Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario

Identificación de Sistemas

Título: Estima de Máxima Verosimilitud
Eficiencia Estadística de un Estimador – Cota de Cramer-Rao
Autor: Dr. Juan Carlos Gómez **Setiembre de 2003**

Estimadores y principio de máxima verosimilitud

Dentro del marco de la identificación de sistemas y la estimación de parámetros, se trata con el problema de extraer información a partir de observaciones que pueden ser no confiables ya que pueden estar corrompidas por ruido. Cuando las perturbaciones son modeladas como procesos aleatorios, las observaciones pueden entonces describirse como realizaciones de variables aleatorias Por lo tanto se pueden describir las observaciones con la variable aleatoria $y^N = (y(1), y(2), \ldots, y(N))$ que toma valores en R^N. Se asume que y^N tiene una función de densidad de probabilidad (PDF):

$$f(\theta; x_1, x_2, \ldots, x_N) = f_y(\theta; x_N) \tag{1}$$

La probabilidad de que se hayan dado las observaciones resulta entonces:

$$P(y^N \in A) = \int_{x^N \in A} f_y(\theta; x^N) dx^N \tag{2}$$

Donde $\theta \in R^d$ es un vector de parámetros que describe propiedades de la variable observada. Este vector supuestamente no se conoce, y el propósito de las observaciones es estimarlo a partir de y^N. Esto es realizado mediante un **estimador** $\hat{\theta}(y^N)$ que es una función $R^N \to R^d$. Si el valor observado de y^N es y_*^N, el resultado de la estimación será $\hat{\theta}_* = \hat{\theta}(y_*^N)$.

Existen muchas funciones de estimación. En particular una que maximiza la probabilidad de que el evento observado haya sido generado por un grupo de datos es el denominado **Estimador de Máxima Verosimilitud** introducido por Fisher (1912). La misma se puede definir como sigue: la función densidad de probabilidad conjunta para el vector aleatorio a ser observado está dada por (1). La probabilidad de que la realización (observación) tome el valor y_*^N es entonces proporcional a

$$f_y(\theta; y_*^N)$$

Notar que ésta es una función determinística de θ una vez que un valor numérico de y_*^N es introducido. Esta función es llamada **Función de Verosimilitud** y refleja la "probabilidad" de que el evento observado ocurra. Sería razonable entonces, seleccionar un estimador de θ de forma tal que los eventos observados resulten "tan probables como sea posible". Esto es, buscamos

$$\hat{\theta}_{ML}(y_*^N) = \arg\max_\theta f_y(\theta; y_*^N) \tag{3}$$

donde la maximización es realizada para un y_*^N fijo. Esta función es conocida como **Estimador de Máxima Verosimilitud (MLE: Maximum Likelihood Estimator).**

<u>Ejemplo:</u> Estima de una Constante

Sean *y(i), i = 1,...,N*, mediciones ruidosas de una constante. Asumimos además, que *y(i)* son variables aleatorias independientes, con una distribución normal, con medias θ_0 desconocida (independiente de *i*), y varianzas conocidas λ_i, entonces podemos escribir:

$$y(i) \in N(\theta_0, \lambda_i) \tag{4}$$

Un estimador muy común de θ_0 es la media muestral (SM: Sample Mean):

$$\hat{\theta}_{SM}(y^N) = \frac{1}{N}\sum_{i=1}^{N} y(i) \tag{5}$$

la cual se prueba, es la estima de mínimos cuadrados.

Para calcular el MLE (Estimador de Máxima Verosimilitud), comenzamos por determinar la PDF conjunta (1) para las observaciones. Sabiendo que la PDF de *y(i)* es

$$\frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(x_i-\theta)^2}{2\lambda_i}\right] \tag{6}$$

y como las *y(i)* son independientes, tenemos que

$$f_y(\theta; x^N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(x_i-\theta)^2}{2\lambda_i}\right] \tag{7}$$

que es la función de verosimilitud. La maximización de la función de verosimilitud es equivalente a la maximización su logaritmo, es decir

$$\begin{aligned} \hat{\theta}_{ML}(y^N) &= \arg\max_\theta \log f_y(\theta; y^N) \\ &= \arg\max_\theta \left\{ -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\frac{1}{2}\log\lambda_i - \frac{1}{2}\sum_{i=1}^{N}\frac{(y(i)-\theta)^2}{\lambda_i} \right\} \end{aligned} \tag{8}$$

Como los dos primeros términos son independientes de θ, podemos maximizar el ultimo termino

$$\hat{\theta}_{ML}(y^N) = \arg\max_\theta \left\{ -\frac{1}{2}\sum_{i=1}^{N}\frac{(y(i)-\theta)^2}{\lambda_i} \right\} \tag{9}$$

Luego, derivando e igualando a cero, obtendremos su máximo. Es decir

$$\frac{\partial}{\partial\theta}\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y(i)-\theta)^2}{\lambda_i} \right] = 0$$

De donde resulta

$$\sum_{i=1}^{N}\frac{y(i)-\theta}{\lambda_i} = 0$$

Despejando θ de la expresión anterior, obtenemos la estima de máxima verosimilitud para el caso propuesto, la cual resulta:

$$\hat{\theta}_{ML}\left(y^N\right)=\frac{1}{\sum_{i=1}^{N}\left(1/\lambda_i\right)}\sum_{i=1}^{N}\frac{y(i)}{\lambda_i} \tag{10}$$

Del resultado anterior podemos notar que:

- Aquellas mediciones que posean mayor varianza λ_i, es decir que tienen mayor error de medición, son menos "pesadas" a la hora del cálculo de la estima.
- Para este ejemplo, el resultado obtenido es el mismo que se obtendría minimizando la varianza del error de predicción. Es decir, que para este caso la Estima de Máxima Verosimilitud resulta igual a la Mejor estima Lineal No Desviada (BLUE: Best Linear Unbiased Estimator).

Propiedades Asintóticas de la MLE

A menudo es dificultoso calcular exactamente las propiedades estadísticas de un estimador, como por ejemplo, la matriz de covarianza del error de estimación. En lugar de esto, se suele realizar un análisis asintótico cuando el número de muestras N, tiende a infinito. Resultados clásicos para MLE, en el caso de observaciones independientes, fueron obtenidos por Wald (1949) y Cramér (1946) y se presentan a continuación.

Suponga que las variables aleatorias *{y(i)}* son independientes y se encuentran idénticamente distribuidas, de modo que

$$f_y\left(\theta, x_1, \ldots, x_N\right)=\prod_{i=1}^{N} f_{y(i)}\left(\theta, x_i\right)$$

Supóngase también, que la distribución de y^N esta dada por $f_y\left(\theta_0 ; x^N\right)$ para algún valor de θ_0. Entonces la variable aleatoria $\hat{\theta}_{ML}\left(y^N\right)$ tiende a θ_0 con probabilidad 1 cuando N tiende a infinito, y la variable aleatoria

$$\sqrt{N}\left(\hat{\theta}_{ML}\left(y^N\right)-\theta_0\right)$$

converge en distribución a una distribución normal, con media cero y matriz de covarianza dada por la cota inferior de Cramér-Rao (M^{-1}), donde M es:

$$M=-E\left[\frac{d^2}{d\theta^2}\log f_y\left(\theta; y^N\right)\right]_{\theta=\theta_0}$$

Eficiencia Estadística de un Estimador - Desigualdad de Cramér-Rao

La calidad de un estimador puede ser cuantificada mediante la matriz de covarianza correspondiente a la estima definida por:

$$P=E\left[\left[\hat{\theta}\left(y^N\right)-\theta_0\right]\left[\hat{\theta}\left(y^N\right)-\theta_0\right]^T\right] \tag{11}$$

Donde θ_0 denota el "verdadero valor" de θ, y puede ser evaluada si asumimos que la PDF de y^N es $f_y\left(\theta_0; y^N\right)$.

Estamos interesados en un estimador tal que haga P lo más pequeña posible. Es interesante por lo tanto buscar el límite inferior de P que se puede obtener con estimadores no desviados, ésta es la denominada **cota de Cramér-Rao**.

Lema:

Sea $\hat{\theta}(y^N)$ un estimador de θ tal que $E\left[\hat{\theta}(y^N)\right]=\theta_0$. Asumimos que la PDF de y^N es $f_y\left(\theta_0;y^N\right)$, válida para toda θ_0, y suponemos que y^N toma valores en un sub-conjunto de R^N cuyo límite no depende de θ.

Entonces:

$$P=E\left[\left[\hat{\theta}\left(y^N\right)-\theta_0\right]\left[\hat{\theta}\left(y^N\right)-\theta_0\right]^T\right]\geq M^{-1} \tag{12}$$

Donde *M* es la matriz de información de Fischer definida de la siguiente forma:

$$M=E\left\{\left[\frac{d}{d\theta}\log f_y\left(\theta;y^N\right)\right]\left[\frac{d}{d\theta}\log f_y\left(\theta;y^N\right)\right]^T\Bigg|_{\theta_0}\right\}=-E\frac{d^2}{d\theta^2}\log f_y\left(\theta;y^N\right)\Bigg|_{\theta_0} \tag{13}$$

Como θ es un vector d-dimensional, $\frac{d}{d\theta}\log f_y\left(\theta;y^N\right)$ es un vector columna d-dimensional y el Hessiano $\frac{d^2}{d\theta^2}\log f_y\left(\theta;y^N\right)$es una matriz de *dxd*. Esta matriz *M* es conocida como la **matriz de información de Fischer**. Debemos notar que evaluar *M* normalmente requiere el conocimiento de θ_0, por lo que el valor exacto de *M* no estará disponible.

Demostración:

Dado que asumimos $E\left[\hat{\theta}(y^N)\right]=\theta_0$ (estimador no desviado) podemos escribir:

$$\theta_0=\int_{R^N}\hat{\theta}\left(x^N\right)f_y\left(\theta;x^N\right)dx^N \tag{14}$$

Por definición:

$$I=\int_{R^N}f_y\left(\theta;x^N\right)dx^N \tag{15}$$

Diferenciando estas expresiones respecto a θ_0:

$$\begin{aligned} I_{dxd}&=\int_{R^N}\hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0}f_y\left(\theta_0;x^N\right)\right]^T dx^N=\\ &=\int_{R^N}\hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T f_y\left(\theta;x^N\right)dx^N=\\ &=E\left[\hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T\right] \end{aligned} \tag{16}$$

$$\begin{aligned} 0=\int_{R^N}\left[\frac{d}{d\theta_0}f_y\left(\theta_0;x^N\right)\right]^T dx^N&=\int_{R^N}\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T f_y\left(\theta_0;x^N\right)dx^N=\\ &=E\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T \end{aligned} \tag{17}$$

Multiplicando por θ_0 la última expresión y restando a la anterior, se obtiene:

$$E\left[\left[\hat{\theta}\left(y^N\right)-\theta_0\right]\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T\right]=I \tag{18}$$

denominando $$\alpha=\hat{\theta}\left(y^N\right)-\theta_0,\quad \beta=\frac{d}{d\theta_0}\log f_y\left(\theta_0,y^N\right) \tag{19}$$

tenemos, $E\alpha\beta^T=I$

Entonces, por construcción

$$E\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}^T=\begin{bmatrix}E\alpha\alpha^T & I\\ I & E\beta\beta^T\end{bmatrix}\geq 0 \tag{20}$$

es semidefinida positiva.

Luego,

$$E\left[\alpha\alpha^T\right]\geq E\left[\beta\beta^T\right]^{-1} \tag{21}$$

Finalmente se obtiene:

$$E\left[\left[\hat{\theta}\left(y^N\right)-\theta_0\right]\left[\hat{\theta}\left(y^N\right)-\theta_0\right]^T\right]\geq E\left[\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0,y^N\right)\right]\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0,y^N\right)\right]^T\right]^{-1} \tag{22}$$

O sea según (13)

$$E\left[\left[\hat{\theta}\left(y^N\right)-\theta_0\right]\left[\hat{\theta}\left(y^N\right)-\theta_0\right]^T\right]\geq M^{-1} \tag{23}$$

que es la varianza mínima que puede obtenerse mediante un estimador no desviado.

Para probar la igualdad de los últimos dos términos de (13) derivando (17)

$$0=\int_{R^N}\left[\frac{d^2}{d{\theta_0}^2}f_y\left(\theta_0;x^N\right)\right]f_y\left(\theta_0,x^N\right)dx^N+\int_{R^N}\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]\left[\frac{d}{d\theta_0}\log f_y\left(\theta_0;x^N\right)\right]^T f_y\left(\theta_0;x^N\right)dx^N$$

obteniendo así (13).

Definición: Un estimador (no desviado) se dice que es **eficiente** desde el punto de vista estadístico si la covarianza de la estima alcanza la cota de Cramér-Rao.