Chapter 8

Likelihood Processes

This chapter describes behavior of likelihood ratios for large sample sizes. We apply results from chapter 4 to show that derivatives of log-likelihoods are additive martingales and results from chapter 7 to show that likelihood ratio processes are multiplicative martingales. This chapter studies settings in which the state vector can be inferred from observed data, while chapter 9 studies settings in which states are hidden but can be imperfectly inferred from observed variables.

8.1 Warmup

This section introduces elementary versions of objects that will play key roles in this chapter. We simplify things by assuming that successive observations of a random vector y are independently and identically distributed draws from a statistical model $\psi(y|\theta_o)$. In this section, we study aspects of the "inverse problem" of inferring θ_o from a sequence observed y's. Remaining sections consider situations in which observations are not identically and independently distributed.

As an instance of a chapter 1, section 1.1 setup, assume that an unknown parameter vector θ resides in a set Θ. Given θ, a vector Y with realizations in $\mathcal{Y}$ is described by a probability density $\psi(y|\theta)$ relative to a measure τ over $\mathcal{Y}$. We also call the density $\psi(y|\theta)$ a likelihood function or statistical model. We assume that a sample $y_1, \ldots, y_T$ is a set of independent draws from $\psi(y|\theta)$. The likelihood function of this sample is

$$L(y_1, \ldots, y_T) = \Pi_{t=1}^{T} \psi(y_t|\theta).$$

We want to study the maximum likelihood estimator of θ as sample size $T \to +\infty$. For this purpose, we define the log density

$$\lambda(y|\theta) = \log \psi(y|\theta)$$

and the log-likelihood function multiplied by T^{-1}

$$l_T(\theta) = T^{-1} \sum_{t=1}^{T} \lambda(y_t|\theta) = T^{-1} \log L(y_1, \ldots, y_T|\theta).$$

We are interested in the behavior of the random variable $l_T(\theta)$ as $T \to +\infty$. For this purpose, let $\psi(y|\theta_o)$ be the density from which the data are actually drawn. Under $\psi(y|\theta_o)$, a law of large numbers implies that

$$l_T(\theta) \to E[\lambda(y|\theta)],$$

where the mathematical expectation is evaluated with respect to the density $\psi(y|\theta_o)$, so that

$$\begin{aligned} E[\lambda(y|\theta)] &= \int \lambda(y|\theta)\psi(y|\theta_o)\tau(dy) \\ &\equiv l_\infty(\theta) \end{aligned}$$

(Here it is understood that $l_T(\theta)$ depends on θ^o.) We can regard $l_\infty(\theta)$ as the population value of the (scaled by T^{-1}) log likelihood function of the parameter vector θ when the data are governed by successive independent draws from the statistical model $\psi(y|\theta_o)$. The following proposition describes a key desirable property of the maximum likelihood estimator of θ for infinite values of T.

Proposition 8.1.1.

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\ l_\infty(\theta) = \theta_o$$

Proof. Let r be the likelihood ratio

$$r = \frac{\psi(y|\theta)}{\psi(y|\theta_o)}.$$

The inequality $\log r \leq r - 1$ (see figure 81) implies

$$\begin{aligned}\int \log\left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)}\right)\psi(y|\theta_o)\tau(dy) &\leq \int\left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)} - 1\right)\psi(y|\theta_o)\tau(dy)\\ &= \int \psi(y|\theta)\tau(dy) - \int \psi(y|\theta_o)\tau(dy)\\ &= 1 - 1 = 0.\end{aligned}$$

Therefore

$$\int \log \psi(y|\theta)\psi(y|\theta_o)\tau(dy) \leq \int \log \psi(y|\theta_o)\psi(y|\theta_o)\tau(dy)$$

or

$$l_\infty(\theta) \leq l_\infty(\theta_o).$$

□

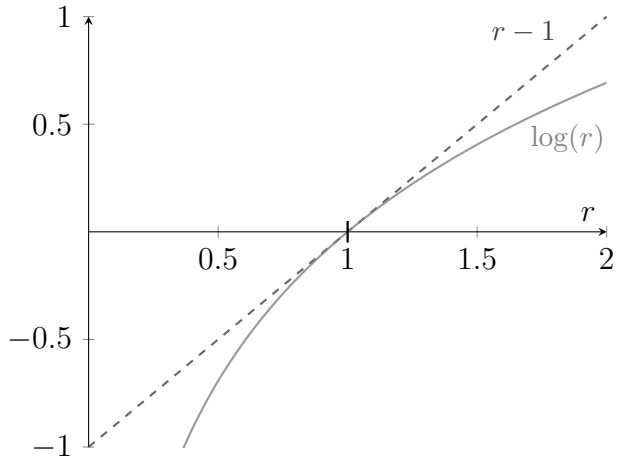

Figure 81: The inequality $(r-1) \geq \log(r)$.

Definition 8.1.2. *The parameter vector θ is said to be* ***identified*** *if* $\underset{\theta\in\Theta}{\operatorname{argmax}}\, l_\infty(\theta)$ *is unique.*

Definition 8.1.3. *The time t element of the* score process *of a likelihood function is defined as*

$$s_t(\theta) = \frac{\partial \lambda(y_t|\theta)}{\partial \theta}.$$

Remark 8.1.4. *The first-order necessary condition for the (population) maximum likelihood estimator is*

$$Es_t(\theta) = 0,$$

where again the mathematical expectation is taken with respect to the density $\psi(y|\theta^o)$.

Definition 8.1.5. *The Fisher information matrix is defined as*

$$I(\theta_o) = Es_t s_t' = \int s_t(\theta_o) s_t(\theta_o)' f(y|\theta_0)\tau(dy)$$

A necessary condition for the parameter vector θ_o to be identified is that the Fisher information matrix $I(\theta_o)$ be a positive definite matrix.

Kullback-Leibler discrepancy

Let

$$m = \frac{\psi(y|\theta)}{\psi(y|\theta_o)}$$

be the likelihood ratio of the θ model relative to the θ_o model. Kullback-Leibler relative entropy is defined as[1]

$$\begin{aligned} \text{ent} = Em\log m &= \int \left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)}\right) \log\left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)}\right) \psi(y|\theta_o)\tau(dy) \\ &= \int \log\left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)}\right) \psi(y|\theta)\tau(dy) \end{aligned}$$

where the mathematical expectation is under the θ_o model. Kullback and Leibler's relative entropy concept is often used to measure the discrepancy between two densities $\psi(y|\theta)$ and $\psi(y|\theta_o)$. This use exploits the important property that $Em\log m$ is nonnegative and that it is zero when $\theta = \theta_o$. To show this, first note the inequality

$$m\log m \geq m - 1.$$

Please see figure 82. After noting that $Em = 1$, it then follows that

$$Em\log m \geq Em - 1 = 1 - 1 = 0$$

[1]See Kullback and Leibler (1951).

So Kullback-Leibler relative entropy is nonnegative. Further, notice that $Em \log m$ can be represented as

$$\int \log\left(\frac{\psi(y|\theta)}{\psi(y|\theta_o)}\right)\psi(y|\theta)\tau(dy) \geq 0$$

Therefore

$$\int \log\psi(y|\theta)\psi(y|\theta)\tau(dy) - \int \log\psi(y|\theta_o)\psi(y|\theta)\tau(dy) \geq 0$$

and equals zero when $\theta = \theta_0$. Thus, setting $\theta = \theta_o$ minimizes population Kullback-Leibler relative entropy.

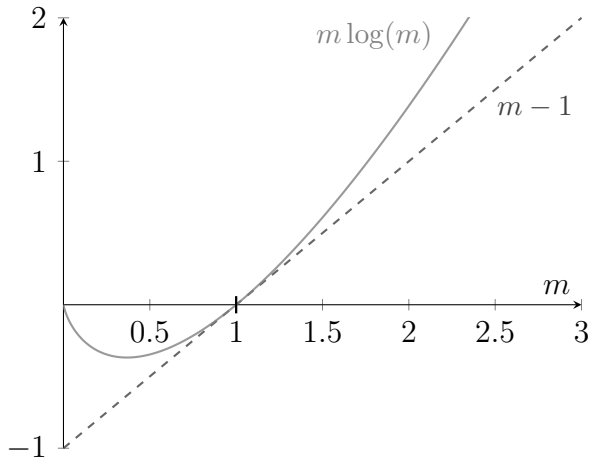

Figure 82: The inequality $(m-1) \leq m\log(m)$.

In the remainder of this chapter and in chapter 9 we extend our study to situations in which observations of y_t are not identically and independently distributed.

8.2 Dependent Processes

As in chapter 4, we let $\{W_{t+1}\}$ be a process of shocks satisfying

$$E\left(W_{t+1}|\mathfrak{F}_t\right) = 0.$$

We let $\{X_t : t = 0, 1, ...\} = \{X_t\}$ be a discrete time stationary Markov process generated by

$$X_{t+1} = \phi(X_t, W_{t+1}),$$

where ϕ is a Borel measurable function. We observe a vector of "signals" whose ith component evolves as an additive functional of $\{X_t\}$

$$Y_{t+1}^{[i]} - Y_t^{[i]} = \kappa_i(X_t, W_{t+1})$$

for $i = 1, 2, ..., k$. Stack these k signals into a vector $Y_{t+1} - Y_t$ and form

$$Y_{t+1} - Y_t = \kappa(X_t, W_{t+1}).$$

In this chapter we impose:

Assumption 8.2.1. *X_0 is observed and there exists a function χ such that*

$$W_{t+1} = \chi(X_t, Y_{t+1} - Y_t).$$

Assumption 8.2.1 is an invertibility condition that asserts that states $\{X_t\}_{t=1}^{\infty}$ can be recovered from signals. To verify this claim, recall that

$$X_{t+1} = \phi(X_t, W_{t+1}) = \phi[X_t, \chi(X_t, Y_{t+1} - Y_t)] \equiv \zeta(X_t, Y_{t+1} - Y_t), \quad (8.1)$$

which allows us to recover a sequence of states from the initial state X_0 and a sequence of signals. In chapter 9, we will relax assumption 8.2.1 and treat states $\{X_t\}_{t=0}^{\infty}$ as hidden.

Let τ denote a measure over the space $\mathcal{Y}$ of admissible signals.

Assumption 8.2.2. *$Y_{t+1} - Y_t$ has a conditional density $\psi(\cdot|x)$ with respect to τ conditioned on $X_t = x$.*

We want to construct the density ψ from κ and the distribution of X_{t+1} conditioned on X_t. One possibility is to construct ψ and τ as follows:

Example 8.2.3. *Suppose that κ is $Y_{t+1} - Y_t = X_{t+1}$ and that the Markov process $\{X_t\}$ has a transition density $p(x^*|x)$ relative to a measure λ. Set $\psi = p$ and $\tau = \lambda$.*

Another possibility is:

Example 8.2.4.

$$\begin{aligned} X_{t+1} &= AX_t + BW_{t+1} \\ Y_{t+1} - Y_t &= DX_t + FW_{t+1}, \end{aligned}$$

where A is a stable matrix, $\{W_{t+1}\}_{t=0}^{\infty}$ is an i.i.d. sequence of $\mathcal{N}(0, I)$ random vectors conditioned on X_0, and F is nonsingular. So

$$X_{t+1} = \left(A - BF^{-1}D\right) X_t + BF^{-1} \left(Y_{t+1} - Y_t\right),$$

which gives the function $\zeta(X_t, Y_{t+1} - Y_t) \doteq X_{t+1}$ in equation 8.1 from assumption 8.2.1 as a linear function of $Y_{t+1} - Y_t$ and X_t. The conditional distribution of $Y_{t+1} - Y_t$ is normal with mean DX_t and nonsingular covariance matrix FF', which gives us ψ. This conditional distribution has a density against Lebesgue measure on $\mathbb{R}^m$, a measure that we can use as τ.

8.3 Likelihood processes

Assumptions 8.2.1 and 8.2.1 imply that associated with a statistical model is a probability specification for $(X_{t+1}, Y_{t+1} - Y_t)$ conditioned X_t. The only information about the distribution of future Y's contained in current and past $(X_t, Y_t - Y_{t-1})$'s and X's is X_t.

The joint density r *likelihood function* of a history of observations $Y_1, ..., Y_t$ o conditioned on X_0 is

$$L_t = \prod_{j=1}^{t} \psi(Y_j - Y_{j-1} | X_{j-1}),$$

so

$$\log L_t = \sum_{j=1}^{t} \log \left[\psi(Y_j - Y_{j-1} | X_{j-1})\right].$$

Because they are functions of signals $\{Y_t - Y_{t-1}\}$ and states $\{X_t\}$, the *likelihood function* and *log-likelihood function* are both stochastic processes. Introducing a density for the initial state X_0 only affects initial conditions for L_0 and $\log L_0$.

Fact 8.3.1. *A log-likelihood process* $\{\log(L_t) : t = 0, 1, \ldots, t\}$ *is an additive functional with increment*

$$\log \psi(y^*|x) = \log \phi[\kappa(x, w^*)|x] \doteq \kappa_\ell(x, w^*),$$

where y^* *denotes a realized value of* $Y_{t+1} - Y_t$.

Fact 8.3.2. *Because a log-likelihood process is an additive functional, a likelihood process is a multiplicative functional.*

Example 8.3.3. *Consider example 8.2.4 again. It follows from the formula for the normal density that*

$$\begin{aligned}\kappa_\ell(X_{t+1}, X_t) = &-\frac{1}{2}(Y_{t+1} - Y_t - DX_t)'(FF')^{-1}(Y_{t+1} - Y_t - DX_t) \\ &-\frac{1}{2}\log\det(FF') - \frac{k}{2}\log(2\pi) \end{aligned} \tag{8.2}$$

and

$$\begin{aligned}\log L_t = &-\frac{1}{2}\sum_{j=1}^{t}(Y_j - Y_{j-1} - DX_{j-1})'(FF')^{-1}(Y_j - Y_{j-1} - DX_{j-1}) \\ &-\frac{t}{2}\log\det(FF') - \frac{kt}{2}\log(2\pi).\end{aligned}$$

If we know the transition density ψ only up to an unknown parameter vector θ, we have a set of transition density functions $\psi(y^*|x, \theta)$ indexed by parameter θ in a space Θ. For each parameter vector θ we construct

$$\log \psi(y^*|x, \theta) = \log \psi[\kappa(x, w^*)|x, \theta] \doteq \kappa_\ell(x, w^*|\theta).$$

Let $\log L_0(\theta)$ be the logarithm of an initial density function for X_0 that we also allow to depend on θ. When it is unique, we sometimes use the stationary distribution as the distribution for X_0 . A log-likelihood process is

$$\log L_t(\theta) = \sum_{j=1}^{t} \log \psi\left(Y_j - Y_{j-1}|X_{j-1}, \theta\right) + \log L_0(\theta).$$

The implied probability distributions for all $t \geq 1$ along with a density for X_0 is the statistical model indexed by parameters θ.

Because a log-likelihood process is an additive functional (see fact 8.3.1), Proposition 4.2.3 tells us that it has a representation

$$\log L_t(\theta) = \nu(\theta)t + \sum_{j=1}^{t} \kappa_a(X_{j-1}, W_j|\theta) - g(X_t|\theta) + g(X_0|\theta) + \log L_0(\theta). \quad (8.3)$$

Let θ_o be the parameter value that generates the data. Applying a law of large numbers and the proposition 4.2.3 properties of its components to (8.3) establishes that under the statistical model with $\theta = \theta_o$

$$\lim_{t \to +\infty} \frac{1}{t} \log L_t(\theta_o) = \nu(\theta_o). \quad (8.4)$$

We will use result (8.4) when we discuss maximum likelihood estimation of the parameter vector θ below.

8.4 Likelihood ratios

Ratios of likelihoods associated with two different parameter vectors, say θ and θ_o, can be constructed by multiplicatively cumulating a sequence of multiplicative increments of the form

$$\exp\left[\kappa_\ell(x, w^*|\theta) - \kappa_\ell(x, w^*|\theta_o)\right] = \frac{\psi(y^*|x, \theta)}{\psi(y^*|x, \theta_o)}.$$

Under the θ_o probability model, the conditional expectation of a multiplicative increment to the likelihood ratio is

$$\int_{\mathcal{Y}} \left[\frac{\psi(y^*|x, \theta)}{\psi(y^*|x, \theta_o)}\right] \psi(y^*|x, \theta_o) \upsilon(dy^*) = \int_{\mathcal{Y}} \psi(y^*|x, \theta) \upsilon(dy^*) = 1 \quad (8.5)$$

for all $x \in \mathcal{X}$ and all $\theta \in \Theta$. This follows because $\psi(y^*|x, \theta)$ is a density for every x and θ. We have established

Theorem 8.4.1. *For each $\theta \in \Theta$, the likelihood ratio process $\left\{ \frac{L_t(\theta)}{L_t(\theta_o)} : t = 0, 1, ... \right\}$ is a multiplicative martingale under the θ_o probability model.*

Thus, under the statistical model with parameter vector θ_o

$$E\left[\frac{L_t(\theta)}{L_t(\theta_o)}\middle|\mathfrak{F}_{t-1}\right] = \frac{L_{t-1}(\theta)}{L_{t-1}(\theta_o)}.$$

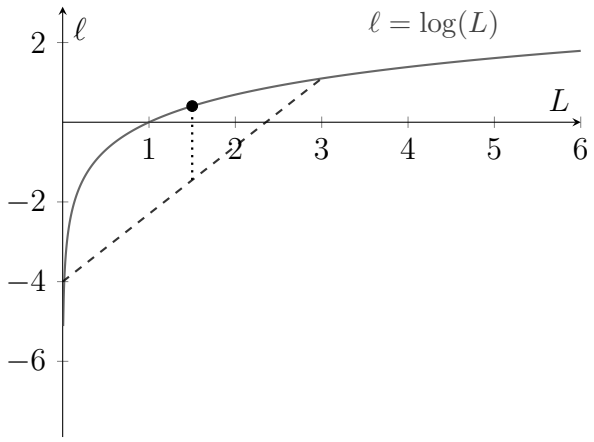

Figure 83: Jensen's Inequality. The logarithmic function is the concave function that equals zero when evaluated at unity. By forming averages using the two endpoints of the straight line below the logarithmic function, we get a point on the line segment that depends on the weights used in the averaging. Jensen's Inequality asserts that the line segment lies below the logarithmic function.

8.5 Log-Likelihoods

Because the logarithmic function is concave, Jensen's Inequality implies that the expectation of the logarithm of a random variable cannot exceed the logarithm of the expectation. (See figure 83.) This reasoning implies the following inequality:

$$\begin{aligned}&\int_{\mathcal{Y}} \left[\log \psi(y^*|x,\theta) - \log \psi(y^*|x,\theta_o)\right] \psi(y^*|x,\theta_o)\tau(dy^*) \\ &\leq \log \int_{\mathcal{Y}} \left[\frac{\psi(y^*|x,\theta)}{\psi(y^*|x,\theta_o)}\right] \psi(y^*|x,\theta_o)\tau(dy^*) = 0. \end{aligned} \tag{8.6}$$

Inequality (8.6) holds with equality only if

$$\log \psi(y^*|x,\theta) = \log \psi(y^*|x,\theta_o)$$

with probability one under the measure $\psi(y^*|x,\theta_o)\tau(dy^*)$. A consequence of inequality 8.6 is that

$$E\left[\log L_{t+1}(\theta) - \log L_{t+1}(\theta_o)|\mathcal{F}_t\right] \leq \log L_t(\theta) - \log L_t(\theta_o). \tag{8.7}$$

Definition 8.5.1. *An additive functional* $\{Y_t\}_{t=0}^{\infty}$ *is said to be a supermartingale with respect to a filtration* $\{\mathfrak{F}_t\}_{t=0}^{\infty}$ *if*

$$E\left(Y_{t+1}|\mathfrak{F}_t\right) \leq Y_t.$$

Inequality (8.7) implies

Theorem 8.5.2. *For each* θ*, the log-likelihood ratio process* $\{\log L_t(\theta) - \log L_t(\theta_o) : t = 0, 1, ...\}$ *is an additive super martingale under the* θ_o *probability model.*

We now combine two facts: (a) that because each log-likelihood process is an additive process, each has a proposition 4.2.3 decomposition of the form (8.3), and (b) that under the θ_o model, a log-likelihood ratio process $\{\log L_t(\theta) - \log L_t(\theta_o) : t = 0, 1, ...\}$ is a supermartingale. It follows that a log-likelihood ratio process has an additive decomposition with a coefficient $\nu(\theta) - \nu(\theta_o)$ on the linear time trend t, that this coefficient is less than or equal to zero, and that it equals zero only when $\theta = \theta_o$. Therefore, the parameter θ_o satisfies

$$\theta_o = \text{argmax}_{\theta \in \Theta} \nu(\theta). \tag{8.8}$$

The method of maximum likelihood appeals to equation (8.4) and the Law of Large Numbers to estimate the time trend of a log-likelihood by the sample average[2]

$$\hat{\nu}_t(\theta) = \frac{1}{t} \log L_t(\theta). \tag{8.9}$$

The *maximum likelihood estimator* of the vector θ based on data up to date t maximizes $\hat{\nu}_t(\theta)$. Heuristic justifications for this assertion come from combining insights from equations (8.4), (8.9), and (8.8).

The trend coefficient $\nu(\theta) - \nu(\theta_o)$ in a proposition 4.2.3 decomposition of the log-likelihood ratio governs the large sample behavior of likelihood ratio statistics for discriminating statistical model θ from statistical model θ_o. We study this in section 8.7.

[2]Use of the Law of Large Numbers pointwise in θ is typically not sufficient to justify statistical consistency. Instead, one has to justify estimation of ν as a function of θ over an admissible parameter space.

8.6 Score processes

In this section, we study first-order necessary necessary conditions for the maximum likelihood estimator. For simplicity suppose that Θ is an open interval of $\mathbb{R}$ containing θ_o. Inequality 8.6 implies that the parameter vector θ_o necessarily maximizes the objective

$$\int_{\mathcal{X}} \log \psi(y^*|x,\theta)\psi(y^*|x,\theta_o)\tau(dy^*). \tag{8.10}$$

Suppose that we can differentiate under the integral sign in (8.10) to get the first-order condition

$$\int_{\mathcal{X}} \left[\frac{d}{d\theta}\log \psi(y^*|x,\theta_o)\right]\psi(y^*|x,\theta_o)\tau(dy^*) = 0. \tag{8.11}$$

Definition 8.6.1. *Where $\ell_t(\theta) = \log L_t(\theta)$, the **score process** $\{S_t : t = 0, 1, ...\}$ is defined as*

$$S_t = \frac{d}{d\theta}\ell_t(\theta)|_{\theta_o} = \sum_{j=1}^{t}\frac{d}{d\theta}\log\psi(Y_j - Y_{j-1}|X_{j-1},\theta_o) + \frac{d}{d\theta}\log L_0(\theta_o).$$

Theorem 8.6.2. *$E\left(S_{t+1} - S_t|X_t\right) = 0$, so the score process is an additive functional; more specifically, it is a martingale with increment $\frac{d}{d\theta}\log\psi(Y_j - Y_{j-1}|X_{j-1},\theta_o)$ under the θ_o probability model. So*

$$\frac{1}{\sqrt{t}}S_t \to \mathcal{N}(0, V)$$

where $V = E\left(\left[\frac{d}{d\theta}\log\psi(Y_{t+1} - Y_t|X_t,\theta_o)\right]^2\right)$.

Proof. This follows directly from equation (8.11) and corollary 4.7.1.[3] □

Theorem 8.6.2 justifies using the martingale central limit theorem outcome stated in corollary 4.7.1 to characterize the large sample behavior of the score process. The associated central limit approximation yields a large sample characterization of the maximum likelihood estimator of θ in a Markov setting. Where θ_t maximizes the log-likelihood function $\ell_t(\theta)$, the

[3] In the formula we use the notation a^2 to refer to the matrix $a'a$ where a is an $n \times 1$ vector.

following result typically prevails under some additional regularity conditions:

$$\sqrt{t}(\theta_t - \theta_o) \to \mathcal{N}(0, V^{-1}).$$

This kind of result motivates interpreting the variance of the martingale increment of the score process

$$V = E\left(\left[\frac{d}{d\theta}\log\psi(Y_{t+1} - Y_t|X_t, \theta_o)\right]^2\right)$$

as a measure of the information in the data about the parameter θ_o; V is called the Fisher information matrix after the statistician R.A. Fisher.

Nuisance parameters

Consider extending the notion of Fisher information to situations in which we want information about one parameter but to get it we have to estimate other parameters too. Assume that there is an unknown scalar parameter θ_o of interest and a vector $\tilde{\theta}_o$ of "nuisance" parameters that also unknown and that we must also estimate. Suppose that the likelihood is parameterized on an open set Θ in a finite dimensional Euclidean space and that the true parameter vector is $(\theta_o, \tilde{\theta}_o) \in \Theta$. Write the multivariate score process as

$$\left\{\begin{bmatrix} S_{t+1} \\ \tilde{S}_{t+1} \end{bmatrix} : t = 0, 1, ...\right\}$$

where $\{S_{t+1} : t = 0, 1, ...\}$ is the partial derivative of the log-likelihood with respect to the parameter of interest θ and $\{\tilde{S}_{t+1} : t = 0, 1, ...\}$ is the partial derivative with respect to the nuisance parameters $\tilde{\theta}$.

Estimating $\tilde{\theta}_o$ simultaneously with θ_o is more difficult than estimating θ_o conditional on knowing $\tilde{\theta}_o$. Fisher's measure of information takes into account additional uncertainty that comes from having to estimate $\tilde{\theta}_o$ simultaneously with θ_o in order to make inferences about θ_o. In particular, Fisher's measure of information about θ_o is the inverse of the $(1,1)$ component of the covariance matrix partitioned conformably with $(\theta', \tilde{\theta}')'$:

$$V = \begin{bmatrix} E(S_{t+1} - S_t)^2 & E(\tilde{S}_{t+1} - \tilde{S}_t)(S_{t+1} - S_t) \\ E(\tilde{S}_{t+1} - \tilde{S}_t)(S_{t+1} - S_t) & E(\tilde{S}_{t+1} - \tilde{S}_t)^2 \end{bmatrix}.$$

To represent the inverse of V, compute the population regression

$$S_{t+1} - S_t = \beta'(\tilde{S}_{t+1} - \tilde{S}_t) + U_{t+1}, \tag{8.12}$$

where β is the population regression coefficient and U_{t+1} is the population regression residual that by construction is orthogonal to the regressor $(\tilde{S}_{t+1} - \tilde{S}_t)$. The population regression induces the representation

$$\begin{bmatrix} S_{t+1} - S_t \\ \tilde{S}_{t+1} - \tilde{S}_t \end{bmatrix} = \begin{bmatrix} I & \beta' \\ 0 & I \end{bmatrix} \begin{bmatrix} U_t \\ \tilde{S}_{t+1} - \tilde{S}_t \end{bmatrix},$$

which, because U_{t+1} is orthogonal to $\tilde{S}_{t+1} - \tilde{S}_t$, implies

$$V = \begin{bmatrix} 1 & \beta' \\ 0 & I \end{bmatrix} \begin{bmatrix} E\left[(U_{t+1})^2\right] & 0 \\ 0 & E\left[(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)'\right] \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta & I \end{bmatrix}.$$

Since

$$\begin{bmatrix} 1 & 0 \\ \beta & I \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\beta & I \end{bmatrix}$$

and

$$\begin{aligned} &\begin{bmatrix} E\left[(U_{t+1})^2\right] & 0 \\ 0 & E\left[(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)'\right] \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \frac{1}{E[(U_{t+1})^2]} & 0 \\ 0 & \left(E\left[(\tilde{S}_{t+1} - \tilde{S}_t)(\tilde{S}_{t+1} - \tilde{S}_t)'\right]\right)^{-1} \end{bmatrix}, \end{aligned}$$

the $(1,1)$ component of the partition of matrix V^{-1} is

$$V_{1,1}^{-1} = \frac{1}{E\left[(U_{t+1})^2\right]}.$$

Thus, the reciprocal of $E\left[(U_{t+1})^2\right]$ gives the Fisher information about θ_o in the presence of nuisance parameters.

The population regression equation (8.12) implies that the inverse of the Fisher information measure $E(U_{t+1})^2$ is no larger than $E(S_{t+1} - S_t)^2$:

$$E[(U_{t+1})^2] \leq E[(S_{t+1} - S_t)^2]. \tag{8.13}$$

This inequality asserts that the likelihood function contains more information about θ_o when $\tilde{\theta}$ is known to be $\tilde{\theta}_o$ than when θ_o and $\tilde{\theta}_o$ are both unknown.

Inequality (8.13) offers reasons to be cautious about interpreting some calibration practices in economics. Pretending that you know $\tilde{\theta}_o$ when you really don't leads you to overstate the information that a data set contains about the parameter θ_o.

Since the multivariate score $\begin{bmatrix} S_{t+1} \\ \tilde{S}_{t+1} \end{bmatrix}$ is a vector of additive martingales, the score regression residual process $\{U_{t+1} : t = 0, 1, ...\}$ is itself an additive martingale.

8.7 Limiting Behavior of Likelihood Ratios

In Chapter 1 we described the problem of selecting between two statistical models, say model θ_1 and model θ_2, based on observed data. As more data become available, the *ex ante* probability of making a mistake becomes smaller. To characterize this formally, we study the tail behavior of the likelihood ratio. We follow Chernoff (1952) and others by using a large deviation theory to measure the difficulty of choosing between models θ_1 and θ_2. Specifically, we study the probability that the likelihood ratio exceeds a given threshold and the probability of making a mistaken model selection via a likelihood ratio test. In line with the decision-theoretic perspective described in chapter 1, the threshold is determined by prior probabilities assigned to the two models and the losses associated with misclassification. For simplicity we assign prior probabilities equal to one half, but our calculations allow for other choices of these probabilities.

We use a large deviation calculation to characterize the limiting behavior of making type I (selecting model θ_2 when model θ_1 is true) and type II (selecting model θ_1 when model θ_2 is true) errors. Construct an additive functional defined by the log-likelihood ratio between two models, one indexed by θ_1 and another by θ_2:

$$Y_t = \log L_t(\theta_2) - \log L_t(\theta_1).$$

We know that the likelihood ratio process is a multiplicative martingale under the θ_1 probability measure. To study a model selection rule that

chooses the model with a higher likelihood, we construct a large deviations inequality that is implied by two elementary ideas:

- Probabilities are expectations of indicator functions.
- Indicator functions are bounded above by exponential functions with positive exponents.

For a scalar $\alpha > 0$, we use these two ideas to deduce the following inequalities:

$$\begin{aligned}\frac{1}{t}\log \text{Prob}\left\{\frac{1}{t}Y_t \geq 0 | X_0 = x\right\} &= \frac{1}{t}\log \text{Prob}\left\{Y_t \geq 0 | X_0 = x\right\} \\ &\leq \frac{1}{t}\log E\left[\exp\left(\alpha Y_t\right) | X_0 = x\right] \\ &= \frac{1}{t}\log E\left[(M_t)^\alpha | X_0 = x\right], \end{aligned} \tag{8.14}$$

where $M_t = \exp(Y_t)$. Let $\eta(M^\alpha)$ be the exponential trend $\tilde{\eta}$ in the Proposition 7.3.2 decomposition of the multiplicative functional M_t^α.

Taking limits of both sides of (8.14) as t gets large,

$$\limsup_{t\to\infty} \frac{1}{t}\log \text{Prob}\left\{\frac{1}{t}Y_t \geq 0 | X_0 = x\right\} \leq \eta\left(M^\alpha\right) \tag{8.15}$$

provided that $\eta\left(M^\alpha\right)$ is well defined and finite. Since $\{M_t\}_{t=0}^\infty$ is a multiplicative martingale, it follows from Jensen's Inequality that $\{Y_t\}_{t=0}^\infty$ and therefore $\{M_t^\alpha\}_{t=0}^\infty$ is an additive supermartingale and hence that

$$\eta\left(M^\alpha\right) \leq 0.$$

It can be shown that $\eta\left(M^\alpha\right)$ is concave in α. Different values of $\alpha > 0$ give rise to different bounds. To get an informative bound we compute:

$$-\varrho(M) = \min_{0\leq\alpha\leq 1} \eta\left(M^\alpha\right). \tag{8.16}$$

It follows from (8.15) and (8.16) that

$$\limsup_{t\to\infty} \frac{1}{t}\log \text{Prob}\left\{\frac{1}{t}Y_t \geq 0 | X_0 = x\right\} \leq -\varrho(M). \tag{8.17}$$

Thus, a positive $\varrho(M)$ bounds the decay rate of the probabilities on the left side of (8.17). Under some more stringent restrictions, the bound (8.17) can be shown to be sharp. In the theory of large deviations, $\varrho(M)$ is called the *rate function*.

Suppose now that θ_2 is the true parameter value. This induces us to use $\{M_t^{-1}\}$ as the likelihood-ratio process. This process is evidently a martingale under the probability measure $\psi(y^*|x,\theta_2)$ implied by θ_2. Following the preceding recipe, after raising M^{-1} to the power α and multiplying by M to change measure, we are led to study $M^{1-\alpha}$ for $0<\alpha<1$. Minimizing $\eta\left(M^{1-\alpha}\right)$ over α verifies that $\varrho(M)$ is again the asymptotic decay rate in the probability of making a mistake. The identical outcomes of these two calculations implies the invariance of $\varrho(M)$ across the two mental experiments, one that assumes that the θ_1 is true, the other that assumes that the θ_2 is true. That allowed Chernoff (1952) to interpret the rate bound $\varrho(M)$ as a measure of statistical discrepancy between two models.

While we have used zero as a threshold for the log-likelihood ratio, the analysis extends immediately to any constant threshold determined by the ratio of prior probabilities put on the two models. Both types of mistake probabilities converge to zero at an exponential rate that is independent of the threshold. For example 8.7.1 to be described next, figure 84 depicts $-\eta\left(M^{\alpha}\right)$ and $-\eta\left(M^{1-\alpha}\right)$ as functions of α. The two functions overlap completely. Both functions are concave in α. After describing the example, we shall tell in detail how we computed $-\eta\left(M^{\alpha}\right)$ and $-\eta\left(M^{1-\alpha}\right)$.

Example 8.7.1. *Statistical model* 1 *with parameter vector* $\theta_1=(\mu_1,\sigma^2)$ *asserts that a scalar random variable has a univariate normal distribution with mean* μ_1 *and variance* σ^2. *Statistical model* 2 *with parameter vector* $\theta_2=(\mu_2,\sigma^2)$ *asserts that a scalar random variable has a univariate normal distribution with mean* μ_2 *and variance* σ^2. *Under model* 1, *a draw* y_t *from the statistical model can be represented* $y_t=\mu_1+\sigma W_t$ *where* $W_t\sim\mathcal{N}(0,1)$. *For a sequence of independent draws* $y_t, t=1,2,\ldots,$ *from model* 1, *the log likelihood ratio* $Y_t=L_t(\theta_2)-L_t(\theta_1)$ *has increment*

$$\begin{aligned}
Y_t-Y_{t-1} &= -\frac{(y_t-\mu_2)^2}{2\sigma^2}+\frac{(y_t-\mu_1)^2}{2\sigma^2}\\
&= -\frac{(\mu_1-\mu_2+\sigma W_t)^2}{2\sigma^2}+\frac{(\sigma W_t)^2}{2\sigma^2}\\
&= -\frac{(\mu_1-\mu_2)^2}{2\sigma^2}-\frac{(\mu_1-\mu_2)W_t}{\sigma}.
\end{aligned}$$

The exponential trend term of the likelihood ratio itself is then $\exp(0) = 1$*, verifying a necessary condition for the likelihood ratio* $M_t = \exp(Y_t)$ *to be a martingale.*

The likelihood ratio $M_t = \exp(Y_t)$ in example 8.7.1 is an instance of a multiplicative functional with a corresponding additive functional sharing the form of those from example 7.3.5. Notice that for such a multiplicative functional M to be a martingale of the example 8.7.1 form, we must have

$$\log M_{t+1} - \log M_t = \nu + F \cdot W_{t+1}$$

with $\nu = -\frac{1}{2}F'F$ in order that the exponential trend term in the decomposition representation for M_t to be identically unity. We want to compute EM^α for $\alpha \in [0, 1]$. To do so, first multiply

$$\log M_{t+1} - \log M_t = -\frac{1}{2}F'F + F \cdot W_{t+1}$$

by $\alpha \in [0, 1]$ to get

$$\alpha \log M_{t+1} - \alpha \log M_t = -\alpha\frac{1}{2}F'F + \alpha F \cdot W_{t+1},$$

the right side of which is the increment in the logarithm of M_{t+1}^α that we seek. It follows that the (negative) growth rate of M_{t+1}^α is the decay rate $\frac{(\alpha^2-\alpha)}{2}F'F$. The rate

$$-\frac{(\alpha^2 - \alpha)}{2}F'F$$

is symmetric about its maximum value equal to $\frac{1}{8}F'F$ that is attained at $\alpha = \frac{1}{2}$, as depicted in figure 84. Thus, the graphs of exponential decay rates of M_t^α and $M_t^{1-\alpha}$ overlap completely.

Example for Dongchen

Recall example 8.7.1. Note that for a sequence of independent draws $y_t, t = 1, 2, \ldots$, from model 2, the log likelihood ratio $\tilde{Y}_t = L_t(\theta_1) - L_t(\theta_2)$ has increment

$$\tilde{Y}_t - \tilde{Y}_{t-1} = -\frac{(\mu_2 - \mu_1)^2}{2\sigma^2} - \frac{(\mu_2 - \mu_1)W_t}{\sigma}.$$