hat{\theta}\|^2 = \text{var}(\hat{\theta}) = \frac{\sigma^2}{N}$ assuming a white Gaussian noise $w_n \sim \mathcal{N}(0, \sigma^2)$. The next sections suggest alternative designs when the sensors are bandwidth constrained to 1 bit transmission, that is $m_n(x_n)$=0 or 1.

Known noise PDF

We begin with an example of a Gaussian noise $w_n \sim \mathcal{N}(0, \sigma^2)$, in which a suggestion for a system design is as follows :

$$m_n(x_n) = I(x_n - \tau) = \begin{cases} 1 & x_n > \tau \\ 0 & x_n \leq \tau \end{cases}$$

$$\hat{\theta} = \tau - F^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} m_n(x_n)\right), \quad F(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{x}^{\infty} e^{-w^2/2\sigma^2}\,dw$$

Here τ is a parameter leveraging our prior knowledge of the approximate location of θ. In this design, the random value of $m_n(x_n)$ is distributed Bernoulli~($q = F(\tau - \theta)$). The processing center averages the received bits to form an estimate $\hat{q}$of q, which is then used to find an estimate of θ. It can be verified that for the optimal (and infeasible) choice of τ = θ the variance of this estimator is $\frac{\pi\sigma^2}{4}$ which is only π / 2 times the variance of MLE without bandwidth constraint. The variance increases as τ deviates from the real value of θ, but it can be shown that as long as | τ − θ | ˜σ the factor in the MSE remains approximately 2. Choosing a suitable value for τ is a major disadvantage of this method since our model does not assume prior knowledge about the approximated location of θ. A coarse estimation can be used to overcome this limitation. However, it requires additional hardware in each of the sensors.

A system design with arbitrary (but known) noise PDF can be found in . In this setting it is assumed that both θ and the noise w_n are confined to some known interval $[-U,U]$. The estimator of also reaches an MSE which is a constant factor times $\frac{\sigma^2}{N}$. In this method, the prior knowledge of U replaces the parameter τ of the previous approach.

Unknown noise parameters

A noise model may be sometimes available while the exact PDF parameters are unknown (e.g. a Gaussian PDF with unknown σ). The idea proposed in for this setting is to use two thresholds τ_1,τ_2, such that $N / 2$ sensors are designed with $m_A(x) = I(x - \tau_1)$, and the other $N / 2$ sensors use $m_B(x) = I(x - \tau_2)$. The processing center estimation rule is generated as follows:

$$\hat{q}_1 = \frac{2}{N}\sum_{n=1}^{N/2} m_A(x_n), \quad \hat{q}_2 = \frac{2}{N}\sum_{n=1+N/2}^{N} m_B(x_n)$$

$$\hat{\theta} = \frac{F^{-1}(\hat{q}_2)\tau_1 - F^{-1}(\hat{q}_1)\tau_2}{F^{-1}(\hat{q}_2) - F^{-1}(\hat{q}_1)}, \quad F(x) = \frac{1}{\sqrt{2\pi}}\int_{x}^{\infty} e^{-v^2/2}dw$$

As before, prior knowledge is necessary to set values for τ_1,τ_2 to have an MSE with a reasonable factor of the unconstrained MLE variance.

Unknown noise PDF

We now describe the system design of for the case that the structure of the noise PDF is unknown. The following model is considered for this scenario:

$$\begin{aligned}&x_n = \theta + w_n, \quad n = 1, \dots, N\\&\theta \in [-U, U]\\&w_n \in \mathcal{P}, \text{ that is } : w_n \text{ is bounded to } [-U, U], \mathbb{E}(w_n) = 0\end{aligned}$$

In addition, the message functions are limited to have the form

$$m_n(x_n) = \begin{cases} 1 & x \in S_n \\ 0 & x \notin S_n \end{cases}$$

where each S_n is a subset of $[-2U, 2U]$. The fusion estimator is also restricted to be linear, i.e. $\hat{\theta} = \sum_{n=1}^{N} \alpha_n m_n(x_n)$.

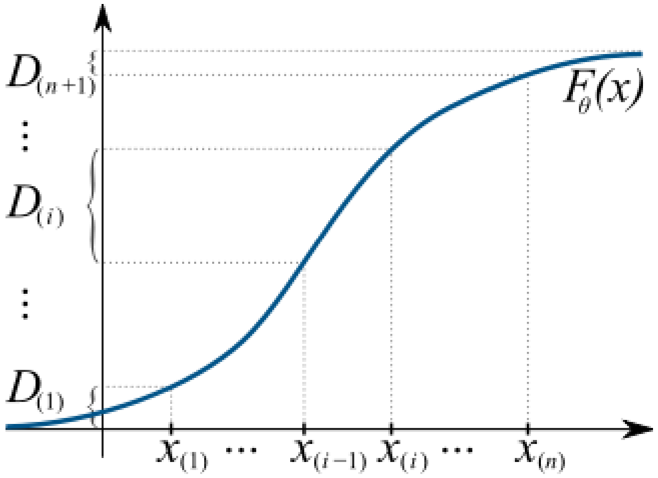
The design should set the decision intervals S_n and the coefficients α_n. Intuitively, we would allocate $N / 2$ sensors to encode the first bit of θ by setting their decision interval to be $[0,2U]$, then $N / 4$ sensors would encode the second bit by setting their decision interval to $[-U, 0] \cup [U, 2U]$and so on. It can be shown that these decision intervals and the corresponding set of coefficients α_n produce a universal δ-unbiased estimator, which is an estimator satisfying $|\mathbb{E}(\theta - \hat{\theta})| < \delta$for every possible value of $\theta \in [-U, U]$and for every realization of $w_n \in \mathcal{P}$. In fact, this intuitive design of the decision intervals is also optimal in the following sense. The above design requires $N \geq \lceil \log \frac{8U}{\delta} \rceil$ to satisfy the universal δ-unbiased property while theoretical arguments show that an optimal (and a more complex) design of the decision intervals would require $N \geq \lceil \log \frac{2U}{\delta} \rceil$, that is: the number of sensors is nearly optimal. It is also argued in that if the targeted MSE $\mathbb{E}\|\theta - \hat{\theta}\| \leq \epsilon^2$uses a small enough ε, then this design requires a factor of 4 in the number of sensors to achieve the same variance of the MLE in the unconstrained bandwidth settings.

Additional information

The design of the sensor array requires optimizing the power allocation as well as minimizing the communication traffic of the entire system. The design suggested in incorporates probabilistic quantization in sensors and a simple optimization program that is solved in the fusion center only once. The fusion center then broadcasts a set of parameters to the sensors that allows them to finalize their design of messaging functions $m_n(\cdot)$as to meet the energy constraints. Another work employs a similar approach to address distributed detection in wireless sensor arrays.

Chapter 13

Maximum Spacing Estimation



The maximum spacing method tries to find a distribution function such that that the spacings, $D_{(i)}$, are all approximately of the same length. This is done by maximizing their geometric mean.

In statistics, **maximum spacing estimation** (**MSE** or **MSP**), or **maximum product of spacing estimation (MPS)**, is a method for estimating the parameters of a univariate statistical model. The method requires maximization of the geometric mean of *spacings* in the data, which are the differences between the cdf values of the closest data points.

The concept underlying the method is based on the probability integral transform, in that a set of independent random samples derived from any random variable should on average be uniformly distributed with respect to the cumulative distribution function of the random variable. The MPS method chooses the parameter values that make the observed data as uniform as possible, according to a specific quantitative measure of uniformity.

One of the most common methods for estimating the parameters of a distribution from data, the method of maximum likelihood (MLE), can break down in various cases, such as involving certain mixtures of continuous distributions. In these cases the method of maximum spacing estimation may be successful.

Apart from its use in pure mathematics and statistics, the method has found applications in such fields as hydrology, econometrics, and others.

History and usage

The MSE method was derived independently by Russel Cheng and Nik Amin at the University of Wales Institute of Science and Technology, and Bo Ranneby at the Swedish University of Agricultural Sciences. The authors explained that due to the probability integral transform at the true parameter, the “spacing” between each observation should be uniformly distributed. This would imply that the difference between the values of the cumulative distribution function at consecutive observations should be equal. This is the case that maximizes the geometric mean of such spacings, so solving for the parameters that maximize the geometric mean would achieve the “best” fit as defined this way. Ranneby (1984) justified the method by demonstrating that it is an estimator of the Kullback–Leibler divergence, similar to maximum likelihood estimation, but with more robust properties for various classes of problems.

There are certain distributions, especially those with three or more parameters, whose likelihoods may become infinite along certain paths in the parameter space. Using maximum likelihood to estimate these parameters often breaks down, with one parameter tending to the specific value that causes the likelihood to be infinite, rendering the other parameters inconsistent. The method of maximum spacings, however, being dependent on the difference between points on the cumulative distribution function and not individual likelihood points, does not have this issue, and will return valid results over a much wider array of distributions.

The distributions that tend to have likelihood issues are often those used to model physical phenomena. Hall & al. (2004) seek to analyze flood alleviation methods, which requires accurate models of river flood effects. The distributions that better model these effects are all three-parameter models, which suffer from the infinite likelihood issue described above, leading to Hall’s investigation of the maximum spacing procedure. Wong & Li (2006), when comparing the method to maximum likelihood, use various data sets ranging from a set on the oldest ages at death in Sweden between 1905 and 1958 to a set containing annual maximum wind speeds.

Definition

Given an iid random sample $\{x_1, \ldots, x_n\}$ of size n from a univariate distribution with cdf $F(x;\theta_0)$, where $\theta_0 \in \Theta$ is an unknown parameter to be estimated, let $\{x_{(1)}, \ldots, x_{(n)}\}$ be the corresponding ordered sample, that is the result of sorting of all observations from smallest to largest. For convenience also denote $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$.

Define the *spacings* as the “gaps” between the values of the distribution function at adjacent ordered points:

$$D_i(\theta) = F(x_{(i)};\,\theta) - F(x_{(i-1)};\,\theta), \quad i = 1,\ldots,n+1.$$

Then the **maximum spacing estimator** of θ_0 is defined as a value that maximizes the logarithm of the geometric mean of sample spacings:

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\max}\ S_n(\theta), \quad \text{where } S_n(\theta) = \ln \sqrt[n+1]{D_1 D_2 \cdots D_{n+1}} = \frac{1}{n+1}\sum_{i=1}^{n+1} \ln D_i(\theta).$$

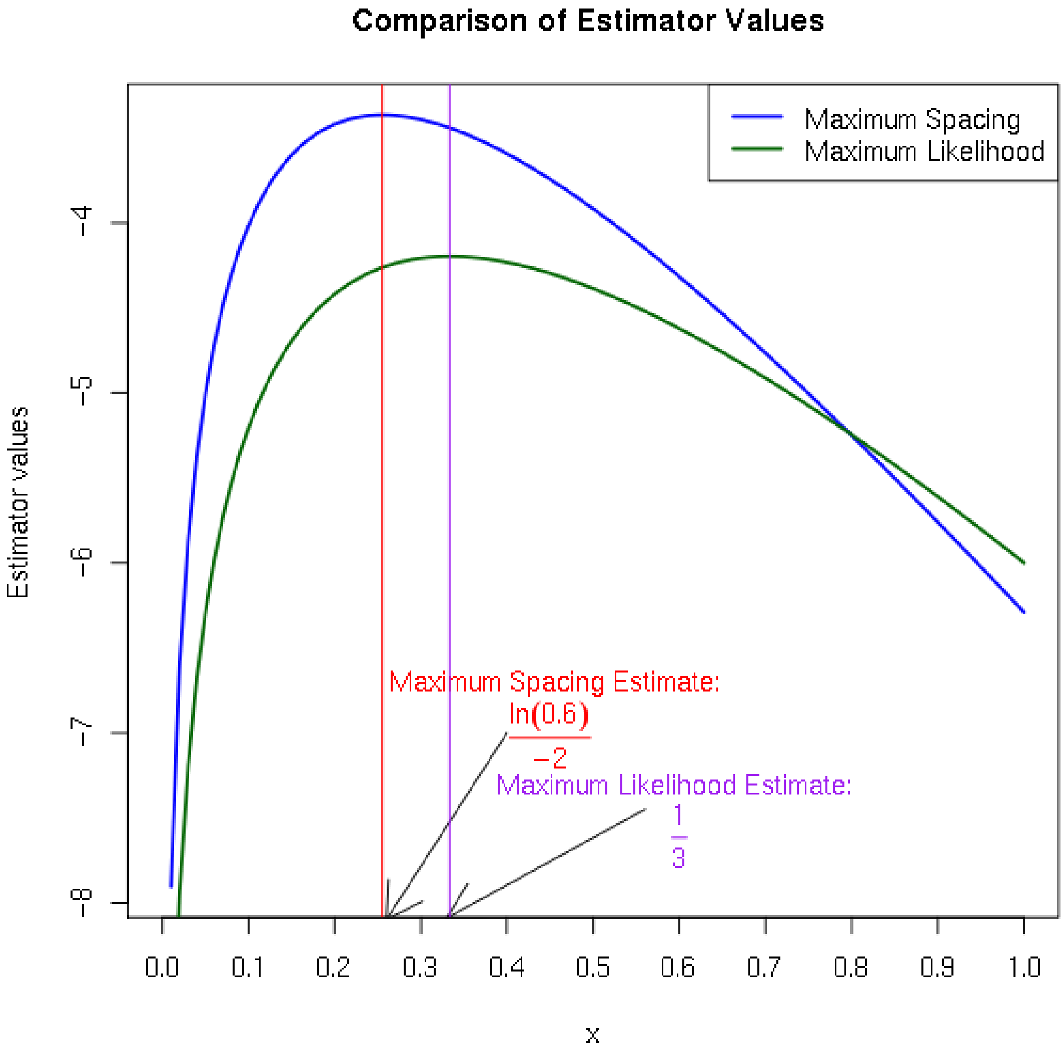
By the inequality of arithmetic and geometric means, function $S_n(\theta)$ is bounded from above by $-\ln(n+1)$, and thus the maximum has to exist at least in the supremum sense.

Note that some authors define the function $S_n(\theta)$ somewhat differently. In particular, Ranneby (1984) multiplies each D_i by a factor of $(n+1)$, whereas Cheng & Stephens (1989) omit the $^1/_{n+1}$ factor in front of the sum and add the “–” sign in order to turn the maximization into minimization. As these are constants with respect to θ, the modifications do not alter the location of the maximum of the function S_n.

Examples

This section presents two examples of calculating the maximum spacing estimator.

Example 1



Plots of the log value of λ for the simplistic example under both likelihood and spacing estimation. The values for which both likelihood and spacing are maximized, the maximum likelihood and maximum spacing estimates, are identified.

Suppose two values $x_{(1)} = 2$, $x_{(2)} = 4$ were sampled from the exponential distribution $F(x;\lambda) = 1 - e^{-x\lambda}$, $x \geq 0$ with unknown parameter $\lambda > 0$. In order to construct the MSE we have to first find the spacings:

i	$F(x_{(i)})$	$F(x_{(i-1)})$	$D_i = F(x_{(i)}) - F(x_{(i-1)})$
1	$1 - e^{-2\lambda}$	0	$1 - e^{-2\lambda}$
2	$1 - e^{-4\lambda}$	$1 - e^{-2\lambda}$	$e^{-2\lambda} - e^{-4\lambda}$
3	1	$1 - e^{-4\lambda}$	$e^{-4\lambda}$

The process continues by finding the λ that maximizes the geometric mean of the "difference" column. Using the convention that ignores taking the $(n+1)^{\text{st}}$ root, this turns into the maximization of the following product: $(1 - e^{-2\lambda}) \cdot (e^{-2\lambda} - e^{-4\lambda}) \cdot (e^{-4\lambda})$. Letting $\mu = e^{-2\lambda}$, the problem becomes finding the maximum of $\mu^5 - 2\mu^4 + \mu^3$. Differentiating, the μ has to satisfy $5\mu^4 - 8\mu^3 + 3\mu^2 = 0$. This equation has roots 0, 0.6, and 1. As μ is actually $e^{-2\lambda}$, it has to be greater than zero but less than one. Therefore, the only acceptable solution is

$$\mu = 0.6 \quad \Rightarrow \quad \lambda_{\text{MSE}} = \frac{\ln 0.6}{-2} \approx 0.255,$$

which corresponds to an exponential distribution with a mean of $^1\!/_\lambda \approx 3.915$. For comparison, the maximum likelihood estimate of λ is the inverse of the sample mean, 3, so $\lambda_{\text{MLE}} = ⅓ \approx 0.333$.

Example 2

Suppose $\{x_{(1)}, \ldots, x_{(n)}\}$ is the ordered sample from a uniform distribution $U(a,b)$ with unknown endpoints a and b. The cumulative distribution function is $F(x;a,b) = (x-a) \div (b-a)$ when $x \in [a,b]$. Therefore individual spacings are given by

$$D_1 = \frac{x_{(1)} - a}{b - a}, \quad D_i = \frac{x_{(i)} - x_{(i-1)}}{b - a} \text{ for } i = 2, \ldots, n, \quad D_{n+1} = \frac{b - x_{(n)}}{b - a}$$

Calculating the geometric mean and then taking the logarithm, statistic S_n will be equal to

$$S_n(a,b) = \tfrac{1}{n+1} \ln(x_{(1)} - a) + \tfrac{1}{n+1} \ln(b - x_{(n)}) - \ln(b-a) + \sum_{i=2}^{n} \ln(x_{(i)} - x_{(i-1)})$$

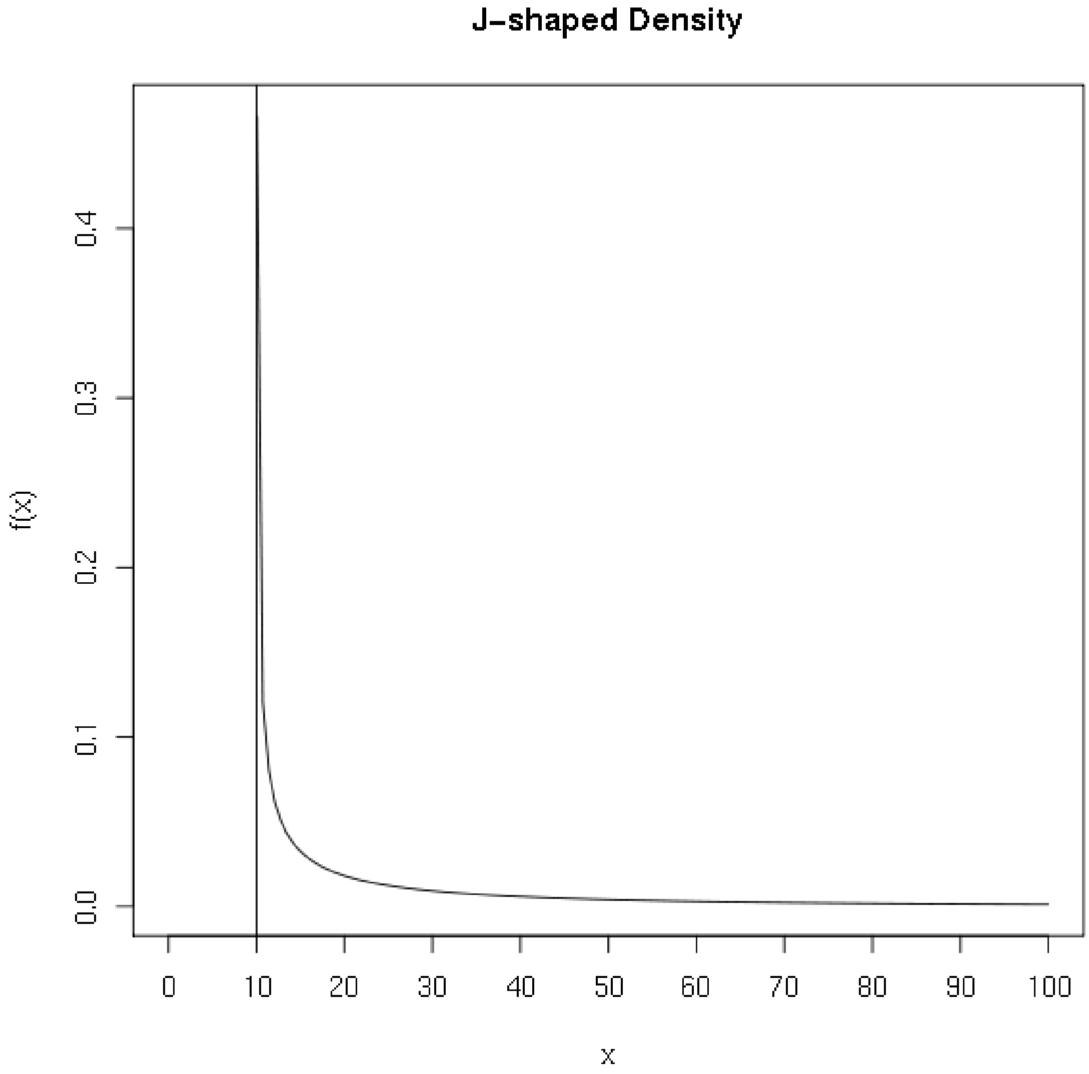
Here only the first three terms depend on the parameters a and b. Differentiating with respect to those parameters and solving the resulting linear system, the maximum spacing estimates will be

$$\hat{a} = \frac{n x_{(1)} - x_{(n)}}{n-1}, \quad \hat{b} = \frac{n x_{(n)} - x_{(1)}}{n-1}.$$

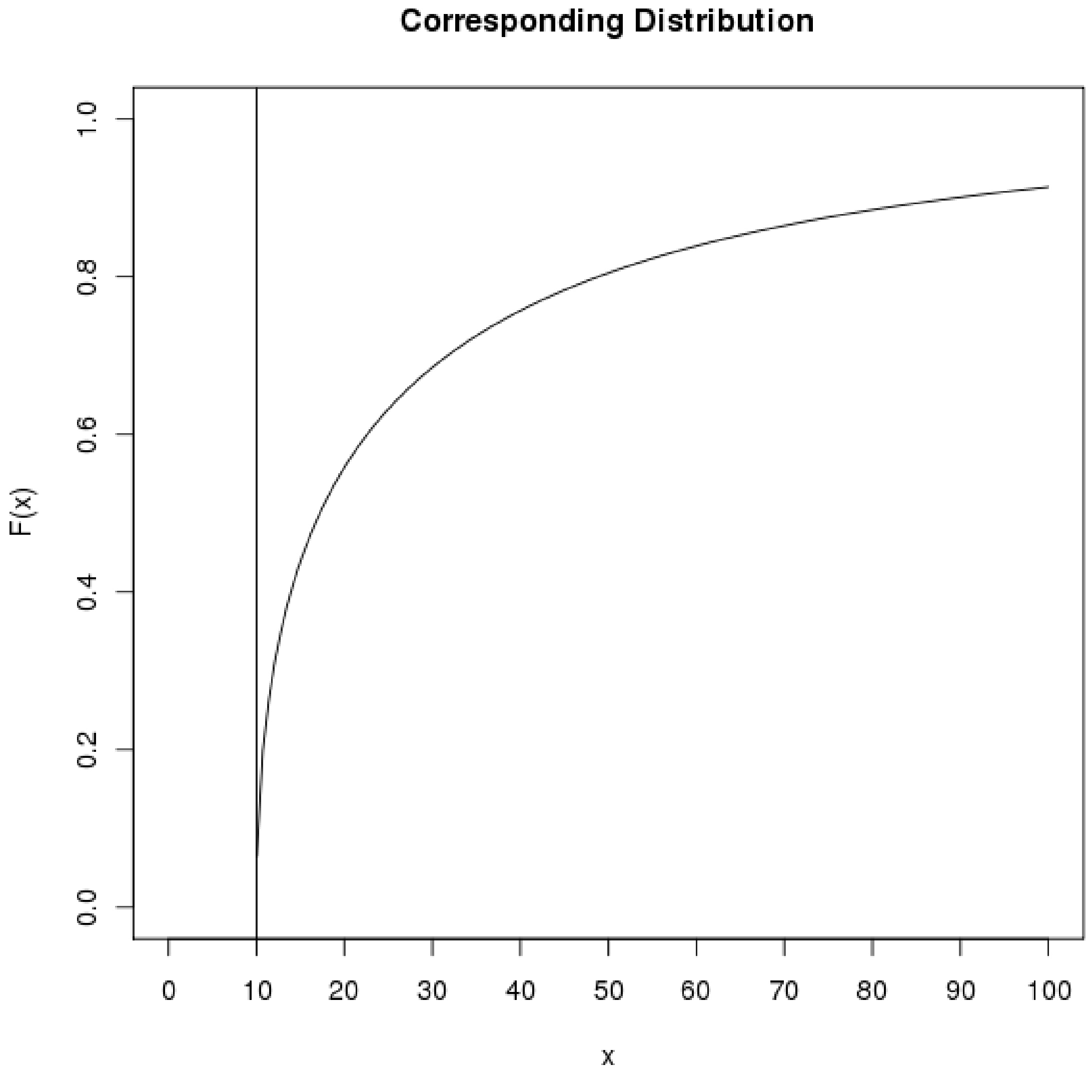
These are known to be the uniformly minimum variance unbiased (UMVU) estimators for the continuous uniform distribution. In comparison, the maximum likelihood estimates for this problem $\hat{a} = x_{(1)}$ and $\hat{b} = x_{(n)}$ are biased and have higher mean-squared error.

Properties

Consistency and efficiency



Density



Distribution

Plot of a “J-shaped” density function and its corresponding distribution. A shifted Weibull with a scale parameter of 15, a shape parameter of 0.5, and a location parameter of 10. The density asymptotically approaches infinity as *x* approaches 10, rendering the estimates of the other parameters inconsistent. Note that there is no inflection point in the graph of the distribution.

The maximum spacing estimator is a consistent estimator in that it converges in probability to the true value of the parameter, θ_0, as the sample size increases to infinity. The consistency of maximum spacing estimation holds under much more general conditions than for maximum likelihood estimators. In particular, in cases where the underlying distribution is J-shaped, maximum likelihood will fail where MSE succeeds. An example of a J-shaped density is the Weibull distribution, specifically a shifted Weibull, with a shape parameter less than 1. The density will tend to infinity as *x* approaches the location parameter rendering estimates of the other parameters inconsistent.

Maximum spacing estimators are also at least as asymptotically efficient as maximum likelihood estimators, where the latter exist. However, MSEs may exist in cases where MLEs do not.

Sensitivity

Maximum spacing estimators are sensitive to closely spaced observations, and especially ties. Given

$$X_{i+k} = X_{i+k-1} = \cdots = X_i,$$

we get

$$D_{i+k}(\theta) = D_{i+k-1}(\theta) = \cdots = D_{i+1}(\theta) = 0.$$

When the ties are due to multiple observations, the repeated spacings (those that would otherwise be zero) should be replaced by the corresponding likelihood. That is, one should substitute $f_i(\theta)$ for $D_i(\theta)$, as

$$\lim_{x_i \to x_{i-1}} (x_i - x_{i-1})^{-1} \int_{x_{i-1}}^{x_i} f(t;\theta)dt = f(x_{i-1},\theta) = f(x_i,\theta),$$

since $x_i = x_{i-1}$.

When ties are due to rounding error, Cheng & Stephens (1989) suggest another method to remove the effects. Given r tied observations from x_i to x_{i+r-1}, let δ represent the round-off error. All of the true values should then fall in the range $x \pm \delta$. The corresponding points on the distribution should now fall between $y_L = F(x-\delta,\hat{\theta})$and $y_U = F(x+\delta,\hat{\theta})$. Cheng and Stephens suggest assuming that the rounded values are uniformly spaced in this interval, by defining

$$D_j = \frac{y_U - y_L}{r-1} \quad (j = i+1,\ldots,i+r-1).$$

The MSE method is also sensitive to secondary clustering. One example of this phenomenon is when a set of observations is thought to come from a single normal distribution, but in fact comes from a mixture normals with different means. A second example is when the data is thought to come from an exponential distribution, but actually comes from a gamma distribution. In the latter case, smaller spacings may occur in the lower tail. A high value of $M(\theta)$ would indicate this secondary clustering effect, and suggesting a closer look at the data is required.

Goodness of fit

The statistic $S_n(\theta)$ is also a form of Moran or Moran-Darling statistic, $M(\theta)$, which can be used to test goodness of fit. It has been shown that the statistic, when defined as

$$S_n(\theta) = M_n(\theta) = -\sum_{j=1}^{n+1} \ln D_j(\theta),$$

is asymptotically normal, and that a chi-squared approximation exists for small samples. In the case where we know the true parameter θ^0, Cheng & Stephens (1989) show that the statistic $M_n(\theta)$has a normal distribution with

$$\mu_M \approx (n+1)(\ln(n+1)+\gamma) - \frac{1}{2} - \frac{1}{12(n+1)},$$
$$\sigma_M^2 \approx (n+1)\left(\frac{\pi^2}{6} - 1\right) - \frac{1}{2} - \frac{1}{6(n+1)},$$

where γ is the Euler–Mascheroni constant which is approximately 0.57722.

The distribution can also be approximated by that of A, where

$$A = C_1 + C_2\chi_{n,}^2$$

in which

$$C_1 = \mu_M - \sqrt{\frac{\sigma_M^2 n}{2}},$$
$$C_2 = \sqrt{\frac{\sigma_M^2}{2n}},$$

and where χ_n^2follows a chi-square distribution with n degrees of freedom. Therefore, to test the hypothesis H_0 that a random sample of n values comes from the distribution $F(x,\theta)$, the statistic $T(\theta) = \frac{M(\theta) - C_1}{C_2}$ can be calculated. Then H_0 should be rejected with significance α if the value is greater than the critical value of the appropriate chi-square distribution.

Where θ_0 is being estimated by $\hat{\theta}$, Cheng & Stephens (1989) showed that $S_n(\hat{\theta}) = M_n(\hat{\theta})$has the same asymptotic mean and variance as in the known case.

However, the test statistic to be used requires the addition of a bias correction term and is:

$$T(\hat{\theta}) = \frac{M(\hat{\theta}) + \frac{k}{2} - C_1}{C_2},$$

where k is the number of parameters in the estimate.

Generalized maximum spacing

Alternate measures and spacings

Ranneby & Ekström (1997) generalized the MSE method to approximate other measures besides the Kullback–Leibler measure. Ekström (1997) further expanded the method to investigate properties of estimators using higher order spacings, where an *m*-order spacing would be defined as $F(X_{j+m}) - F(X_j)$.

Multivariate distributions

Ranneby & al. (2005) discuss extended maximum spacing methods to the multivariate case. As there is no natural order for $\mathbb{R}^k\,(k > 1)$, they discuss two alternative approaches: a geometric approach based on Dirichlet cells and a probabilistic approach based on a “nearest neighbor ball” metric.

Chapter 14

M-Estimator

In statistics, **M-estimators** are a broad class of estimators, which are obtained as the minima of sums of functions of the data. Least-squares estimators and many maximum-likelihood estimators are M-estimators. The definition of M-estimators was motivated by robust statistics, which contributed new types of M-estimators. The statistical procedure of evaluating an M-estimator on a data set is called **M-estimation**.

More generally, an M-estimator may be defined to be a zero of an estimating function. This estimating function is often the derivative of another statistical function: For example, a maximum-likelihood estimate is often defined to be a zero of the derivative of the likelihood function with respect the parameter: thus, a maximum-likelihood estimator is often a critical point of the score function. In many applications, such M-estimators can be thought of as estimating characteristics of the population.

Historical motivation

The method of least squares is a prototypical M-estimator, since the estimator is defined as a minimum of the sum of squares of the residuals.

Another popular M-estimator is maximum-likelihood estimation. For a family of probability density functions *f* parameterized by θ, a maximum likelihood estimator of θ is computed for each set of data by maximizing the likelihood function over the parameter space $\{ \theta \}$. When the observations are independent and identically distributed, a ML-estimate $\hat{\theta}$satisfies

$$\widehat{\theta} = \arg\max_{\theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right)$$

or, equivalently,

$$\widehat{\theta} = \arg\min_{\theta} \left(-\sum_{i=1}^{n} \log \left(f(x_i, \theta) \right) \right).$$

Maximum-likelihood estimators are often inefficient and biased for finite samples. For many regular problems, maximum-likelihood estimation performs well for "large samples", being an approximation of a posterior mode. If the problem is "regular", then any bias of the MLE (or posterior mode) decreases to zero when the sample-size increases to infinity. The performance of maximum-likelihood (and posterior-mode) estimators drops when the parametric family is mis-specified.

Definition

In 1964, Peter Huber proposed generalizing maximum likelihood estimation to the minimization of

$$\sum_{i=1}^{n} \rho(x_i, \theta),$$

where ρ is a function with certain properties (see below). The solutions

$$\hat{\theta} = \arg\min_{\theta} \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right)$$

are called **M-estimators** ("M" for "maximum likelihood-type" (Huber, 1981, page 43)); other types of robust estimator include L-estimators, R-estimators and S-estimators. Maximum likelihood estimators are thus a special case of M-estimators. With suitable rescaling, M-estimators are special cases of extremum estimators (in which more general functions of the observations can be used).

The function ρ, or its derivative, ψ, can be chosen in such a way to provide the estimator desirable properties (in terms of bias and efficiency) when the data are truly from the assumed distribution, and 'not bad' behaviour when the data are generated from a model that is, in some sense, *close* to the assumed distribution.

Types of M-estimators

M-estimators are solutions, θ, which minimize

$$\sum_{i=1}^{n} \rho(x_i, \theta).$$

This minimization can always be done directly. Often it is simpler to differentiate with respect to *θ* and solve for the root of the derivative. When this differentiation is possible, the M-estimator is said to be of **ψ-type**. Otherwise, the M-estimator is said to be of **ρ-type**.

In most practical cases, the M-estimators are of ψ-type.

ρ-type

For positive integer *r*, let $(\mathcal{X}, \Sigma)$and $(\Theta \subset \mathbb{R}^r, S)$be measure spaces. $\theta \in \Theta$is a vector of parameters. An M-estimator of ρ-type *T* is defined through a measurable function $\rho : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$. It maps a probability distribution *F* on $\mathcal{X}$to the value $T(F) \in \Theta$(if it exists) that minimizes $\int_{\mathcal{X}} \rho(x, \theta) dF(x)$:

$$T(F) := \arg\min_{\theta \in \Theta} \int_{\mathcal{X}} \rho(x, \theta) dF(x)$$

For example, for the maximum likelihood estimator, ρ(*x*,θ) = − log(*f*(*x*,θ)), where $f(x, \theta) = \frac{\partial F(x, \theta)}{\partial x}$.

ψ-type

If ρ is differentiable, the computation of $\widehat{\theta}$is usually much easier. An M-estimator of ψ-type *T* is defined through a measurable function $\psi : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^r$. It maps a probability distribution *F* on $\mathcal{X}$to the value $T(F) \in \Theta$(if it exists) that solves the vector equation: $\int_{\mathcal{X}} \psi(x, \theta) dF(x) = 0$

$$\int_{\mathcal{X}} \psi(x, T(F)) dF(x) = 0$$

For example, for the maximum likelihood estimator, $\psi(x,\theta) = \left(\frac{\partial \log(f(x,\theta))}{\partial \theta^1}, \dots, \frac{\partial \log(f(x,\theta))}{\partial \theta^p}\right)^{\mathrm{T}}$, where u^{T} denotes the transpose of vector *u* and $f(x,\theta) = \frac{\partial F(x,\theta)}{\partial x}$.

Such an estimator is not necessarily an M-estimator of ρ-type, but if ρ has a continuous first derivative with respect to θ, then a necessary corresponding M-estimator of ψ-type to be an M-estimator of ρ-type is $\psi(x,\theta) = \nabla_\theta \rho(x,\theta)$. The previous definitions can easily be extended to finite samples.

If the function ψ decreases to zero as $x \rightarrow \pm\infty$, the estimator is called redescending. Such estimators have some additional desirable properties, such as complete rejection of gross outliers.

Computation

For many choices of ρ or ψ, no closed form solution exists and an iterative approach to computation is required. It is possible to use standard function optimization algorithms, such as Newton-Raphson. However, in most cases an iteratively re-weighted least squares fitting algorithm can be performed; this is typically the preferred method.

For some choices of ψ, specifically, *redescending* functions, the solution may not be unique. The issue is particularly relevant in multivariate and regression problems. Thus, some care is needed to ensure that good starting points are chosen. Robust starting points, such as the median as an estimate of location and the median absolute deviation as a univariate estimate of scale, are common.

Properties

Distribution

It can be shown that M-estimators are asymptotically normally distributed. As such, Wald-type approaches to constructing confidence intervals and hypothesis tests can be used. However, since the theory is asymptotic, it will frequently be sensible to check the distribution, perhaps by examining the permutation or bootstrap distribution.

Influence function

The influence function of an M-estimator of ψ-type is proportional to its defining ψ function.

Let *T* be an M-estimator of ψ-type, and *G* be a probability distribution for which *T*(*G*) is defined. Its influence function IF is

$$\mathrm{IF}(x;T,G) = -\frac{\psi(x,T(G))}{\int\left[\frac{\partial\psi(y,\theta)}{\partial\theta}\right]\mathrm{d}y}$$

A proof of this property of M-estimators can be found in Huber (1981, Section 3.2).

Applications

M-estimators can be constructed for location parameters and scale parameters in univariate and multivariate settings, as well as being used in robust regression .

Examples

Mean

Let (X_1, ... , X_n) be a set of independent, identically distributed random variables, with distribution *F*.

If we define

$$\rho(x,\theta) = \frac{(x-\theta)^2}{2},$$

we note that this is minimized when θ is the mean of the *X*s. Thus the mean is an M-estimator of ρ-type, with this ρ function.

As this ρ function is continuously differentiable in θ, the mean is thus also an M-estimator of ψ-type for $\psi(x, \theta) = \theta - x$.