Z-estimators (generalized method of moments)

Consider the estimation of an unknown parameter θ in a set Θ, based on data $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Each function $h(x, \cdot)$ on Θ defines a ***Z-estimator*** $\widehat{\theta}_n = \widehat{\theta}_n(x_1, \ldots, x_n)$ as a zero of a random ***criterion function***

$$H_n(\theta) := H_n(\theta, \mathbf{x}) := \frac{1}{n}\sum\nolimits_{i \le n} h(x_i, \theta).$$

That is, $\widehat{\theta}_n$ is defined by the equality $H_n(\widehat{\theta}_n) = 0$. For different choices of h we get different estimators—different functions of the data. The choice of h can be suggested by a model or by various optimality criteria.

For simplicity I will consider only the case where Θ is a subset of the real line. Vector-valued parameters can be handled by taking h as a vector-valued function.

<1> **Example.** Suppose the data $x_1, \ldots, x_n$ are modelled as independent observations from a density belonging to a family $\{ f_\theta(x) : \theta \in \Theta\}$.

The maximum likelihood estimator (MLE) is defined as the value that maximizes the joint density $p(x_1, \ldots, x_n, \theta) = \prod_{i \le n} f_\theta(x_i)$. If f_θ is a smooth function of θ, and if the maximum occurs at the point where $\partial p/\partial\theta$ is zero, the MLE corresponds to the Z-estimator defined by

$$h(x, \theta) = \frac{\partial}{\partial\theta} \log f(x, \theta).$$

For example, for fitting a $N(\theta, 1)$ model the function $h(x, \theta) = x - \theta$ generates the MLE, which happens to coincide with the ***method of moments estimator***. In general, if $m(\theta) := \int x f_\theta(x)\, dx$, the method of moments estimator for a one-dimensional parameter θ is defined as the solution to $m(\widehat{\theta}_n) = \sum_{i \le n} x_i/n$, which corresponds to the function $h(x, \theta) := x - m(\theta)$. □

For the purposes of numerical illustration I will work with the function

<2>
$$\bar{h}(x, \theta) = \begin{cases} x - \theta & \text{if } |x - \theta| \le 1 \\ +1 & \text{if } x > \theta + 1 \\ -1 & \text{if } x < \theta - 1 \end{cases}$$

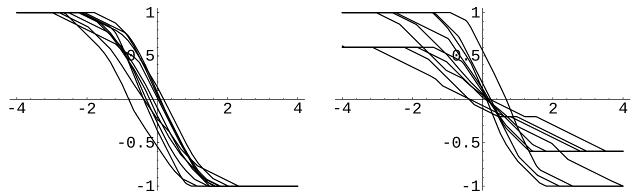
I choose this particular function for two reasons: there is no simple closed-form expression for the corresponding Z-estimator $\bar{\theta}_n$; and similar h functions have played an important role in the modern theory of "robust statistics". The first property shows why it is important to have some general theory for the behaviour of Z-estimators, to cover cases where we cannot analyze a closed-form representation. For this handout, whenever I write a bar over a function or estimator you will know that I am referring to this particular choice for h. Thus $\overline{H}_n$ denotes the corresponding criterion function for a sample of size n, and $\bar{\theta}_n$ is defined by $\overline{H}_n(\bar{\theta}_n) = 0$.

Suppose the data are generated as independent observations from some fixed density f. (The method also works for discrete distributions. I leave the substitution of sums for integrals to you.) It is seldom possible to calculate the exact distribution of the Z-estimator $\widehat{\theta}_n$. But, as I will soon explain,

> *if the function h is smooth enough in θ and the sample size n is large enough, then $\widehat{\theta}_n$ will typically have an approximately normal distribution, with variance of order $1/n$.*

To understand why $\widehat{\theta}_n$ behaves well for large samples from a fixed f, we first have to understand what the random criterion function H_n is doing. Different realizations of the data generate different H_n functions. For example, the following pictures were obtained by superimposing the $\overline{H}_n$ functions (corresponding to the $\bar{h}$ from <2>) for 10 independent samples of size $n = 5$:

from the $N(0,1)$ distribution on the left-hand side, and from the standard Cauchy distribution on the right-hand side.



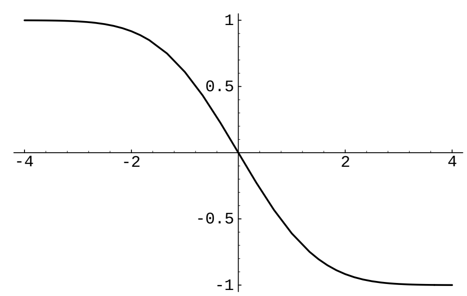
Look at the 10 realizations of the estimator $\bar{\theta}_n$ (the points at which the $\overline{H}_n$ curves intersect the horizontal axis) in each picture. Notice the spread around the origin. The estimator $\bar{\theta}_n$ has a distribution that depends on the joint distribution of the data.

Consistency

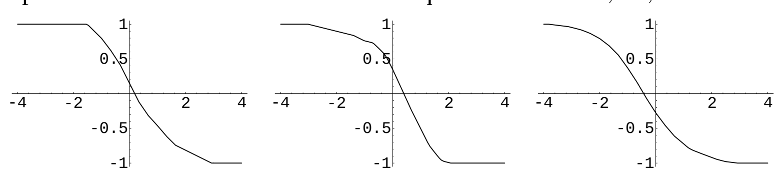
Suppose the x_i are independent observations from some density f. For each fixed θ, the random variable $H_n(\theta)$ is an average of the n independent random variables $h(x_i,\theta)$, for $i = 1,2,\ldots,n$. By the law of large numbers, $H_n(\theta)$ should be close to its expected value, $H(\theta,\, f) := \mathbb{E}_f h(x,\theta) = \int h(x,\theta)\, f(x)\, dx$.

For example, if f equals ϕ, the $N(0,1)$ density, with distribution function $\Phi(x)$, then the $\overline{H}(\theta,\phi)$ corresponding to the $\bar{h}$ from <2> is given by

$$\begin{aligned}\overline{H}(\theta,\phi) &= \int_{\theta+1}^{\infty}\phi(x)\,dx - \int_{-\infty}^{\theta-1}\phi(x)\,dx + \int_{\theta-1}^{\theta+1}(x-\theta)\phi(x)\,dx \\ &= 1 - \Phi(\theta+1) - \Phi(\theta-1) \\ &\qquad - \theta(\Phi(\theta+1) - \Phi(\theta-1)) - (\phi(\theta+1) - \phi(\theta-1)).\end{aligned}$$



With a little imagination you might convince yourself that each of the 10 superimposed plots for $\overline{H}_5(\cdot)$ from the $N(0,1)$ density look like $\overline{H}(\cdot\,,\phi)$. Perhaps the effect is more obvious in a sequence for $n = 5, 10, 20$:



As n gets larger, $\overline{H}_n$ converges to $\overline{H}(\cdot\,,\phi)$. With probability tending to one, the estimator $\bar{\theta}_n$ concentrates around the solution $\theta = 0$ for the equation $\overline{H}(\theta,\phi) = 0$. That is, $\bar{\theta}_n$ converges in probability to 0 for independent samples from the $N(0,1)$ density.

For general h with data $x_1, x_2, \ldots$ generated independently from a density f, the estimator $\widehat{\theta}_n$ converges in probability (as $n \to \infty$) to $z = z(f,h)$, the root—which I assume is unique—of the equation $H(z,\, f) = 0$.

If the data $x_1, x_2 \ldots$ are modelled as independent observations from a density belonging to some f from a family $\{f_\theta(x) : \theta \in \Theta\}$, it is traditional to consider behaviour of $\widehat{\theta}_n$ under each f_θ. If the equation $H(z, f_\theta) = 0$ has its solution $z(f_\theta, h)$ equal to θ, for each θ, then $\widehat{\theta}_n$ will converge in probability under the f_θ model to θ; if the data actually are generated from an f_{θ_0}, for an unknown "true" value θ_0, then the Z-estimator will converge to that θ_0. This consistent requirement places a constraint on h.

Asymptotic normality

How closely will $\widehat{\theta}_n$ be distributed about its limiting value $z = z(f, h)$, for data generated independently from a density f? A Taylor expansion about z gives the answer.

$$0 = H_n(\widehat{\theta}_n) \approx H_n(z) + (\theta - z)\textstyle\sum_{i \le n} h'(x_i, z)/n,$$

where h' denotes the partial derivative of h with respect to θ. Solve.

$$\sqrt{n}(\widehat{\theta}_n - z) = \frac{-\sqrt{n} H_n(z)}{\sum_{i \le n} h'(x_i, z)/n}.$$

You'll see in a moment why I have multiplied through by a $\sqrt{n}$. By the law of large numbers, the average in the denominator has large probability of being close to

$$J(f, h) := \mathbb{E}_f h'(x, z) = \int h'(x, z) f(x)\, dx$$

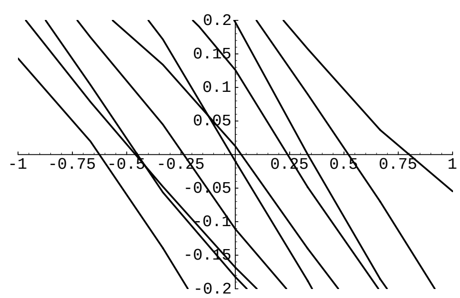
where z is defined by the equality $\int h(x, z) f(x)\, dx = 0$.

A similar argument shows that $H_n(z)$ should lie close to $0 = H(z, f)$, which merely reconfirms that $\widehat{\theta}_n - z$ should get close to zero. We can do better. As an average of independent random variables $h(x_i, z)$ with zero expected values, the random variable $H_n(z)$ will have an approximate $N(0, \sigma^2(f, h)/n)$ distribution, where

$$\sigma^2(f, h) := \text{var}_f h(x, z) = \int h(x, z)^2 f(x)\, dx \qquad \text{because } H(z, f) = 0.$$

The extra factor of $\sqrt{n}$ magnifies the $H_n(z)$ up to a quantity distributed roughly $N(0, \sigma^2(f, h))$, leaving $\sqrt{n}(\widehat{\theta}_n - z)$ with an approximate normal distribution with zero mean and variance equal to $\sigma^2(f, h)/J(f, h)^2$.

A magnified version of the picture for ten realizations of the $\overline{H}_{20}$ function, generated from samples of size 20 with f equal to the standard Cauchy, shows what is going on.



In the region near zero, where the Z-estimator lies with high probability, the $\overline{H}_n$ function is roughly linear, with slope equal to $J(f, h)$, with the intercept shifted around by the random variable $H_n(z)$, which has roughly a normal distribution with standard deviation of order $1/\sqrt{n}$. For the asymptotics at the $1/\sqrt{n}$-level, the Z-estimation problem reduces to a simple linear equation, through the workings of the law of large numbers and the central limit theorem.

Optimal choice of h

Once again consider the situation where the data $x_1, \ldots, x_n$ are modelled as independent observations from a density belonging to a family $\{f_\theta(x) : \theta \in \Theta\}$. Let me write $\mathbb{E}_\theta$ and var_θ, instead of $\mathbb{E}_{f_\theta}$ and var_{f_θ}, to denote calculations carried out under the f_θ model. Thus $\mathbb{E}_\theta g(x)$ will be shorthand for $\int g(x) f_\theta(x)\,dx$; the x is treated both as a generic x_i and as a dummy variable of integration.

In order that the Z-estimator $\widehat{\theta}_n$ should converge to θ under the f_θ model, *for every* θ, we must have

<3> $$\mathbb{E}_\theta h(x, \theta) = \int h(x, \theta) f_\theta(x)\,dx = 0 \qquad \text{for every } \theta.$$

The problem is to find the h function that minimizes

$$\frac{\sigma^2(f_\theta, h)}{J(f_\theta, h)^2} := \frac{\mathbb{E}_\theta h(x, \theta)^2}{(\mathbb{E}_\theta h'(x, \theta))^2}$$

for each θ, subject to the constraint <3>.

I will show that the minimum is achieved when $h(x, \theta)$ equals

$$\ell_\theta(x) := \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{f'_\theta(x)}{f_\theta(x)}.$$

That is, the lower bound for asymptotic variance is achieved when h defines the MLE. Some authors call ℓ the ***score function*** for the model. The corresponding $J(f_\theta, \ell_\theta)$ is given by

$$-J(f_\theta, \ell_\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial\theta^2} \log f_\theta(x) \right) =: \mathcal{I}(\theta) = \text{var}_\theta(\ell_\theta) = \mathbb{E}_\theta \ell_\theta^2,$$

the ***(Fisher) information function*** for the model.

To establish the optimality property for ℓ, we can argue as in the proof of the information inequality to derive first a lower bound for $\sigma^2(f_\theta, h)/J(f_\theta, h)^2$, and then show that ℓ achieves that lower bound.

Differentiate <3> with respect to θ, assuming appropriate smoothness for the densities (and ignoring the question of whether we are allowed to take the derivative inside the integral sign):

<4> $$\int h'(x, \theta) f_\theta(x)\,dx + \int h(x, \theta) f'_\theta(x)\,dx = 0.$$

Recognize the first integral as $J(f_\theta, h)$. Rewrite $f'_\theta(x)$ as $\ell_\theta(x) f_\theta(x)$ to recognize the second integral as $\mathbb{E}_\theta h(x, \theta)\ell_\theta(x)$, and thereby deduce that $J(f_\theta, h) = -\mathbb{E}_\theta h(x, \theta)\ell_\theta(x)$ and

$$\sigma^2(f_\theta, h)/J(f_\theta, h)^2 = \frac{\mathbb{E}_\theta h(x, \theta)^2}{(\mathbb{E}_\theta h(x, \theta)\ell_\theta(x))^2}.$$

The Cauchy-Schwarz inequality asserts

$$\left(\mathbb{E}_\theta h(x, \theta)^2\right)\left(\mathbb{E}_\theta \ell_\theta(x)^2\right) \geq \left(\mathbb{E}_\theta h(x, \theta)\ell_\theta(x)\right)^2,$$

with equality when $h(x, \theta)$ equals $\ell_\theta(x)$. Thus

$$\sigma^2(f_\theta, h)/J(f_\theta, h)^2 \geq 1/\mathbb{E}_\theta \ell_\theta(x)^2 = 1/\mathcal{I}(\theta),$$

with equality when $h(x, \theta) = \ell_\theta(x)$, in which case $\sigma^2(f_\theta, \ell) = \mathcal{I}(\theta) = -J(f_\theta, \ell)$ and $\sqrt{n}(\widehat{\theta}_n - \theta)$ is approximately $N(0, 1/\mathcal{I}(\theta))$ distributed under the f_θ model, for each θ.

In summary: If we require that the Z-estimator converge in probability to θ under independent sampling from f_θ, *for every* θ, then the asymptotic variance cannot be smaller than $1/\mathbb{I}_\theta$. The asymptotic normal distribution for the MLE has variance equal to the lower bound.

Warnings

I have not been rigorous about the conditions required for the arguments leading to "asymptotic optimality" of the MLE amongst the class of consistent Z-estimators. For example, the argument surely fails when f_θ denotes the Uniform$(0, \theta)$ density, which is not everywhere differentiable. A completely rigorous treatment is quite difficult. The development of the rigorous theory has been a major theme in modern theoretical statistics.