Likelihood Ratio and Score Processes

Contents

Download PDF here

Authors: Lars Peter Hansen (University of Chicago) and Thomas J. Sargent (NYU)

6.1. Introduction

In this chapter we study the behavior of likelihood ratio processes and score processes. We treat these as processes so that we can study their dynamic behavior when additional observations are included.

6.2. Multiplicative martingales

Let $\{M_t : t \geq 0\}$ denote a process whose logarithm evolves as:

$$\log M_{t+1} - \log M_t = \kappa(X_t, W_{t+1}).$$

Thus the logarithmic counterpart is recognizable as an additive functional as studied in Chapter 4. We call such an $\{M_t : t \geq 0\}$ a multiplicative functional. We explore properties of such processes in some generality in Chapter 7. In this chapter, the processes of interest are multiplicative martingales:

Definition 6.1

The process M is a **multiplicative martingale** if

$$E\left(\exp\left[\kappa(X_t, W_{t+1})\right] \mid X_t\right) = 1,$$

and thus $E\left(M_{t+1} \mid M_t, X_t\right) = M_t$.

Multiplicative martingales provide a convenient way to construct alternative probabilities.

Proposition 6.1

Suppose that $\{M_t : t \geq 0\}$ is a multiplicative martingale. The ratio M_{t+1}/M_t implies a transition probability conditioned on X_t via the formula:

$$\widetilde{E}\left[f(X_t, W_{t+1}) \mid X_t\right] = E\left[\left(\frac{M_{t+1}}{M_t}\right) f(X_t, W_{t+1}) \mid X_t\right].$$

The resulting process for X remains Markovian, and the implied τ-period transition probability can be represented as:

$$\widetilde{E}\left[f(X_{t+\tau-1}, W_{t+\tau}) \mid X_t\right] = E\left[\left(\frac{M_{t+\tau}}{M_t}\right) f(X_{t+\tau-1}, W_{t+\tau}) \mid X_t\right].$$

Proof. The linear operator:

$$\widetilde{E}\left(\cdot \mid X_t = x\right)$$

maps nonnegative functions into nonnegative functions and the unit function into itself. Thus this operator is a conditional expectation. Moreover, it maps functions of (w, x) into functions of x alone. This implies that $\{X_t : t \geq 0\}$ is a first-order Markov process under the implied change of probability. The representation of the implied τ-period conditional expectation operator follows from the Law of Iterated Expectations.

Remark 6.1

Under the change of measure induced by M_{t+1}/M_t, the shock W_{t+1} typically does not have conditional mean zero.

So far we have not restricted the initial condition M_0. We have only characterized its stochastic evolution. Suppose that the process $\{X_t\}$ is stationary with probability denoted Q, and consider $M_0 = \tilde{q}(X_0)$ where $\tilde{q}$ satisfies:

$$E\left[M_0 f(X_0)\right] = E\left[M_1 f(X_1)\right]$$
$$\int \tilde{q}(x) dQ(x) = 1$$

for all bounded (measurable) functions f of the Markov state x. Then $\widetilde{Q}(dx) \stackrel{\text{def}}{=} \tilde{q}(x)Q(dx)$ is a stationary distribution under the implied $\tilde{\cdot}$ probability distribution.

In what follows we will impose such an initial condition on M_0 so that we can apply a Law of Large Numbers as characterized in Chapter 1. In addition we will impose the ergodic restriction that the only solutions to the equation:

$$\widetilde{E}\left[\left(\frac{M_{t+1}}{M_t}\right) f(X_{t+1}) \mid X_t = x\right] = f(x)$$

are constant functions with $\tilde{Q}$ measure one as investigated in Chapter 2.

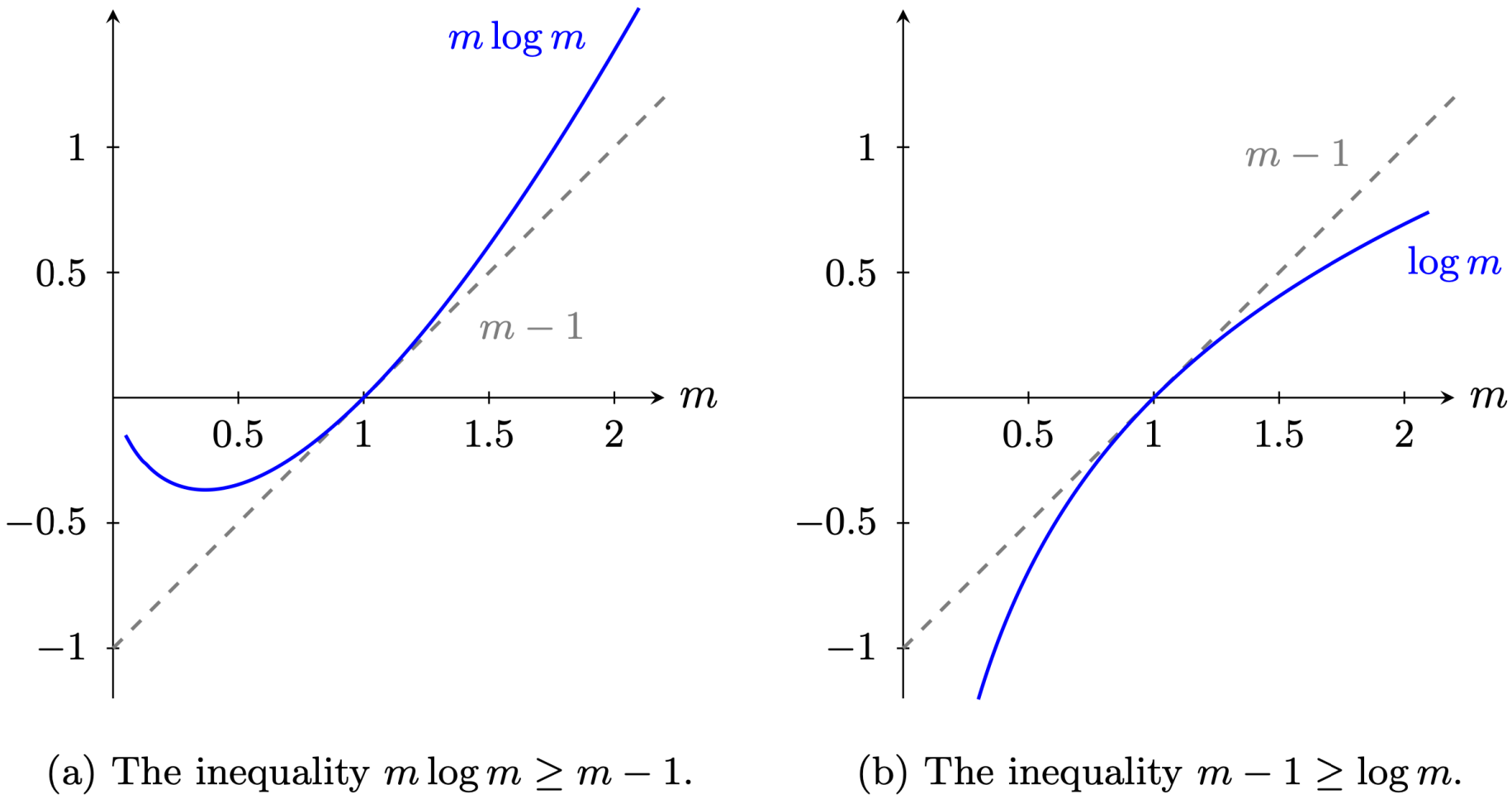
6.3. Multiple models

So far, we have considered two models, an initial one and a second one implied by a multiplicative martingale. Now suppose we have ℓ such models, given an initial one and $\ell - 1$ multiplicative martingales in addition to the initial one with multiplicative martingale equal to one. We denote each such martingale as $\{M_t^i : t \geq 0\}$ where each one induces a process that is stationary and ergodic, with $M_t^1 = 1$ for all $t \geq 0$. In terms of Chapter 1, think of this as partitioning sample space into ℓ partitions. Each partition is matched to one of the martingales, which in turn gives the probabilities conditioned on being in that partition.

In this construction, we chose to normalize model one to be the probability distribution used for representing all of the probabilities of the martingales. Suppose instead we had used model two for this purpose. Now the $\{M_t^i : t \geq 0\}$'s cease to be martingales because we have changed the underlying probability from model one to model two. The ratio processes $\{M_t^i / M_t^2 : t \geq 0\}$ are multiplicative martingales under this change in the underlying probability. This change in modeling perspective will be central in the discussion that follows.

6.4. Some Large Sample Properties

We next consider two gradient inequalities that we will use prominently. The function $m \log m$ is convex and the function $\log m$ is concave. A convex function lies above its gradient approximation and a concave function lies below its gradient. Observe that the gradient is one for both functions when $m = 1$ and the gradient approximation is $m - 1$. Fig. 6.1 illustrates these two inequalities.



(a) The inequality $m \log m \geq m - 1$. (b) The inequality $m - 1 \geq \log m$.

Fig. 6.1 Gradient inequalities

We use these inequalities in conjunction with conditional expectations to obtain the targets of interest. This amounts to an application of what is called Jensen's Inequality and implies

$$\begin{aligned} E\left[(\log M_{t+1} - \log M_t) M_{t+1}\right] &\geq 0 \\ E\left(\log M_{t+1} - \log M_t\right) &\leq 0. \end{aligned} \tag{6.1}$$

These weak inequalities become strict when $\log M_{t+1} - \log M_t$ is not equal to zero with probability one. Observe that $\{\log M_t : t \geq 0\}$ is an additive functional of the type studied in Chapter 4 with a negative trend coefficient. A consequence of the Law of Large Numbers and the second inequality in (6.1) is that:

$$\lim_{t \to \infty} \frac{1}{t} \log M_t = E\left(\log M_{t+1} - \log M_t\right) \leq 0$$

Suppose now we consider expectations under the probabilities implied by the multiplicative martingale $\{M_t \geq 0\}$. Then the first inequality in (6.1) implies that

$$\lim_{t\to\infty}\frac{1}{t}\log M_t = \widetilde{E}\left(\log M_{t+1} - \log M_t\right) \geq 0$$

These two limits give a large sample justification for using the criterion

$$\lim_{t\to\infty}\frac{1}{t}\log M_t$$

to determine which of the two models generates the data since the inequalities are strict when the multiplicative martingale is not degenerate. This insight extends directly to the case of multiple models captured by alternative multiplicative martingales. It leads to the use of

$$\max_{\ell=1,2,\ldots,\ell}\frac{1}{t}\log M_t^i$$

as a way to select models with a large sample justification. This has a direct extension to the method of maximum likelihood.

We conclude this section by pointing to unusual sample path properties of multiplicative martingales.

Theorem 6.1

For a nondegenerate multiplicative martingale, $\{M_t : t \geq 0\}$ converges almost surely to zero even though $E\left(\frac{M_t}{M_0} \mid X_0\right) = 1$ for each t.

Proof. We have observed that by the Law of Large Numbers:

$$\lim_{t\to\infty}\frac{1}{t}\log M_t$$

converges to a negative limit almost surely. This in turn implies that

$$\lim_{t\to\infty} M_t = 0$$

almost surely.

6.5. Loglikelihoods

We now give a particular construction of the function κ. Let θ denote a parameter vector and suppose that a vector of data at date $t+1$, Z_{t+1}, is given by

$$Z_{t+1} = \phi(X_t, W_{t+1}, \theta).$$

$$\begin{aligned}\kappa(x, w^* \mid \theta) &= \log \psi(z^* \mid x, \theta) - \log \psi(z^* \mid x, \theta_1) \\ &= \log \psi\left[\phi(x, w^*) \mid x, \theta\right] - \log \psi\left[\phi(x, w^*) \mid x, \theta_1\right]\end{aligned}$$

where $\psi(z^* \mid x, \theta)$ is the density of Z_{t+1} given X_t and θ. We further assume that given (x, θ), Z_{t+1} is informationally equivalent to W_{t+1}. This can be justified by assuming that we can invert ϕ as a function of W_{t+1}. We assume that given θ, X_t is revealed by past Z_t and possibly a finite number of lags. With this construction, we think of model θ_1 as the baseline used for computing expectations.

By factoring a joint density, we obtain a likelihood process conditioned on (X_0, θ) as:

$$L_t(\theta) = \sum_{j=1}^{t} \log \psi(Z_j \mid X_{j-1}, \theta).$$

We apply our previous arguments to justify finding the

$$\max_{i=1,2,\ldots\ell} L_t(\theta^i)$$

for a discrete set of models.[1]

6.6. Score Processes

In this section, we study first-order necessary conditions associated with the maximum likelihood estimator. For simplicity, suppose that the parameter space Θ is an open interval of $\mathbb{R}$ containing θ_1. Inequality (6.1) implies that the parameter value θ_1 necessarily maximizes the objective

$$\int_{\mathcal{X}} \log \psi(z^*|x, \theta)\psi(z^*|x, \theta_1)\tau(dz^*). \tag{6.2}$$

Suppose that we can differentiate under the integral sign in (6.2) to get the first-order condition

$$\int_{\mathcal{X}} \left[\frac{d}{d\theta} \log \psi(z^*|x, \theta_o)\right] \psi(z^*|x, \theta_1)\tau(dz^*) = 0. \tag{6.3}$$

Definition 6.2

The **score process** $\{S_t : t = 0, 1, \ldots\}$ is defined as

$$S_t = \frac{d}{d\theta} L_t(\theta)\Big|_{\theta_1} = \sum_{j=1}^{t} \frac{d}{d\theta} \log \psi(Z_j | X_{j-1}, \theta_1).$$

Theorem 6.2

$E\left(S_{t+1} - S_t \mid X_t\right) = 0$, so the score process is an additive functional; in particular, it is a martingale with increment $\frac{d}{d\theta} \log \psi(Z_j | X_{j-1}, \theta_1)$ under the θ_1 probability model. So

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

where $V = E\left(\left[\frac{d}{d\theta} \log \psi(Z_{t+1} | X_t, \theta_1) \right]^2 \right)$.

Proof. This follows directly from equation (6.3) and Proposition 3.2.

Theorem 6.2 justifies using the martingale central limit theorem Proposition 3.2 to characterize the large sample behavior of the score process. The resulting central limit approximation yields a large sample characterization of the maximum likelihood estimator of θ in a Markov setting. With additional regularity conditions, the following result is typical:

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

where θ_t maximizes the log likelihood function $L_t(\theta)$. This analysis has a direct extension to the case where θ is a vector rather than a scalar. This kind of result motivates interpreting the covariance of the martingale increment of the score process

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

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

6.6.1. Nuisance Parameters

Consider extending the notion of Fisher information to the following situation. There is a vector θ of unknown parameters. We mainly want information about one component of the parameter vector,

although to get that information we have to estimate other components too. Call the first component of θ the 'parameters of interest' θ_o and call the other component the 'nuisance parameters' ϑ_o. Suppose that the likelihood is parameterized on an open set Θ in a finite dimensional Euclidean space and that the true parameter vector is $(\theta_o, \vartheta_o) \in \Theta$. Write the multivariate score process as

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

where $\{S_{t+1} : t = 0, 1, \ldots\}$ is the partial derivative of the log-likelihood with respect to θ and $\{\tilde{S}_{t+1} : t = 0, 1, \ldots\}$ is the partial derivative with respect to ϑ. Estimating ϑ_o simultaneously with θ_o is more difficult than estimating θ_o when ϑ_o is known. Fisher's measure of information corrects for the additional challenge of estimating ϑ_o simultaneously in order to make inferences about θ_o. In particular, Fisher's measure of information about θ is the inverse of the $(1,1)$ component of the covariance matrix partitioned conformably with θ, ϑ:

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

To represent this inverse in a useful way, compute the population regression

$$S_{t+1} - S_t = \beta'(\tilde{S}_{t+1} - \tilde{S}_t) + U_{t+1}, \tag{6.4}$$

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

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

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

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

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

$$V_{1,1}^{-1} = E(U_{t+1}^2)^{-1}.$$

The population regression equation (6.4) implies that the Fisher information measure $E[(U_{t+1})^2]$ is no larger than $E[(S_{t+1} - S_t)^2]$. Thus,

$$E[(U_{t+1})^2] \leq E[(S_{t+1} - S_t)^2]. \tag{6.5}$$

This inequality asserts that the likelihood function contains more information about θ when ϑ is known to be ϑ_o than when θ and ϑ are both unknown.

Inequality (6.5) offers reasons to be cautious when interpreting calibration exercises in macroeconomics. Pretending that you know ϑ leads you to overstate the information in the sample about the parameter θ.

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 $\{U_{t+1} : t = 0, 1, \ldots\}$ is itself an additive martingale.

---

[1] This inequality analysis extends pointwise in θ. A pointwise use of the Law of Large Numbers is typically not sufficient to justify statistical consistency.