Modelado y Simulación II

Enrique Baeyens
enrbae@eis.uva.es

Departamento de Ingeniería de Sistemas y Automática
Universidad de Valladolid
Escuela de Ingenierías Industriales
Paseo del Cauce 59
47011 Valladolid

Índice

1. Introducción a la teoría de la estimación

En teoría de la estimación se pretende obtener el modelo matemático de un un conjunto de observaciones de una variable aleatoria. El modelo más natural es la distribución de probabilidad de la variable aleatoria que generó los datos.

Hay dos metodologías de estimación:

1. Estimación clásica.
2. Estimación bayesiana.

En estimación clásica, la función de densidad de probabilidad de la variable aleatoria observada se considera que viene modelada por una función matemática que queda completamente representada por un número finito de parámetros desconocidos pero constantes, representados por el vector θ.

Por el contrario, en estimación bayesiana, la variable aleatoria observada es una función matemática de otra variable aleatoria cuya función de distribución de probabilidad se quiere estimar a partir del conjunto de observaciones. Para ello se determinará la distribución de probabilidad condicionada a las observaciones y a partir de ella se obtendrá información de la variable aleatoria a estimar.

1.1. Notación y resultados básicos

Sea X una variable aleatoria definida sobre el espacio de probabilidad $(\Omega_x, \mathscr{F}_x, P_x)$. Su función de densidad de probabilidad se denota $p(x)$ y se define como:

$$p(x) = \frac{d}{dx}\mathbf{Prob}[X \leq x]$$

Su esperanza se denota $\mathbf{E}(X)$ y viene definida como:

$$\mathbf{E}(X) = \int_{x\in\Omega_x} x p(x) dx$$

Sean X e Y dos variables aleatorias definidas sobre el espacio conjunto de probabilidad $(\Omega_x \times \Omega_y, \mathscr{F}_{xy}, P_{xy})$. La función de densidad conjunta se denota $p(x,y)$ y se define

$$p(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}\mathbf{Prob}(X \leq x, Y \leq y)$$

La covarianza cruzada de X e Y se denota $\mathbf{Cov}(X,Y)$ y viene definida por la expresión:

$$\mathbf{Cov}(X,Y) = \mathbf{E}\left[(X-\mathbf{E}(X))(Y-\mathbf{E}(Y))^T\right] = \int_{x\in\Omega_x}\int_{y\in\Omega_y}(x-\mathbf{E}(x))(y-\mathbf{E}(y))^T p(x,y)dydx$$

La covarianza de X se denota $\mathbf{Cov}(X)$ y se define

$$\mathbf{Cov}(X) = \mathbf{Cov}(X,X) = \int_{x\in\Omega_x}(x-\mathbf{E}(x))(x-\mathbf{E}(x))^T p(x)dx$$

La función de densidad de probabilidad de la variable aleatoria X condicionada a la variable aleatoria Y se denota $p(x|y)$ y se obtiene aplicando el teorema de Bayes

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

La esperanza de la variable aleatoria X condicionada a Y se denota $\mathbf{E}(X|Y)$ y se obtiene mediante la expresión

$$\mathbf{E}(X|Y) = \int_{x\in\Omega_x} xp(x|y)dx$$

El teorema del límite central. Sea $\{X_k\}_{k=1}^n$ una secuencia de variables aleatorias independientes entre sí e identicamente distribuidas con esperanza μ y covarianza C, entonces la distribución de probabilidad de la variable aleatoria promedio $S_n = \sum_{i=1}^n X_i/n$ converge asintóticamente a una gaussiana de media μ y covarianza C/n.

La ley de los grandes números. Sea $\{X_k\}_{k=1}^n$ una secuencia de observaciones de una variable aleatoria X y sea $g(X)$ una función de la variable aleatoria X, entonces $g(\bar{X}_n)$, con $\bar{X}_n = g(\sum_{i=1}^n X_i/n)$, converge en probabilidad a la esperanza $\mathbf{E}(g(X))$, es decir para todo $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{Prob}\left[\left|g(\bar{X}_n) - \mathbf{E}(g(X))\right| > \epsilon\right] = 0$$

1.2. Estimación clásica

Sea $\{x_k\}_{k=1}^N$ una secuencia de observaciones, con x_k el dato registrado en el instante de tiempo t_k. Las observaciones son una realización de una variable aleatoria vectorial $x = \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}^T$ con función de densidad de probabilidad $p(x)$ desconocida que se desea estimar a partir de las observaciones.

En teoría de estimación clásica, la función de densidad de probabilidad buscada pertenece a una familia de funciones de densidad de probabilidad parametrizadas por un vector θ de parámetros constantes $\theta \in \mathbb{R}^n$ pero desconocidos

$$\{f(x;\theta) : \theta \in \mathbf{\Theta}\}$$

La función de densidad que generó los datos debe ser determinada a partir del conjunto de observaciones $x \in \mathbb{R}^N$.

Formulación del problema: Determinar una función que proporcione una estimación del vector de parámetros $\theta \in \mathbf{\Theta}$ a partir del vector de observaciones x,

$$\hat{\theta} = g(x) \tag{1}$$

de forma que $p(x,\hat{\theta})$ aproxime la función de densidad de probabilidad que generó el vector de observaciones $x \in \mathbb{R}$.

El estimador de parámetros $\hat{\theta}$ es una variable aleatoria vectorial, ya que depende de las observaciones y toma un valor diferente para cada realización.

Medida de la bondad del estimador: Una medida natural de la calidad de un estimador es su error cuadrático medio, definido de la siguiente forma

$$\mathbf{MSE}(\hat{\theta}) = \mathbf{E}[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)] = \mathrm{Tr}\mathbf{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T].$$

Sin embargo, el error cuadrático medio no es adecuado para obtener estimadores ya no solo depende de los datos, sino también del verdadero valor del vector de los parámetros que no es conocido.

El error cuadrático medio puede expresarse de la siguiente manera

$$\begin{aligned}\mathbf{MSE}(\hat{\theta}) &= \mathbf{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \\ &= \mathbf{E}[(\hat{\theta}-\mathbf{E}[\hat{\theta}]+\mathbf{E}[\hat{\theta}]-\theta)(\hat{\theta}-\mathbf{E}[\hat{\theta}]+\mathbf{E}[\hat{\theta}]-\theta)^T] \\ &= \mathbf{E}[(\hat{\theta}-\mathbf{E}(\hat{\theta}))(\hat{\theta}-\mathbf{E}(\hat{\theta}))]^T + (\mathbf{E}[\hat{\theta}]-\theta)(\mathbf{E}[\hat{\theta}]-\theta)^T + \\ &\quad 2\mathbf{E}[(\hat{\theta}-\mathbf{E}(\hat{\theta}))](\mathbf{E}[\hat{\theta}]-\theta)^T \\ &= \mathbf{Cov}(\hat{\theta}) + (\mathbf{B}(\hat{\theta}))(\mathbf{B}(\hat{\theta}))^T + 2\underbrace{(E[\hat{\theta}]-E[\hat{\theta}])}_{=0}B(\hat{\theta})\end{aligned}$$

siendo $\mathbf{Cov}(\hat{\theta})$ la covarianza del estimador y $\mathbf{B}(\hat{\theta})$ el sesgo. La varianza del estimador depende sólo de los datos, mientras que el sesgo depende también del verdadero valor del vector de parámetros. Si el estimador tiene sesgo nulo, entonces el error cuadrático medio coincide con la varianza del estimador.

Por tanto, un estimador realizable es el estimador insesgado de mínima varianza. Sin embargo este estimador no siempre existe, o puede ser difícil de obtener.

1.3. El estimador insesgado y eficiente

El estimador insesgado y eficiente minimiza el error cuadrático medio de estimación. Siempre que exista, esta es la mejor opción. Sin embargo este estimador puede no existir o ser muy difícil de calcular. Para determinar este estimador, comenzaremos calculando una cota inferior de la covarianza de un estimador insesgado, la cota de Cramer-Rao. El estimador insesgado eficiente es aquel que alcanza la cota de Cramer-Rao.

Función de verosimilitud. La función de densidad de probabilidad $p(x;\theta)$ particularizada para el vector de observaciones x es función solo del vector de parámetros y se denomina función de verosimilitud.

$$L(\theta) := p(x;\theta).$$

El logaritmo neperiano de la función de verosimilitud se denomina función logarítmica de verosimilitud.

Matriz de información de Fisher $\mathbf{I}(\theta)$. Se define de la siguiente manera

$$\mathbf{I}_{ij}(\theta) = -\mathbf{E}\left[\frac{\partial^2 \ln p(x;\theta)}{\partial\theta_i \partial\theta_j}\right] = \mathbf{E}\left[\frac{\partial \ln p(x;\theta)}{\partial\theta_i}\frac{\partial \ln p(x;\theta)}{\partial\theta_j}\right]$$

Para comprobar la igualdad de ambas esperanzas

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}\ln p(x;\theta) &= \frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_j}p(x;\theta)\\
\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\ln p(x;\theta) &= \frac{\partial}{\partial\theta_i}\left(\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_j}p(x;\theta)\right)\\
&= -\frac{1}{(p(x;\theta))^2}\frac{\partial}{\partial\theta_i}p(x;\theta)\frac{\partial}{\partial\theta_j}p(x;\theta)+\frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)\\
&= -\frac{\partial}{\partial\theta_i}\ln p(x;\theta)\frac{\partial}{\partial\theta_j}\ln p(x;\theta)+\frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)
\end{aligned}$$

Tomando ahora esperanzas matemáticas a ambos lados de la igualdad

$$-\mathbf{E}\left[\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\ln p(x;\theta)\right]=\mathbf{E}\left[\frac{\partial}{\partial\theta_i}\ln p(x;\theta)\frac{\partial}{\partial\theta_j}\ln p(x;\theta)\right]$$

ya que

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{p(x;\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)\right] &= \int_{x\in\Omega}\frac{p(x;\theta)}{p(x;\theta)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(x;\theta)dx\\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int_{x\in\Omega}p(x;\theta)dx\\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}1=0
\end{aligned}$$

1.3.1. La cota inferior de Cramer-Rao

Establece una cota inferior de la matriz de covarianza que puede alcanzarse para cualquier estimador insesgado. Esta cota inferior viene definida en términos de la matriz de información de Fisher. Para cualquier estimador insesgado $\hat{\theta}$,

$$\mathbf{Cov}(\hat{\theta})\geq\mathbf{I}^{-1}(\theta).$$

Obtención de la cota de Cramer-Rao. Sea $\hat{\theta}=g(x)$ un estimador insesgado del vector de parámetros θ entonces $\mathbf{E}[\hat{\theta}]=\theta$. La obtención se realiza haciendo uso de la desigualdad de Cauchy-Schwarz[1]

$$\mathbf{E}(AA^T)\geq\mathbf{E}(AB^T)(\mathbf{E}(BB^T))^{-1}\mathbf{E}(BA^T)$$

[1]Puede demostrarse utilizando la fórmula del complemento de Schur para $\mathbf{E}(ZZ^T)$ siendo $Z=[A^T\ B^T]^T$.

Eligiendo $A = \hat{\theta} - \theta$ y $B = \frac{\partial}{\partial x} \ln p(x;\theta)$

$$\begin{aligned}
\mathbf{E}(AB^T) &= \mathbf{E}\left[(\hat{\theta}(x)-\theta)\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)^T\right] \\
&= \int_{x\in\Omega}(\hat{\theta}(x)-\theta)\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)^T p(x;\theta)dx \\
&= \int_{x\in\Omega}\hat{\theta}(x)\left(\frac{\partial}{\partial\theta}p(x;\theta)\right)^T - \theta\int_{x\in\Omega}\left(\frac{\partial}{\partial\theta}p(x;\theta)\right)^T dx \\
&= \frac{\partial}{\partial\theta}\left(\int_{x\in\Omega}\hat{\theta}(x)p(x;\theta)dx\right) - \theta\left(\frac{\partial}{\partial\theta}\int_{x\in\Omega}p(x;\theta)dx\right)^T \\
&= \frac{\partial}{\partial\theta}\mathbf{E}(\hat{\theta}) - \theta\left(\frac{\partial}{\partial\theta}1\right)^T = I = (\mathbf{E}(BA^T))^T
\end{aligned}$$

$$\mathbf{E}(BB^T) = \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)^T\right] = \mathbf{I}(\theta)$$

Finalmente, sustituyendo en la desigualdad de Cauchy-Schwarz se obtiene la expresión buscada

$$\mathbf{Cov}(\hat{\theta}) \geq \mathbf{I}^{-1}(\theta)$$

El estimador de insesgado de mínima varianza es aquel que alcanza el valor de la cota de Cramer-Rao.

1.3.2. Obtención del estimador insesgado y eficiente mediante el método de la cota de Cramer-Rao

La cota de Cramer-Rao establece un método constructivo para determinar el estimador insesgado de mínima varianza. Si el gradiente de la función logarítmica de versimilitud puede expresarse de la forma siguiente:

$$\frac{\partial \ln L(\theta)}{\partial\theta} = \frac{\partial \ln p(x;\theta)}{\partial\theta} = \mathbf{I}(\theta)(g(x)-\theta)$$

entonces el estimador insesgado de mínima varianza es:

$$\hat{\theta}_{\text{MVU}} = g(x)$$

y su matriz de covarianzas es

$$\mathbf{Cov}(\hat{\theta}_{\text{MVU}}) = \mathbf{I}^{-1}(\theta)$$

En primer lugar, comprobaremos que el estimador $\hat{\theta} = g(x)$ es insesgado. Nótese que

$$\mathbf{E}(g(x)-\theta) = \mathbf{I}^{-1}(\theta)\,\mathbf{E}\left[\frac{\partial \ln p(x;\theta)}{\partial\theta}\right]$$

pero

$$\begin{aligned}\mathbf{E}\left[\frac{\partial \ln p(x;\theta)}{\partial \theta}\right] &= \int_{x\in\Omega_x} p(x;\theta)\frac{\partial \ln p(x;\theta)}{\partial \theta}dx \\ &= \int_{x\in\Omega_x} p(x;\theta)\frac{1}{p(x;\theta)}\frac{\partial p(x;\theta)}{\partial \theta}dx \\ &= \frac{\partial}{\partial \theta}\int_{x\in\Omega_x} p(x;\theta)dx = \frac{\partial}{\partial \theta}1 = 0\end{aligned}$$

lo que garantiza que $B(\hat{\theta}) = \mathbf{E}[g(x) - \theta] = 0$.

Para comprobar que el estimador es además eficiente, calculamos la matriz de información de Fisher:

$$\begin{aligned}\mathbf{I}(\theta) &= \mathbf{E}\left[\left(\frac{\partial \ln p(x;\theta)}{\partial \theta}\right)\left(\frac{\partial \ln p(x;\theta)}{\partial \theta}\right)^T\right] \\ &= \mathbf{I}(\theta)\mathbf{E}[(g(x)-\theta)(g(x)-\theta)^T]\mathbf{I}(\theta)\end{aligned}$$

de donde tomando como estimador $\hat{\theta} = g(x)$

$$\mathbf{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] = \mathbf{I}^{-1}(\theta)$$

la covarianza del estimador iguala a la cota de Cramer-Rao, que es el menor valor que puede tomar y esto garantiza que el estimador $\hat{\theta} = g(x)$ es eficiente.

Ejemplo. Sea el modelo estático

$$x_k = \theta + w_k, \quad k = 1,\ldots,N,\ \theta \in \mathbb{R},$$

siendo w_k una variable aleatoria gaussiana de media cero y varianza σ^2. Entonces, sea $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}$ el vector de observaciones, cuya función de verosimilitud es

$$\begin{aligned}L(\theta) &= p(x;\theta) \\ &= \prod_{k=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x_k-\theta)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\theta)^2\right]\end{aligned}$$

ya que $\{w_k\}$ es un ruido blanco.

La función logarítmica de verosimilitud es

$$\ln L(\theta) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\theta)^2$$

y sus derivadas respecto al parámetro desconocido θ son

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta} &= \frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{k=1}^{N}x_k - \theta\right) \\ \frac{\partial^2 L(\theta)}{\partial \theta^2} &= -\frac{N}{\sigma^2}\end{aligned}$$

Ahora igualando términos:

$$\begin{aligned}\frac{\partial \ln p(x;\theta)}{\partial \theta} &= \frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{k=1}^{N} x_k - \theta\right) \\ &= I(\theta)(g(x)-\theta)\end{aligned}$$

De donde el estimador insesgado de mínima varianza es

$$\hat{\theta} = g(x) = \frac{1}{N}\sum_{k=1}^{N} x_k$$

cuya varianza es

$$\mathbf{Cov}(\hat{\theta}) = \mathbf{I}^{-1}(\theta) = \frac{\sigma^2}{N}$$

Ejemplo (Modelo lineal en los parámetros) Considérese el modelo matemático

$$b(x) = A(x)\theta + w$$

Siendo $A(x)$ y $b(x)$ transformaciones del vector de observación x y w una variable aleatoria vectorial gaussiana de media cero y matriz de covarianza C. El vector de parámetros que se desea estimar es θ. Las dimensiones de los diferentes elementos del modelo son:

$$\begin{aligned}A(x) &\in \mathbb{R}^{m\times n} \\ b(x) &\in \mathbb{R}^{m\times 1} \\ \theta &\in \mathbb{R}^{n\times 1} \\ w &\in \mathbb{R}^{m\times 1}\end{aligned}$$

En general $m \ll n$, ya que las observaciones son más numerosas que la dimensión del vector de parámetros.

Para obtener el estimador insesgado de mínima varianza calculamos la función de densidad de probabilidad de las observaciones

$$p(x;\theta) = \frac{1}{\sqrt{(2\pi)^m \det(C)}} \exp\left[-\frac{1}{2}(b(x)-A(x)\theta)^T C^{-1}(b(x)-A(x)\theta)\right]$$

Entonces haciendo uso de la cota de Cramer-Rao, el estimador insesgado de mínima varianza será

$$\hat{\theta} = g(x)$$

si

$$\frac{\partial}{\partial \theta}\ln p(x;\theta) = I(x)(g(x)-\theta)$$

Derivando la función logarítmica de densidad de probabilidad respecto al vector de parámetros:

$$\begin{aligned}\frac{\partial}{\partial \theta}\ln p(x;\theta) &= \frac{\partial}{\partial \theta}\left[-\ln((2\pi)^m \det(C))^{1/2} - \frac{1}{2}(b(x)-A(x)\theta)^T C^{-1}(b(x)-A(x)\theta)\right] \\ &= A^T(x)C^{-1}(b(x)-A(x)\theta) \\ &= A^T(x)C^{-1}A(x)\left[(A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}b(x)-\theta\right]\end{aligned}$$

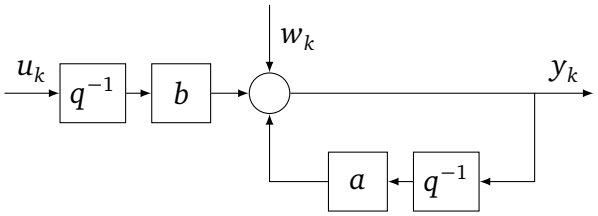

Figura 1: Diagrama de bloques del sistema dinámico del ejemplo

Por tanto, el estimador insesgado de mínima varianza es:

$$\hat{\theta} = (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}b(x)$$

y su covarianza es:

$$\mathbf{Cov}(\hat{\theta}) = (A^T(x)C^{-1}A(x))^{-1}$$

Un caso particular es el célebre modelo lineal con ruido blanco

$$x = A\theta + w$$

con $C = \mathbf{Cov}(w) = \sigma^2 I_{N\times N}$. Para este caso $A(x) = A$, $b(x) = x$, y $m = N$. El estimador insesgado de mínima varianza del vector de parámetros es:

$$\hat{\theta} = (A^T A)^{-1}A^T x$$

y su matriz de covarianza es:

$$\mathbf{Cov}(\theta) = (A^T A)^{-1}$$

Ejemplo (Sistema dinámico lineal) Sea el sistema dinámico lineal representado por el diagrama de bloques de la figura 1 y caracterizado por la siguiente ecuación en diferencias

$$y_k = a y_{k-1} + b u_{k-1} + w_k$$

donde los parámetros a y b son constantes pero desconocidos y w_k es un ruido blanco gaussiano con esperanza $\mathbf{E}(w_k) = 0$ y autocovarianza $\mathbf{E}(w_k w_j) = \sigma^2\delta_{k,j}$. Las observaciones se recopilan en el vector x definido de la siguiente manera

$$x = \begin{bmatrix} x_1^T & x_2^T & \cdots & x_N^T \end{bmatrix}^T,\ x_k = \begin{bmatrix} u_k & y_k \end{bmatrix}^T$$

Se desea obtener estimadores de los parámetros a y b a partir del vector de observaciones. El modelo que representa el problema de estimación es lineal en los parámetros y corresponde a la ecuación

$$b(x) = A(x)\theta + w$$

siendo

$$b(x) = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix},\ A(x) = \begin{bmatrix} y_1 & u_1 \\ y_2 & u_2 \\ \vdots & \vdots \\ y_{N-1} & u_{N-1} \end{bmatrix},\ \theta = \begin{bmatrix} a \\ b \end{bmatrix},\ w = \begin{bmatrix} w_2 \\ w_3 \\ \vdots \\ w_N \end{bmatrix}$$

El vector w satisface $\mathbf{E}(w) = 0$, $\mathbf{E}(ww^T) = \sigma^2 I$, siendo I la matriz identidad de orden $N-1$. Entonces, el vector de estimadores insesgados y eficientes se puede obtener aplicando la fórmula obtenida en el ejemplo anterior

$$\hat{\theta} = (A^T(x)A(x))^{-1}A^T(x)b(x)$$

y su matriz de covarianza es

$$\mathbf{Cov}(\hat{\theta}) = (A^T(x)A(x))^{-1}$$

1.3.3. Método general de obtención de estimadores insesgados y eficientes

Cuando la cota de Cramer-Rao no puede ser utilizada para obtener el estimador insesgado de mínima varianza es posible que pueda hacerse uso todavía de la factorización de Neyman-Fisher y del teorema de Rao-Blackwell-Lehmann-Scheffe.

Comencemos introduciendo el concepto de estadístico suficiente. Sea $\{x_k\}_{k=1}^N$ un conjunto de observaciones de una variable aleatoria con función de densidad de probabilidad $p(x;\theta)$, siendo θ un vector de parámetros desconocido que se desea estimar.

Concepto de estadístico. Un estadístico $T(x)$ es una función de los datos observados.

Estadístico suficiente. Un estadístico $T(x)$ es suficiente para el vector de parámetros θ si la función de densidad de probabilidad condicionada a la observación del estadístico $p(x|T(x) = T_0;\theta)$ no depende del vector de parámetros θ. Es decir,

$$p(x|T(x) = T_0;\theta) = w(x)\delta(T(x) - T_0)$$

Un estadístico suficiente contiene toda la información necesaria para estimar el vector de parámetros θ, por tanto una vez que el estadístico ha sido observado, la densidad de probabilidad condicional no tiene información adicional sobre el vector de parámetros.

La definición anterior no suele ser apropiada para comprobar si un estadístico es suficiente ya que calcular la función de densidad de probabilidad condicionada suele ser una tarea ardua excepto en ciertos casos como cuando las variables aleatorias son gaussianas. Un método alternativo es la factorización de Neyman-Fisher.

Factorización de Neyman-Fisher. La función de densidad de probabilidad $p(x,\theta)$ puede factorizarse de la siguiente forma:

$$p(x;\theta) = g(T(x);\theta)h(x)$$

si y solo si $T(x)$ es un estadístico suficiente para el vector de parámetros θ.

Demostración de la parametrización de Neyman-Fisher Si la factorización se cumple entonces $T(x)$ es un estadístico suficiente. En efecto, aplicando el teorema de Bayes

$$\begin{aligned} p(x|T(x) = T_0;\theta) &= \frac{p(x, T(x) = T_0;\theta)}{p(T(x) = T_0;\theta)} \\ &= \frac{p(x;\theta)\delta(T(x) - T_0)}{\int p(x;\theta)\delta(T(x) - T_0)dx} \end{aligned}$$

sustituyendo la factorización de Neyman-Fisher en el numerador y denominador se obtiene

$$\begin{aligned} p(x|T(x)=T_0;\theta) &= \frac{g(T(x)=T_0,\theta)h(x)\delta(T(x)-T_0)}{g(T(x)=T_0,\theta)\int h(x)\delta(T(x)-T_0)dx} \\ &= \frac{h(x)\delta(T(x)-T_0)}{\int h(x)\delta(T(x)-T_0)dx} \end{aligned}$$

La expresión anterior no depende de θ, por tanto $T(x)$ es un estadístico suficiente.

Ahora se demostrará que si $T(x)$ es un estadístico suficiente, entonces existe la factorización de Neyman Fisher.

Debido a que $T(x)$ es un estadístico suficiente, se cumple que

$$p(x|T(x)=T_0;\theta) = w(x)\delta(T(x)-T_0)$$

y sin pérdida de generalidad puede tomarse

$$w(x) = \frac{h(x)}{\int h(x)\delta(T(x)-T_0)dx}$$

lo que garantiza que $p(x|T(x)=T_0;\theta)$ es una función de densidad con integral igual a uno,

$$\begin{aligned} \int p(x|T(x)=T_0;\theta) &= \int w(x)\delta(T(x)-T_0)dx \\ &= \frac{\int h(x)\delta(T(x)-T_0)dx}{\int h(x)\delta(T(x)=T_0)dx} = 1. \end{aligned}$$

Aplicando el teorema de Bayes

$$\begin{aligned} p(x;\theta)\delta(T(x)-T_0) &= p(x,T(x)=T_0;\theta) \\ &= p(x|T(x)=T_0;\theta)p(T(x)=T_0;\theta) \\ &= w(x)\delta(T(x)-T_0)p(T(x)=T_0;\theta) \\ &= p(T(x);\theta)w(x)\delta(T(x)-T_0) \end{aligned}$$

y sustituyendo la expresión de $w(x)$,

$$p(x;\theta)\delta(T(x)-T_0) = \frac{p(T(x);\theta)}{\int h(x)\delta(T(x)-T_0)dx}h(x)\delta(T(x)-T_0)$$

de donde puede despejarse

$$p(x;\theta) = g(T(x);\theta)h(x)$$

con

$$g(T(x);\theta) = \frac{p(T(x);\theta)}{\int h(x)\delta(T(x)-T_0)dx}$$

Un estadístico suficiente $T(x)$ es completo si existe una única función del estadístico $T(x)$ que proporciona un estimador insesgado del vector de parámetros θ.

Estadístico completo: Un estadístico suficiente $T(x)$ es completo si para cualquier función ϕ

$$\mathbf{E}[\phi(T)] = \int \phi(T)p(T;\theta)dT = 0,\ \forall\theta \quad \Rightarrow \quad \phi(T) = 0,\ \forall T.$$

Sea $g(T(x))$ una función del estadístico suficiente $T(x)$ tal que $\mathbf{E}(g(T(x)) = \theta$ entonces eligiendo $\phi(T) = \mathbf{E}(g(T) - \theta)$, la propiedad de ser completo garantiza unicidad de la función $g(T)$ que hace que el estadístico suficiente $T(x)$ sea insesgado. Por tanto, si $T(x)$ es un estadístico suficiente y completo y $g(T(x))$ es insesgado, entonces $g(T(x))$ es el único estadístico suficiente insesgado.

El teorema de Rao-Blackwell-Lehmann-Scheffe. Sea $\breve{\theta}$ un estimador insesgado del parámetro θ, entonces $\hat{\theta} = \mathbf{E}[\breve{\theta}|T(x)]$ es

1. Un estimador insesgado de θ.
2. Con varianza menor o igual que la de $\breve{\theta}$ para todo x.
3. Si $T(x)$ es un estadístico suficiente y completo entonces es el estimador insesgado de mínima varianza.

Demostración del teorema de Rao-Blackwell-Lehman-Scheffe. El estimador se obtiene como

$$\begin{aligned}\hat{\theta} &= \mathbf{E}(\breve{\theta}|T(x)) \\ &= \int \breve{\theta}(x)p(x|T(x);\theta)dx \\ &= \int \breve{\theta}(x)p(x|T(x))dx\end{aligned}$$

La función de densidad condicional $p(x|T(x);\theta) = p(x|T(x))$ no depende de θ al ser $T(x)$ un estadístico suficiente. Para demostrar que $\hat{\theta}$ es insesgado, tomamos esperanza matemática:

$$\begin{aligned}\mathbf{E}(\hat{\theta}) &= \int\int \breve{\theta}(x)p(x|T(x))dx\,p(T(x);\theta)dT \\ &= \int \breve{\theta}(x)\int p(x|T(x))p(T(x);\theta)dT\,dx \\ &= \int \breve{\theta}(x)p(x;\theta)dx \\ &= \mathbf{E}(\breve{\theta}) = \theta\end{aligned}$$

ya que $\breve{\theta}$ es insesgado.

Para demostrar que $\mathbf{Cov}(\hat{\theta}) \leq \mathbf{Cov}(\breve{\theta})$

$$\begin{aligned}\mathbf{Cov}(\breve{\theta}) &= \mathbf{E}(\breve{\theta} - \theta)(\breve{\theta} - \theta)^T \\ &= \mathbf{E}(\breve{\theta} - \hat{\theta} + \hat{\theta} - \theta)(\breve{\theta} - \hat{\theta} + \hat{\theta} - \theta)^T \\ &= \mathbf{E}(\breve{\theta} - \hat{\theta})(\breve{\theta} - \hat{\theta})^T + \mathbf{E}(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T + 2\mathbf{E}(\breve{\theta} - \hat{\theta})(\hat{\theta} - \theta)^T\end{aligned}$$

El término cruzado se anula ya que

$$\begin{aligned}\mathbf{E}(\breve{\theta}-\hat{\theta})(\hat{\theta}-\theta)^T &= \int\int(\breve{\theta}-\hat{\theta})(\hat{\theta}-\theta)^T p(x,T(x);\theta)dTdx \\ &= \int\int(\breve{\theta}-\hat{\theta})(\hat{\theta}-\theta)^T p(x|T(x);\theta)p(T(x);\theta)dTdx \\ &= \int\int(\breve{\theta}-\hat{\theta})p(x|T(x);\theta)dx(\hat{\theta}-\theta)^T p(T(x);\theta)dT\end{aligned}$$

donde la última igualdad resulta del hecho de que $\hat{\theta}$ es función unicamente del estadístico $T(x)$. La integral más interna de la expresión anterior es

$$\begin{aligned}\int(\breve{\theta}-\hat{\theta})p(x|T(x);\theta)dx &= \int\breve{\theta}p(x|T(x);\theta)dx-\hat{\theta} \\ &= \hat{\theta}-\hat{\theta}=0\end{aligned}$$

Por tanto,

$$\mathbf{E}(\breve{\theta}-\hat{\theta})(\hat{\theta}-\theta)^T=0$$

$$\begin{aligned}\mathbf{Cov}(\breve{\theta}) &= \mathbf{E}(\breve{\theta}-\hat{\theta})(\breve{\theta}-\hat{\theta})^T+\mathbf{E}(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T \\ &= \mathbf{E}(\breve{\theta}-\hat{\theta})(\breve{\theta}-\hat{\theta})^T+\mathbf{Cov}(\hat{\theta})\geq\mathbf{Cov}(\hat{\theta})\end{aligned}$$

Además, si $T(x)$ es un estadístico suficiente y completo, entonces solo hay una transformación $g(T(x))$ que proporciona un estimador insesgado. Por tanto, el estimador $\hat{\theta}$ es el estimador insesgado y eficiente.

Obtención del estimador insesgado y eficiente mediante el método general: Hay dos formas de obtener el estimador insesgado de mínima varianza a partir de un estadístico suficiente y completo $T(x)$:

1. Encontrar una función g tal que $\hat{\theta}=g(T(x))$ sea un estimador insesgado de θ, es decir tal que $\mathbf{E}[g(T(x))]=\theta$ entonces $\hat{\theta}$ es el estimador insesgado de mínima varianza.
2. Si $\breve{\theta}$ un estimador insesgado de θ, entonces $\hat{\theta}=\mathbf{E}(\breve{\theta}|T(x))=\int\breve{\theta}p(\breve{\theta}|T(x))d\breve{\theta}$ es el estimador insesgado de mínima varianza.

Ejemplo Para el modelo $x_k=\theta+w_k$ siendo w el vector columna que agrupa todas las variables aleatorias $\{w_k\}_{k=1}^N$ con esperanza cero y matriz de covarianza σ^2, la función de densidad de probabilidad $p(x;\theta)$ es:

$$p(x;\theta)=\frac{1}{\sqrt{(2\pi\sigma^2)^N}}\exp\left[\frac{1}{2\sigma^2}(x-\mathbf{1}\theta)(x-\mathbf{1}\theta)^T\right]$$

completando cuadrados

$$\begin{aligned}(x-\mathbf{1}\theta)^T(x-\mathbf{1}\theta) &= x^Tx-2\theta\mathbf{1}^T\theta+\mathbf{1}^T\mathbf{1}\theta^2 \\ &= x^Tx-2\theta\sum_{i=1}^N x_i+N\theta^2 \\ &= (\theta-T(x)/N)N(\theta-T(x)/N)+\sum_{i=1}^N x_i^2-(T(x)/N)^2\end{aligned}$$

siendo $T(x) = \sum_{i=1}^{N} x_i$. Entonces la función de densidad de probabilidad puede expresarse como

$$\begin{aligned} p(x;\theta) &= \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left[-\frac{1}{2\sigma^2/N}(\theta - T(x)/N)\right] \exp\left[\frac{1}{2\sigma^2}(T(x)/N)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right] \\ &= g(T(x),\theta)h(x) \end{aligned}$$

con

$$\begin{aligned} g(T(x),\theta) &= \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left[-\frac{1}{2\sigma^2/N}(\theta - T(x)/N)^2\right] \exp\left[\frac{1}{2\sigma^2}(T(x)/N)^2\right] \\ h(x) &= \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right] \end{aligned}$$

Para obtener un estimador insesgado y eficiente calculamos la esperanza del estadístico $T(x)$,

$$\begin{aligned} \mathbf{E}[T(x)] &= \sum_{i=1}^{N} \mathbf{E}(x_i) \\ &= \sum_{i=1}^{N} \mathbf{E}(\theta + w_i) \\ &= N\theta + \sum_{i=1}^{N} \mathbf{E}(w_i) \\ &= N\theta \end{aligned}$$

Entonces $\phi(T(x)) = \frac{1}{N}T(x)$ es un estadístico suficiente y completo y proporciona el estimador insesgado y eficiente del parámetro θ:

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

El estimador insesgado y eficiente podría haberse obtenido haciendo uso del teorema de Rao-Blackwell-Lehmann-Scheffe. En este caso, puede elegirse como estimador insesgado de θ $\breve{\theta} = x_1$ ya que

$$\begin{aligned} \mathbf{E}(x_1) &= \mathbf{E}(\theta + w_1) \\ &= \theta + \mathbf{E}(w_1) \\ &= \theta \end{aligned}$$

Entonces, como $T(x) = \sum_{i=1}^{N} x_i$ es un estadístico suficiente y completo, el estimador insesgado y eficiente se obtiene como sigue

$$\hat{\theta} = \mathbf{E}[\breve{\theta}|T]$$

Teniendo en cuenta que

$$z = \begin{bmatrix} \breve{\theta} \\ T \end{bmatrix} = Lx$$

siendo

$$L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \dots \\ x_N \end{bmatrix}$$

entonces, z es un vector de v.a. gaussianas con esperanzas y covarianzas dadas por las siguientes expresiones

$$\mathbf{E}(z) = \begin{bmatrix} 1 \\ N \end{bmatrix} \theta, \quad \mathbf{Cov}(z) = C = \sigma^2 L L^T = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix}$$

La función de densidad de probabilidad conjunta de $\breve{\theta}$ y T es

$$p(\breve{\theta}, T; \theta) = p(z) = \frac{1}{\sqrt{(2\pi\sigma^2)^2(N-1)}} \exp\left[-\frac{1}{2\sigma^2} \begin{bmatrix} \breve{\theta} - \theta \\ T - N\theta \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix} \begin{bmatrix} \breve{\theta} - \theta \\ T - N\theta \end{bmatrix} \right]$$

que puede descomponerse

$$\begin{aligned} p(\breve{\theta}, T; \theta) &= p(\breve{\theta}|T; \theta) p(T; \theta) \\ &= \frac{1}{\sqrt{2\pi\sigma^2(1-1/N)}} \exp\left[-\frac{(\breve{\theta} - T/N)^2}{2(1-1/N)\sigma^2} \right] \frac{1}{\sqrt{2\pi\sigma^2 N}} \exp\left[-\frac{(T - N\theta)^2}{2N\sigma^2} \right] \end{aligned}$$

Por tanto, la densidad de probabilidad condicionada es

$$p(\breve{\theta}|T; \theta) = \frac{1}{\sqrt{2\pi\sigma^2(1-1/N)}} \exp\left[-\frac{(\breve{\theta} - T/N)^2}{2(1-1/N)\sigma^2} \right]$$

y el estimador insesgado y eficiente es la esperanza condicionada

$$\hat{\theta} = \mathbf{E}[\breve{\theta}|T] = \frac{T}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Las propidedades estadísticas de este estimador son

$$\begin{aligned} \mathbf{E}(\hat{\theta} - \theta) &= \frac{1}{N}\mathbf{E}(T) - \theta = \frac{1}{N} N\theta - \theta = 0 \\ \mathbf{Cov}(\hat{\theta}) &= \mathbf{Cov}\left[\frac{T}{N} \right] = \frac{1}{N^2}\mathbf{Cov}(T) = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N} \end{aligned}$$

1.4. El estimador de máxima verosimilitud

En los casos en que no existe el estimador insesgado de mínima varianza o bien es difícil de calcular por los métodos explicados anteriormente se propone utilizar el estimador de máxima verosimilitud definido de la siguiente forma

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \mathbf{\Theta}} \ln p(x; \theta)$$

siendo $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}^T$ el vector de observaciones de una variable aleatoria con función de densidad de probabilidad $f(x; \theta_0)$ y θ_0 el verdadero valor del vector de parámetros, que no es conocido y desea determinarse.

La idea que subyace es que el estimador calcula el valor del vector de parámetros que maximiza la probabilidad de ocurrencia de los datos observados. El estimador de máxima verosimilitud satisface las siguientes propiedades:

1. Es consistente, es decir converge asintóticamente al verdadero valor de los parámetros.

$$\lim_{N \to \infty} \hat{\theta} = \theta_0$$

2. Es asintóticamente insesgado.
$$\lim_{N\to\infty} \mathbf{E}[\hat{\theta}] = \theta_0$$
3. Es asintóticamente eficiente.
$$\lim_{N\to\infty} \mathbf{Cov}[\hat{\theta}] = \mathbf{I}_N^{-1}(\theta_0)$$
siendo $\mathbf{I}_N(\theta)$ la matriz de Fisher obtenida con N observaciones.
4. Está asintóticamente distribuido según una distribución de probabilidad $\mathcal{N}(\theta_0, \mathbf{I}_N^{-1}(\theta_0))$.
5. Coincide con el estimador insesgado de mínima varianza, si este último existe.
6. Puede obtenerse mediante métodos numéricos (Newton-Raphson, algoritmo EM).

Consistencia del estimador de máxima verosimilitud La demostración de consistencia del estimador de máxima verosimilitud se basa en la ley de los grandes números que establece que si $\{x_k\}_{k=1}^N$ es un conjunto de observaciones de una variable aleatoria X, entonces
$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} h(x_k, \theta)$$
converge en probabilidad a $\mathbf{E}(h(X,\theta))$.

Debido a que las observaciones x_k son independientes entre sí, se cumple que
$$\begin{aligned} \ln p(x;\theta) &= \ln \prod_{k=1}^{N} p(x_k;\theta) \\ &= \sum_{k=1}^{N} \ln p(x_k;\theta) \end{aligned}$$
y tomando límites cuando el número de observaciones N tiende a infinito, por la ley de los grandes números se obtiene
$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \ln p(x_k;\theta) &= \mathbf{E}[\ln p(X;\theta)] \\ &= \int_{x_k} \ln p(x_k;\theta) p(x_k;\theta_0) dx_k \end{aligned}$$
siendo $p(x_k;\theta_0)$ la función de densidad de probabilidad que generó las observaciones. Teniendo en cuenta la siguiente propiedad[2]
$$\int_{x_k} \ln \frac{p(x_k;\theta_0)}{p(x_k;\theta)} p(x_k;\theta_0) dx_k \geq 0$$

[2]La divergencia de Kullback-Leiber entre dos funciones de densidad de probabilidad se define $D(p_1,p_2) := \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx$ y es una medida de la disimilaridad entre dos distribuciones de probabilidad. La divergencia de Kullback-Leiber satisface que $D(p_1,p_2) \geq 0$ para cualesquiera p_1 y p_2 y solo se anula cuando $p_1 = p_2$. Para comprobarlo se utiliza la desigualdad $\ln x \leq x - 1$, que solo es una igualdad para $x = 1$, entonces si $p_1(x)$ y $p_2(x)$ son dos funciones de densidad de probabilidad cualesquiera
$$\int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx = -\int p_1(x) \ln \frac{p_2(x)}{p_1(x)} dx \geq -\int p_1(x) \left(\frac{p_2(x)}{p_1(x)} - 1 \right) dx = -\int p_2(x) dx + \int p_1(x) dx = -1 + 1 = 0$$

para todo $\theta \neq \theta_0$, se cumple

$$\int_{x_k} \ln p(x_k;\theta_0)p(x_k;\theta_0)d\theta \geq \int_{x_k} \ln p(x_k;\theta)p(x_k;\theta_0)d\theta$$

La igualdad se obtiene cuando $\theta = \theta_0$, lo que implica que el máximo de la función logarítmica de verosimilitud cuando $N \to \infty$ es el verdadero valor del vector de parámetros. Por tanto el estimador de máxima verosimilitud es consistente ya que tiende asintóticamente al verdadero valor de los parámetros,

$$\lim_{N\to\infty} \hat{\theta} = \theta_0$$

Estimador asintóticamente insesgado. La consistencia del estimador implica que es asintóticamente insesgado, ya que tomando la esperanza

$$\lim_{N\to\infty} \mathbf{E}[\hat{\theta}] = \mathbf{E}[\theta_0] = \theta_0$$

Eficiencia asintótica. Para obtener la matrix de covarianza se hace uso de la propiedad de que el gradiente del logaritmo de la función de verosimilitud se anula en el máximo, es decir

$$\begin{aligned} 0 &= \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\hat{\theta}} \\ &= \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta_0} + \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\bar{\theta}} (\hat{\theta}-\theta_0) \end{aligned}$$

donde la segunda igualdad se obtiene aplicando el teorema del valor medio[3] para $\bar{\theta} = (1-\alpha)\hat{\theta} + \alpha\theta_0$. con $\alpha \in (0,1)$.

Elevando al cuadrado la expresión anterior, se obtiene

$$\left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\bar{\theta}} (\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T \left(\left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\bar{\theta}}\right)^T = \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta_0} \left(\left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta_0}\right)^T$$

Tomando ahora límite cuando N tiende a infinito y teniendo en cuenta que $\lim_{N\to\infty} \hat{\theta} = \theta_0$, entonces

$$\begin{aligned} \lim_{N\to\infty} \bar{\theta} &= (1-\alpha) \lim_{N\to\infty} \hat{\theta} + \alpha\theta_0 \\ &= (1-\alpha)\theta_0 + \alpha\theta_0 = \theta_0 \end{aligned}$$

Aplicando la ley de los grandes números

$$\begin{aligned} \lim_{N\to\infty} \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\bar{\theta}} &= \lim_{N\to\infty} \frac{N}{N} \sum_{k=1}^{N} \left.\frac{\partial^2 \ln p(x_k;\theta)}{\partial \theta^2}\right|_{\bar{\theta}} \\ &= \lim_{N\to\infty} N\mathbf{E}\left[\left.\frac{\partial^2 \ln p(x_k;\theta)}{\partial \theta^2}\right|_{\theta_0}\right] \\ &= \lim_{N\to\infty} N\mathbf{I}_1(\theta_0) = \lim_{N\to\infty} \mathbf{I}_N(\theta_0) \end{aligned}$$

[3]El teorema del valor medio de Lagrange establece que $f(x,\hat{\theta}) = f(x,\theta_0) + \left.\frac{\partial f(x,\theta)}{\partial x}\right|_{\bar{\theta}} (\hat{\theta}-\theta_0)$ para $\bar{\theta} = (1-\alpha)\hat{\theta} + \alpha\theta_0$ y algún $\alpha \in (0,1)$.

siendo $\mathbf{I}_1(\theta_0)$ la matriz de Fisher asociada a la función de verosimilitud de una observación e $\mathbf{I}_N(\theta_0)$ la matriz de Fisher asociada a la función de verosimilitud de N observaciones que están relacionadas por la expresión $\mathbf{I}_N(\theta_0) = N\mathbf{I}_1(\theta_0)$, ya que $\ln p(x;\theta_0) = \sum_{k=1}^{N} \ln p(x_k,\theta_0)$.

Por tanto, la expresión del error del estimador al cuadrado, cuando N tiende a infinito, queda

$$\lim_{N\to\infty} \mathbf{I}_N(\theta_0)(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T\mathbf{I}_N(\theta_0) = \lim_{N\to\infty} \left.\frac{\partial \ln p(x;\theta)}{\partial\theta}\right|_{\theta_0} \left(\left.\frac{\partial \ln p(x;\theta)}{\partial\theta}\right|_{\theta_0}\right)^T$$

y tomando experanzas a ambos lados de la igualdad

$$\lim_{N\to\infty} \mathbf{I}_N(\theta_0)\mathbf{E}[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T]\mathbf{I}_N(\theta_0) = \lim_{N\to\infty} \mathbf{I}_N(\theta_0)$$

de donde se obtiene finalmente

$$\lim_{N\to\infty} \mathbf{E}\left[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T\right] = \lim_{N\to\infty} \mathbf{I}_N^{-1}(\theta_0)$$

por tanto el estimador de máxima verosimilitud es asintóticamente eficiente ya que su covarianza iguala la cota de Cramer-Rao cuando número de observaciones tiende a infinito.

Distribución asintótica gaussiana. Para demostrar que el estimador de máxima verosimilitud está distribuido según una gaussiana tomamos límites cuando N tiende a infinito en la expresión del gradiente de la función de verosimilitud y obtenemos

$$\lim_{N\to\infty} \mathbf{I}_N(\theta_0)(\hat{\theta}-\theta_0) = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \left.\frac{\partial \ln p(x_k;\theta)}{\partial\theta}\right|_{\theta_0}$$

El término de la derecha es una suma de variables aleatorias independientes y según el teorema del límite central converge a una variable aleatoria con distribución gaussiana.

Por tanto, como ya hemos demostrado que cuando el número de observaciones tiende a infinito

$$\begin{aligned} \lim_{N\to\infty} \mathbf{E}(\hat{\theta}-\theta_0) &= 0 \\ \lim_{N\to\infty} \mathbf{Cov}(\hat{\theta}) &= \mathbf{I}^{-1}(\theta_0) \end{aligned}$$

el estimador de máxima verosimilitud está distribuido asintóticamente según una gaussiana de media θ_0 y matriz de covarianza $\mathbf{I}^{-1}(\theta)$

$$\hat{\theta}-\theta_0 \sim \mathcal{N}(0,\mathbf{I}^{-1}(\theta_0))$$

Ejemplo Consíderese de nuevo el ejemplo

$$x_k = a + w_k$$

con w_k un ruido blanco gaussiano de media cero y varianza σ^2 entonces:

$$\begin{aligned} p(x;\theta) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\theta)^2\right] \\ \ln p(x;\theta) &= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\theta)^2 \end{aligned}$$

entonces, maximizando la función logarítmica de verosimilitud se obtiene

$$0 = \frac{\partial \ln p(x;\theta)}{\partial \theta} = \frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{k=1}^{N} x_k - \theta\right)$$

de donde el estimador de máxima verosimilitud es

$$\hat{\theta} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

El estimador de máxima verosimilitud coincide para este caso con el estimador insesgado eficiente por lo que su sesgo y varianza son

$$\begin{aligned} \mathbf{E}[\hat{\theta} - a] &= 0 \\ \mathbf{Cov}[\hat{\theta}] &= \frac{\sigma^2}{N} \end{aligned}$$

Ejemplo (Modelo lineal en los parámetros) Sea el modelo lineal

$$b(x) = A(x)\theta + w$$

con $x \in \mathbb{R}^N$ el vector de observaciones, $\theta \in \mathbb{R}^n$ el vector de parámetros desconocidos y w un vector de dimensión m de variables aleatorias gaussianas de media cero y matriz de covarianza $\mathbf{Cov}(w) = C$. La función de verosimilitud es:

$$L(\theta) = \ln p(x;\theta) = \frac{1}{2}\ln((2\pi)^N \det(C)) - \frac{1}{2}(b(x) - A(x)\theta)^T C^{-1}(b(x) - A(x)\theta)$$

El máximo de la función logarítmica de verosimilitud se obtiene cuando su gradiente se anula, es decir cuando

$$0 = \frac{\partial}{\partial \theta} L(\theta) = A^T(x)C^{-1}(b(x) - A(x)\theta) = 0$$

Por tanto el estimador de máxima verosimilitud es:

$$\hat{\theta} = (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}b(x)$$

Notar que el estimador de máxima verosimilitud coincide con el estimador insesgado de mínima varianza, y por tanto su covarianza es igual a la inversa de la matriz de información de Fisher $\mathbf{I}(\theta) = -\mathbf{E}(\frac{\partial^2}{\partial \theta^2}\ln p(x;\theta))$.

$$\begin{aligned} \mathbf{E}(\hat{\theta}) &= \theta \\ \mathbf{Cov}(\hat{\theta}) &= \theta = (A^T(x)C^{-1}A(x))^{-1} \end{aligned}$$

Es interesante destacar que maximizar la función de verosimilitud coincide con minimizar la función

$$V(\theta) = \frac{1}{2}(b(x) - A(x)\theta)^T C^{-1}(b(x) - A(x)\theta)$$

que es una función cuadrática y que puede expresarse de forma equivalente como

$$\begin{aligned} V(\theta) = {} & \frac{1}{2}(A^T(x)C^{-1}b(x) - A^T(x)C^{-1}A(x)\theta)^T(A^T(x)C^{-1}A(x))^{-1}(A^T(x)C^{-1}b(x) - A^T(x)C^{-1}A(x)\theta) + \\ & \frac{1}{2}b^T(x)(C^{-1} - C^{-1}A(x)(A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1})b(x) \end{aligned}$$

La expresión anterior es una suma de cuadrados, y solo el primer término depende de θ, el máximo se obtiene para

$$\hat{\theta} = (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}b(x)$$

En este caso, de nuevo el estimador de máxima verosimilitud coincide con el estimador insesgado de mínima varianza. Este resultado permitirá relacionar el estimador de máxima verosimilitud con el estimador de mínimos cuadrados.

Sustituyendo en la expresión del estimador de máxima verosimilitud

$$b(x) = A(x)\theta + w$$

se obtiene

$$\begin{aligned}\hat{\theta} &= (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}b(x)\\ &= (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}(A(x)\theta + w)\\ &= \theta + (A^T(x)C^{-1}A(x))^{-1}A^T(x)C^{-1}w\end{aligned}$$

de donde se obtiene facilmente que el estimador es insesgado y eficiente

$$\begin{aligned}\mathbf{E}(\hat{\theta} - \theta) &= 0\\ \mathbf{Cov}(\hat{\theta}) &= (A^T(x)C^{-1}A(x))^{-1} = \mathbf{I}^{-1}(\theta)\end{aligned}$$

Ejemplo (Variable aleatoria gaussiana con la misma esperanza y varianza) Sea $\{x_k\}_{k=1}^N$ una secuencia de observaciones de una variable aleatoria de media y varianza a. Obtener el estimador de máxima verosimilitud de a.

El parámetro a estimar es $\theta = a$. Sea $x = \begin{bmatrix} x_1 & x_2 & \dots & x_N \end{bmatrix}$ el vector que agrupa las observaciones. La función logarítmica de verosimilitud para este problema es:

$$\ln p(x;\theta) = -\frac{N}{2}\ln(2\pi\theta) - \frac{1}{2\theta}(x - \mathbf{1}\theta)^T(x - \mathbf{1}\theta)$$

siendo $\mathbf{1}$ un vector de unos de dimensión N. La derivada de la función logarítmica de verosimilitud es:

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln p(x;\theta) &= -\frac{N}{2\theta} + \frac{1}{2\theta^2}(x - \mathbf{1}\theta)^T(x - \mathbf{1}\theta) + \frac{1}{\theta}\mathbf{1}^T(x - \mathbf{1}\theta)\\ &= \frac{N}{2\theta^2}\left(-\theta + \frac{1}{N}(x - \mathbf{1}\theta)^T(x - \mathbf{1}\theta) + \frac{2}{N}\mathbf{1}^T\theta(x - \mathbf{1}\theta)\right)\\ &= \frac{N}{2\theta^2}\left(-\theta + \frac{1}{N}(x + \mathbf{1}\theta)^T(x - \mathbf{1}\theta)\right)\end{aligned}$$

El estimador se obtiene resolviendo la ecuación:

$$\begin{aligned}0 &= -\theta + \frac{1}{N}(x + \mathbf{1}\theta)^T(x - \mathbf{1}\theta)\\ &= -\theta + \frac{1}{N}x^Tx - \frac{1}{N}\mathbf{1}\mathbf{1}^T\theta^2\\ &= -\theta + \frac{1}{N}x^Tx - \theta^2\end{aligned}$$

Es una ecuación de segundo orden cuya única solución positiva (ya que la varianza debe ser siempre positiva) es:

$$\hat{\theta} = -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{N}x^Tx}$$

En primer lugar comprobaremos que el estimador es asintóticamente insesgado. Cuando N tiende a infinito

$$\frac{1}{N}x^T x = \frac{1}{N}\sum_{k=1}^{N} x_k^2 \to \mathbf{E}[x^2] = a + a^2$$

entonces

$$\begin{aligned}\lim_{N\to\infty} \hat{\theta} &= -\frac{1}{2} + \sqrt{\frac{1}{4} + a + a^2} \\ &= -\frac{1}{2} + \sqrt{\left(a + \frac{1}{2}\right)^2} = a\end{aligned}$$

Además se puede demostrar que este estimador es asintóticamente eficiente y su covarianza cuando N tiende a infinito es

$$\mathbf{Cov}(\hat{x}) = \mathbf{I}_N(a) = \frac{2a^2}{N}$$

1.4.1. Método numérico de optimización

Una ventaja del estimador de máxima verosimilitud es que pueden implementarse métodos numéricos de tipo Newton Raphson para su obtención.

El estimador de máxima verosimilitud satisface las siguientes condiciones

$$\begin{aligned}0 &= g(\hat{x}, \theta) = \frac{\partial}{\partial \theta} \ln p(x;\theta)\bigg|_{\hat{\theta}} \\ 0 &\leq H(x, \hat{\theta}) = \frac{\partial}{\partial \theta} g(x, \theta)\bigg|_{\hat{\theta}}\end{aligned}$$

Si θ_k es un vector próximo al estimador buscado, entonces la función $g(x, \theta)$ puede aproximarse linealmente en un entorno del vector θ_k

$$g(x, \theta) \approx g(x, \theta_k) + H(x, \theta_k)(\theta - \theta_k) = 0$$

El siguiente vector de parámetros θ_{k+1} es aquel que anula la aproximación lineal de $g(x, \theta)$,

$$\theta_{k+1} = \theta_k - H^{-1}(x, \theta_k) g(x, \theta_k)$$

El proceso se repite hasta que θ_k satisface $g(x, \theta_k) = 0$.

1.4.2. Estimador de máxima verosimilitud de un modelo no lineal

Sea el modelo no lineal en los parámetros:

$$b(x) = f(x, \theta) + w$$

siendo $x \in \mathbb{R}^N$ el vector de obsersevaciones, $\theta \in \mathbb{R}^n$ el vector de parámetros desconocidos y $w \in \mathbb{R}^m$ un vector de ruido blanco gaussiano de media cero y matriz de varianza C. Los vectores de funciones $b(x)$ y $f(x, \theta)$ son conocidos y tienen dimensión m. Sea θ_0 el verdadero valor del vector de parámetros que se desea estimar y que no es conocido.

La función logarítmica de verosimilitud asociada al modelo es:

$$\ln p(x;\theta) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(b(x) - f(x,\theta)\right)^T C^{-1}\left(b(x) - f(x,\theta)\right)$$

cuyo gradiente respecto al vector de parámetros θ es:

$$\frac{\partial}{\partial \theta} \ln p(x;\theta) = -\left(\frac{\partial f(x,\theta)}{\partial \theta}\right)^T C^{-1}\left(b(x) - f(x,\theta)\right)$$

El estimador de máxima verosimilitud es el vector de parámetros $\hat{\theta}$ que resuelve la ecuación:

$$-\left(\frac{\partial f(x,\theta)}{\partial \theta}\right)^T C^{-1}\left(b(x) - f(x,\theta)\right) = 0$$

y satisface que el hessiano

$$\frac{\partial^2}{\partial \theta^2} \ln p(x;\theta)$$

es una matriz semidefinida negativa.

La ecuación anterior puede ser muy difícil de resolver analíticamente y se propone la utilización de un método numérico. La idea básica es la siguiente, a partir de un vector de parámetros θ_k próximo a la solución buscada, se intentará obtener un nuevo vector θ_{k+1} más cercano al estimador buscado. Para ello se aproxima linealmente la función $f(x;\theta)$ en un entorno del vector θ_k

$$f(x,\theta) \approx f(x,\theta_k) + A(x,\theta_k)(\theta - \theta_k)$$

siendo

$$A(x,\theta_k) = \left.\frac{\partial f(x,\theta)}{\partial \theta}\right|_{\theta_k}$$

La ecuación que caracteriza el estimador de máxima verosimilitud en función de la aproximación de la función $f(x,\theta)$ es ahora

$$0 = -A^T(x,\theta_k)C^{-1}\left(b(x) - f(x,\theta_k) - A(x,\theta_k)(\theta - \theta_k)\right)$$

cuya solución proporciona el siguiente vector de parámetros θ_{k+1}.

$$\theta_{k+1} = \theta_k + (A^T(x,\theta_k)C^{-1}A(x,\theta_k))^{-1}A^T(x,\theta_k)C^{-1}(b(x) - f(x,\theta_k))$$

El proceso se repite hasta que el gradiente de la función logarítmica de verosimilitud se anule

$$\frac{\partial}{\partial \theta} \ln p(x;\theta_k) = 0$$

Las propiedades estadísticas del estimador se obtienen a partir de la aproximación lineal de la función no lineal $f(x,\theta)$ en el entorno del verdadero valor del vector de parámetros. El estimador de máxima verosimilitud satisface la ecuación

$$\begin{aligned}
0 &= -A^T(x,\hat{\theta})C^{-1}\left(b(x) - f(x,\hat{\theta})\right) \\
&= -A^T(x,\hat{\theta})C^{-1}\left(b(x) - f(x,\theta_0) - A(x,\bar{\theta})(\theta_0 - \hat{\theta})\right) \\
&= -A^T(x,\hat{\theta})C^{-1}\left(w + f(x,\theta_0) - f(x,\theta_0) - A(x,\bar{\theta})(\theta_0 - \hat{\theta})\right) \\
&= -A^T(x,\hat{\theta})C^{-1}\left(w - A(x,\bar{\theta})(\theta_0 - \hat{\theta})\right)
\end{aligned}$$

para $\bar{\theta} = (1-\alpha)\hat{\theta} + \alpha\theta$. Para la segunda igualdad de la expresión anterior se ha hecho uso del teorema del valor medio.

Despejando se obtiene

$$\hat{\theta} - \theta_0 = (A(x,\hat{\theta})^T C^{-1} A(x,\bar{\theta}))^{-1} C^{-1} A(x,\bar{\theta}) w$$

Teniendo en cuenta que el estimador de máxima verosimilitud es asintóticamente insesgado, entonces cuando el número de observaciones tiende a infinito $\hat{\theta} = \bar{\theta} = \theta_0$ y las propiedades estadísticas son:

$$\begin{aligned} \mathbf{E}(\theta_0 - \hat{\theta}) &= 0 \\ \mathbf{Cov}(\theta) &= (A(x,\theta_0)^T C^{-1} A(x,\theta_0))^{-1} \geq \mathbf{I}^{-1}(\theta_0) \end{aligned}$$

La matriz de covarianza coincide con la inversa de la matriz de Fisher, ya que

$$\begin{aligned} \mathbf{I}(\theta) &= \mathbf{E}\left[\left(\frac{\partial p(x;\theta)}{\partial \theta} \right) \left(\frac{\partial p(x;\theta)}{\partial \theta} \right)^T \right] \\ &= \mathbf{E}\left[-A^T(x,\theta) C^{-1} (b(x) - f(x,\theta)) (-A^T(x,\theta) C^{-1} (b(x) - f(x,\theta)))^T \right] \\ &= A^T(x,\theta) C^{-1} \mathbf{E}[ww^T] C^{-1} A(x,\theta) \\ &= A^T(x,\theta) C^{-1} A(x,\theta) \end{aligned}$$

y por tanto el estimador es asintóticamente eficiente.

La matriz de covarianza depende del verdadero valor del vector de parámetros θ que no es conocido, sin embargo en virtud de las propiedades del estimador de máxima verosimilitud, puede aproximarse por el valor estimado obteniéndose

$$\mathbf{Cov}(\hat{\theta}) \approx (A(x,\hat{\theta})^T C^{-1} A(x,\hat{\theta}))^{-1}$$

la aproximación es suficientemente buena cuando el número de observaciones es grande.

Finalmente, como cuando el número de observaciones tiende a infinito, el estimador puede ponerse como:

$$\hat{\theta} - \theta_0 = (A(x,\theta_0)^T C^{-1} A(x,\theta_0))^{-1} C^{-1} A(x,\theta_0) w$$

El estimador es una combinación lineal de procesos aleatorios gaussianos y por tanto tiene distribución de probabilidad gaussiana con esperanza θ_0 y covarianza $(A(x,\theta_0)^T C^{-1} A(x,\theta_0))^{-1}$.

1.5. El mejor estimador lineal insesgado

Se considera que el estimador es una función lineal del vector de observación:

$$\hat{\theta} = Kx$$

Para que sea insesgado

$$\mathbf{E}[\hat{\theta}] = K\mathbf{E}[x] = \theta$$

lo que solo puede cumplirse si

$$\mathbf{E}[x] = A\theta$$

en cuyo caso K debe ser tal que $KA = I$.

La covarianza es $\mathbf{E}(\hat{\theta}) = K^T CK$ con $C = \mathbf{Cov}(x) = \mathbf{E}[(x - \mathbf{E}(x))(x - \mathbf{E}(x))^T]$. El estimador se obtiene resolviendo el problema de optimización

$$\hat{\theta} = \arg \min_K \{ \mathrm{Tr} KCK^T \mid KA = I \}$$

cuya solución es

$$K = (A^T C^{-1} A)^{-1} A^T C^{-1}$$

por tanto el estimador es:

$$\hat{\theta} = (A^T C^{-1} A)^{-1} A^T C^{-1} x$$

Este estimador se denomina mejor estimador lineal insesgado. Nótese que no hace uso de la función de densidad de probabilidad de las observaciones, sino solamente de los momentos de primer y segundo orden (media y matriz de covarianza). Sin embargo, produce el mismo resultado que el estimador insesgado de mínima varianza si el ruido fuera gaussiano.

1.6. El estimador de mínimos cuadrados

Se basa en suponer conocido el modelo

$$b(x) = A(x)\theta + e$$

siendo e un vector de errores de modelado sobre el que no se dispone de información probabilística. Se pretende minimizar la suma ponderada de errores al cuadrado:

$$\hat{\theta} = \arg\min\{(b(x) - A(x)\theta)^T W (b(x) - A(x)\theta)\}$$

siendo $W = W^T > 0$ la matriz de peso de los errores.

La solución se obtiene completando cuadrados

$$\begin{aligned}(b(x) - A(x)\theta)^T W (b(x) - A(x)\theta) \;=\; & ((A^T(x)WA(x))^{-1}A^T(x)W\,b(x) - \theta)^T (A^T(x)WA(x)) \times \\ & ((A^T(x)WA(x))^{-1}A^T(x)W\,b(x) - \theta) + \\ & b^T(x)(W - WA(x)(A^T(x)WA(x))^{-1}A^T(x)W)b(x)\end{aligned}$$

El mínimo se obtiene para

$$\hat{\theta} = (A^T(x)WA(x))^{-1}A^T(x)W\,b(x)$$

y el valor mínimo de la suma de errores al cuadrado es $b^T(x)(W - WA(x)(A^T(x)A(x))^{-1}A^T(x)W)b(x)$.

El estimador de mínimos cuadrados es un estimador determinista que no hace uso de la información de los errores. Sin embargo, se obtiene la misma expresión que la obtenida para el estimador insesgado de mínima varianza si los errores fueran un proceso aleatorio con matriz de covarianza $\mathbf{Cov}(e) = W^{-1}$.

Sin embargo conviene notar que si no se dispone de la información estadística de los errores, no es posible elegir adecuadamente la función de peso $W = (\mathbf{Cov}(e))^{-1}$ y en general el estimador de mínimos cuadrados tendrá sesgo. Para el caso particular de que los errores sean ruido blanco, la matriz de peso puede elegirse igual a la identidad $W = I$ y el estimador coincidirá con el estimador insesgado de mínima varianza.

1.6.1. El estimador recursivo de mínimos cuadrados

El estimador de mínimos cuadrados puede expresarse de forma recursiva. Supongamos que la matriz $A(x)$ y el vector $b(x)$ se particionan conforme a las observaciones y que la matriz de pesos de los errores es diagonal por bloques cuyos elementos se denotan por W_i siendo cada uno

de ellos una matriz simétrica y semidefinida positiva de dimensión igual al vector de observación x_i.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} A(x_1) \\ A(x_2) \\ \vdots \\ A(x_k) \end{bmatrix} \theta_k + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}$$

Entonces el estimador de mínimos cuadrados calculado con k observaciones se denota como θ_k y se obtiene resolviendo de forma recursiva el siguiente sistema de ecuaciones lineales, cada vez que se obtiene una nueva observación.

$$\begin{aligned} M_k\hat{\theta}_k &= N_k \\ M_k &= M_{k-1} + A_k^T W_k A_k \\ N_k &= N_{k-1} + A_k^T W_k x_k \end{aligned}$$

siendo $A_k = A(x_k)$ y con $M_0 = 0$, $N_0 = 0$.

Suele ser interesante expresar la fórmula recursiva en términos del propio vector de parámetros, para ello se hace lo siguiente

$$\begin{aligned} M_k\hat{\theta}_{k-1} &= M_{k-1}\hat{\theta}_{k-1} + A_k^T W_k A_k \hat{\theta}_{k-1} \\ &= N_{k-1} + A_k^T W_k A_k \hat{\theta}_{k-1} \end{aligned}$$

despejando N_{k-1}

$$N_{k-1} = M_k\hat{\theta}_{k-1} - A_k^T W_k A_k \hat{\theta}_{k-1}$$

y sustituyendo en la recursión de N_k

$$N_k = M_k\hat{\theta}_{k-1} - A_k^T W_k A_k \hat{\theta}_{k-1} + A_k^T W_k x_k$$

La expresión del estimador es

$$\begin{aligned} M_k\hat{\theta}_k &= N_k \\ &= M_k\hat{\theta}_{k-1} - A_k^T W_k A_k \hat{\theta}_{k-1} + A_k^T W_k x_k \\ &= M_k\hat{\theta}_{k-1} - A_k^T W_k (x_k - A_k \hat{\theta}_{k-1}) \end{aligned}$$

quedando finalmente

$$\begin{aligned} M_k\hat{\theta}_k &= M_k\hat{\theta}_{k-1} + A_k^T W_k (x_k - A_k \hat{\theta}_{k-1}) \\ M_k &= M_{k-1} + A_k^T W_k A_k \\ M_0 &= 0 \end{aligned}$$

El procedimiento recursivo puede hacerse todavía más eficiente mediante la identidad de Woodbury[4]

$$\begin{aligned} P_k &= M_k^{-1} \\ &= (M_{k-1} + A_k^T W_k A_k)^{-1} \\ &= M_{k-1}^{-1} - (M_{k-1}^{-1} A_k^T)(W_k^{-1} + A_k M_{k-1}^{-1} A_k^T)^{-1}(A_k M_{k-1}^{-1}) \\ &= P_{k-1} - (P_{k-1} A_k^T)(W_k^{-1} + A_k P_{k-1} A_k^T)^{-1}(A_k P_{k-1}) \end{aligned}$$

[4]La identidad de Woodbury establece que para cuatro matrices A, B, C y D de dimensiones compatibles y siendo A y D regulares

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}$$

La fórmula recursiva de mínimos cuadrados es

$$\begin{aligned}\hat{\theta}_k &= \hat{\theta}_{k-1} + P_k A_k^T W_k (x_k - A_k \hat{\theta}_{k-1}) \\ P_k &= P_{k-1} - (P_{k-1} A_k^T)(W_k^{-1} + A_k P_{k-1} A_k^T)^{-1}(A_k P_{k-1}) \\ P_0 &= \lambda I, \quad \lambda \gg 1\end{aligned}$$

1.7. El estimador basado en momentos

El método de los momentos calcula los momentos estadísticos de de las observaciones de un determinado modelo matemático caracterizado por un vector de parámetros θ para obtener una determinada relación (no necesariamente lineal) entre los momentos y el vector de parámetros buscado. A partir de esta relación se obtiene la expresión del vector de parámetros como función de los momentos.

El estimador de parámetros se obtiene sustituyendo los momentos teóricos por sus estimaciones a partir de las observaciones.

El estimador de los momentos se realiza en dos pasos. Sea $x(\theta)$ la observación que depende del vector de parámetros θ.

1. Se calculan los momentos $m_k = \mathbf{E}[(x(\theta) - \mathbf{E}(x^k) = h_k(\theta)$ de la variable aleatoria de observación y se forma el sistema de ecuaciones

$$m_k = h_k(\theta), \quad k = 1, \ldots, \ell$$

con $\ell \geq \dim(\theta)$ que puede expresarse en forma compacta como $m = H(\theta)$.

2. Se calculan estadísticos consistentes $T_k(x)$ que estimen los momentos a partir de las observaciones, por ejemplo pueden utilizarse los estimadores naturales

$$\hat{m}_k = T_k(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \quad k = 1, \ldots, \ell$$

y se resuelve el sistema de ecuaciones

$$\hat{m} = H(\theta)$$

para obtener el estimador $\hat{\theta} = H^{-1}(\hat{m})$. La función H no siempre puede invertirse de forma sencilla, en ese caso se recurrirá a métodos numéricos del tipo Newton-Raphson.

El método de los momentos no es óptimo, pero suele ser consistente y proporciona buenos estimadores cuando se dispone de un gran número de datos.

Para calcular las propiedades estadísticas de estos estimadores, supongamos que el estimador es

$$\hat{\theta} = h(T(x))$$

siendo $T(x) = [T_1(x) \cdots T_\ell(x)]^T$ el vector de estimadores de los momentos. Suponiendo que las covarianzas de $T(x)$ son pequeñas, y aproximando linealmente el estimador en el entorno de la media del vector T, $\mu = E(T)$

$$\hat{\theta} = h(T) \approx h(\mu) + \left.\frac{\partial h}{\partial T}\right|_{\mu} (T - \mu)$$

Una aproximación a la esperanza y a la matriz de covarianza del del estimador del vector de parámetros son:

$$\begin{aligned} \mathbf{E}(\hat{\theta}(T(x))) &\approx h(\mu) = \theta \\ \mathbf{Cov}(\hat{\theta}(T(x))) &\approx \left(\frac{\partial h}{\partial T}\bigg|_{\mu} \right) \mathbf{Cov}(T(x)) \left(\frac{\partial h}{\partial T}\bigg|_{\mu} \right)^T \end{aligned}$$

Ejemplo Sea el modelo lineal

$$x_k = a + w_k, \quad k = 1, \ldots, N$$

El momento de primer orden es

$$m_1 = \mathbf{E}(a + w) = a$$

El estimador del momento es la media muestral

$$\hat{m}_1 = \frac{1}{N} \sum_{k=1}^{N} x_k$$

Sustituyendo

$$\hat{a} = \hat{m}_1 = \frac{1}{N} \sum_{k=1}^{N} x_k$$

1.8. Estimadores bayesianos

La estimación bayesiana difiere de la estimación clásica en que el vector de parámetros desconocidos que se desea estimar no se considera un vector determinista sino un vector de variables aleatorias para el cual se desea obtener su realización o bien sus propiedades estadísticas. Se denomina estimación bayesiana porque para su desarrollo se hace uso del teorema de Bayes.

Entre sus características cabe destacar que requiere información a priori de la distribución de probabilidad del vector que se desea estimar. Esto puede ser una dificultad y ha sido muy criticado en la literatura, ya que la información previa condiciona el valor final de la estimación, especialmente cuando se dispone de un número escaso de observaciones. Sin embargo, cuando se dispone de esta información, el método bayesiano permite integrarla en el proceso de forma natural. Si no se dispone de esta información, podría ser más adecuado utilizar métodos clásicos de estimación.

Una ventaja del estimador bayesiano frente a los estimadores clásicos es que permite obtener estimadores en situaciones donde no existe el estimador insesgado eficiente. La clave radica en que el estimador bayesiano obtiene un estimador que es óptimo en promedio, es decir que el estimador bayesiano obtiene un mejor error cuadrático medio que cualquier otro estimador para la mayor parte de los valores del vector de parámetros.

Obtención del estimador bayesiano de mínimo error cuadrático medio Los elementos necesarios para establecer un método bayesiano de estimación son:

1. La distribución de probabilidad del vector de parámetros a estimar, $p(\theta)$.
2. El criterio de bondad que se minimizará, normalmente el error cuadrático medio de estimación definido como:
$$\mathbf{BMSE}(\hat{\theta}) = \mathbf{E}[(\hat{\theta} - \theta)^T(\hat{\theta} - \theta)]$$

3. Un vector de N observaciones x.

El error cuadrático medio se puede expresar de la siguiente forma

$$\begin{aligned}\mathbf{BMSE}(\hat{\theta}) &= \mathbf{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \\ &= \int_{\theta\in\Omega_\theta}\int_{x\in\Omega_x}(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)p(x,\theta)dxd\theta\end{aligned}$$

Notar que $p(x,\theta)$ es la función de densidad conjunta del vector de observaciones y del vector de parámetros. Aplicando el teorema de bayes

$$p(x,\theta)=p(\theta|x)p(x)$$

Aquí, $p(\theta|x)$ es la probabilidad del vector de parámetros θ, condicionada a la realización del vector de observación x.

El error cuadrático medio es

$$\begin{aligned}\mathbf{BMSE}(\hat{\theta}) &= \int_x\left(\int_\theta(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)p(\theta|x)d\theta\right)p(x)dx \\ &= \int_x\mathbf{E}\left[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)|x\right]p(x)dx\end{aligned}$$

La función de densidad $p(x)\geq 0$ para todo $x\in\Omega_x$, entonces el error cuadrático medio alcanza un mínimo cuando el error cuadrático medio condicionado a la observación es mínimo, por tanto

$$\hat{\theta}=\arg\min_{\hat{\theta}\in\mathbb{R}^n}\mathbf{E}\left[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)|x\right]$$

Para calcular el mínimo desarrollamos la expresión del error cuadrático medio condicionado a la observación

$$\begin{aligned}\mathbf{E}\left[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)|x\right] &= \int_\theta(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)p(\theta|x)d\theta \\ &= \hat{\theta}^T\hat{\theta}\int_\theta p(\theta|x)d\theta-2\hat{\theta}^T\int_\theta\theta p(\theta|x)d\theta+\int_\theta\theta^T\theta p(\theta|x)d\theta \\ &= \hat{\theta}^T\hat{\theta}-2\hat{\theta}^T\mathbf{E}(\theta|\theta)+\mathbf{E}(\theta^T\theta|\theta)\end{aligned}$$

Completando cuadrados

$$\begin{aligned}\mathbf{E}\left[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)|x\right] &= \hat{\theta}^T\hat{\theta}-2\hat{\theta}^T\mathbf{E}(\theta|\theta)+\mathbf{E}(\theta^T\theta|\theta) \\ &= (\hat{\theta}-\mathbf{E}(\theta|x))^T(\hat{\theta}-\mathbf{E}(\theta|x))+\mathbf{E}(\theta^T\theta|x)-(\mathbf{E}(\theta|x))^T(\mathbf{E}(\theta|x))\end{aligned}$$

y teniendo en cuenta que

$$\mathbf{E}\left[(\theta-\mathbf{E}(\theta|x))^T(\theta-\mathbf{E}(\theta|x))|x\right]=\mathbf{E}(\theta^T\theta|x)-(\mathbf{E}(\theta|x))^T(\mathbf{E}(\theta|x))$$

se obtiene finalmente

$$\mathbf{E}\left[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)|x\right]=(\hat{\theta}-\mathbf{E}(\theta|x))^T(\hat{\theta}-\mathbf{E}(\theta|x))+\mathbf{E}\left[(\theta-\mathbf{E}(\theta|x))^T(\theta-\mathbf{E}(\theta|x))|x\right]$$

De la expresión anterior, teniendo en cuenta que esta formada por dos sumandos ambos no negativos y que solo el primero depende de $\hat{\theta}$ el mínimo se alcanza en

$$\hat{\theta} = \mathbf{E}(\theta|x)$$

y el valor mínimo del error cuadrático medio es

$$\begin{aligned}
\mathbf{BMSE}(\hat{\theta}) &= \int_x \mathbf{E}\left[(\theta - \mathbf{E}(\theta|x))^T(\theta - \mathbf{E}(\theta|x))|x\right] p(x)dx \\
&= \int_x \mathrm{Tr}\mathbf{E}\left[(\theta - \mathbf{E}(\theta|x))(\theta - \mathbf{E}(\theta|x))^T|x\right] p(x)dx \\
&= \int_x \mathrm{Tr}\left[\mathbf{Cov}(\theta|x)\right] p(x)dx \\
&= \mathrm{Tr}\mathbf{E}\left[\mathbf{Cov}(\theta|x)\right] = \mathrm{Tr}\mathbf{Cov}(\theta)
\end{aligned}$$

El estimador bayesiano de error cuadrático medio mínimo (MMSE) es la esperanza del vector buscado condicionada a la observación y el error cuadrático medio mínimo que alcanza es la esperanza de la suma de las varianzas condicionadas de las componentes del vector de parámetros.

El estimador bayesiano es insesgado y eficiente pero de una forma diferente a los estimadores clásicos. En estimación bayesiana el vector que se desea estimar es un vector de variables aleatorias y su optimalidad es un promedio con respecto a su espacio muestral. Para ciertos valores del vector a estimar su varianza puede ser superior a la que proporcionaría un estimador clásico, sin embargo en promedio la varianza del estimador bayesiana es menor cuando se tiene en cuenta su espacio muestral.

Características del método bayesiano

1. La técnica bayesiana parte de la premisa que el vector que se desea estimar corresponde a la realización de un vector de variables aleatorias θ.

2. El estimador es la esperanza del vector de variables aleatorias θ condicionada al vector de observaciones. Esta esperanza condicional se calcula utilizando la función de densidad de probabilidad a posteriori.
$$\hat{\theta} = \mathbf{E}(\theta|x) = \int \theta p(\theta|x)d\theta$$

3. La función de densidad de probabilidad a posteriori se calcula haciendo uso del teorema de Bayes:
$$\begin{aligned}
p(\theta|x) &= \frac{p(x,\theta)}{p(x)} \\
&= \frac{p(x|\theta)p(\theta)}{p(x)} \\
&= \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}
\end{aligned}$$

4. El estimador de error cuadrático medio mínimo $\hat{\theta}$ depende tanto del conocimiento a priori codificado en la función de densidad de probabilidad a priori $p(\theta)$ como del vector de observación. La influencia de la información a priori disminuye al aumentar el número de observaciones.

5. La elección de la función de densidad de probabilidad a priori es crítica en la estimación bayesiana. Una elección equivocada puede influir produciendo un pobre estimador. La función de densidad de probabilidad a priori juega un papel similar a la elección de un modelo incorrecto en la estimación clásica. Si no se dispone de buena información a priori sobre el valor del vector de parámetros suele ser preferible utilizar un estimador clásico.

Ejemplo: Distribución marginal uniforme Sea $\{x_k\}_{k=1}^N$ una secuencia de observaciones generada por el modelo $x_k = a + w_k$, siendo w_k un ruido blanco gaussiano de esperanza cero y varianza $\sigma^2 = 100$.

La función de densidad de probabilidad marginal del parámetro θ es uniforme entre 1.0 y 4.0.

$$p(\theta) = \begin{cases} 1/3, & \text{si } 1{,}0 \leq \theta \leq 4{,}0 \\ 0, & \text{en otro caso.} \end{cases}$$

La función de densidad de probabilidad a posteriori es

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

siendo

$$p(x|\theta) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left[-\frac{1}{2\sigma^2}(x - \mathbf{1}\theta)^T(x - \mathbf{1}\theta)\right]$$

Ahora definiendo $\bar{x} = \mathbf{1}^T x/N$ y teniendo en cuenta que

$$(x - 1\theta)^T(x - 1\theta) = (\theta - \bar{x})^T N(\theta - \bar{x}) + x^T x - \bar{x}^2$$

se obtiene

$$\begin{aligned} p(x|\theta) &= \frac{\exp(x^T x - \bar{x}^2)}{(\sqrt{2\pi\sigma^2})^N} \exp\left[-\frac{1}{2\sigma^2/N}(\theta - \bar{x})^2\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\left[-\frac{1}{2\sigma^2/N}(\theta - \bar{x})^2\right] \end{aligned}$$

la segunda igualdad proviene de que la integral de la función de densidad de probabilidad condicionada debe ser uno. La función de densidad a posteriori puede ponerse como

$$p(\theta|x) = \begin{cases} c \exp\left[-\frac{1}{2\sigma^2/N}(\theta - \bar{x})^2\right] & \text{si } 1{,}0 \leq \theta \leq 4{,}0 \\ 0 & \text{en otro caso.} \end{cases}$$

donde c es la constante de normalización de la función de densidad, para que su integral sea uno.

$$c = \left(\int_{1,0}^{4,0} \exp\left[-\frac{1}{2\sigma^2/N}(\theta - \bar{x})^2 d\theta\right]\right)^{-1}$$

En la figura 2 se muestra la función de densidad marginal y a posteriori y dos estimadores para $a = 2{,}0$, $\sigma = 10$ y dos conjuntos de observaciones de tamaños $N_1 = 100$ y $N_2 = 400$.

En la figura se ve una característica de los estimadores bayesianos, según el número de observaciones aumenta, la varianza de la distribución de probabilidad a posteriori es menor y se concentra en torno al valor de su esperanza, de forma que la influencia de la distribución marginal es cada vez menor al aumentar el número de observaciones. Sin embargo, si el conocimiento previo codificado en la distribución de probabilidad es erróneo, también lo será el valor estimado.

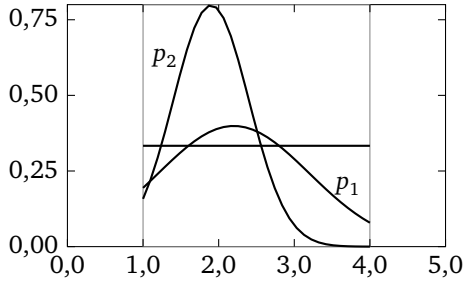

Figura 2: Distribuciones de probabilidad marginal (uniforme entre 1 y 4) y a posteriori (gaussianas truncadas) para un conjunto de observaciones generado con $a = 2$, $\sigma = 10$, y tamaños de muestra $N_1 = 100$ y $N_2 = 400$.

1.8.1. Distribución de probabilidad condicionada para variables aleatorias gaussianas

La distribución de probabilidad normal tiene la propiedad de que puede descomponerse analíticamente de forma simple para obtener la distribución de probabilidad condicionada.

El resultado es el siguiente.

Sean x e y dos vectores de variables aleatorias de dimensiones N_x y N_y, respectivamente con distribución de probabilidad conjunta gaussiana caracterizada por la función de densidad de probabilidad

$$p(z) = \frac{1}{\sqrt{(2\pi)^N \det C}} \exp\left[-\frac{1}{2}(z-\mu)^T C^{-1}(z-\mu)\right]$$

siendo $z = [x^T \; y^T]^T \in \mathbb{R}^N$ con $N = N_x + N_y$, vector de esperanzas $\mu = [\mu_x^T \; \mu_y^T]^T$ y matriz de covarianza

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$$

La función de densidad de probabilidad conjunta se puede factorizar como

$$p(z) = p(x,y) = p(x|y)p(y)$$

siendo $p(x|y)$ la función de densidad de probabilidad condicionada que también es gaussiana caracterizada por el vector de medias $\mu_{y|x}$ y la matriz de covarianza $C_{x|y}$

$$\begin{aligned} \mu_{x|y} &= \mu_x + C_{xy}C_{yy}^{-1}(y-\mu_y) \\ C_{x|y} &= C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} \end{aligned}$$

Para ello se utiliza la descomposición de Schur

$$\begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = \begin{bmatrix} I & C_{xy}C_{yy}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} I & 0 \\ C_{yy}^{-1}C_{yx} & I \end{bmatrix}$$

de donde se obtiene $\det(C) = \det(C_{xx} - C_{xy}C_{yy}^{-1}C_{yx})\det(C_{yy})$, además la inversa de la matriz de covarianza C es:

$$\begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -C_{yy}^{-1}C_{yx} & I \end{bmatrix} \begin{bmatrix} C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} & 0 \\ 0 & C_{yy} \end{bmatrix}^{-1} \begin{bmatrix} I & -C_{xy}C_{yy}^{-1} \\ 0 & I \end{bmatrix}$$

entonces llamando $C_{x|y} = C_{xx} - C_{xy}C_{yy}^{-1}C_{yx}$ y $\mu_{x|y} = \mu_x + C_{xy}C_{yy}^{-1}(y - \mu_y)$

$$\begin{aligned}
&\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^T \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix} \\
&\qquad = \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -C_{yy}^{-1}C_{yx} & I \end{bmatrix} \begin{bmatrix} C_{x|y} & 0 \\ 0 & C_{yy} \end{bmatrix}^{-1} \begin{bmatrix} I & -C_{xy}C_{yy}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix} \\
&\qquad = \begin{bmatrix} x-\mu_{x|y} \\ y-\mu_y \end{bmatrix}^T \begin{bmatrix} C_{x|y} & 0 \\ 0 & C_{yy} \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_{x|y} \\ y-\mu_y \end{bmatrix}
\end{aligned}$$

lo que demuestra que $p(x, y) = p(x|y)p(y)$ siendo $p(x|y)$ una función de densidad de probabilidad gaussiana de media $\mu_{x|y}$ y matriz de covarianzas $C_{x|y}$.

1.8.2. Estimador bayesiano gaussiano

Sean x e y vectores de variables aleatorias gaussianas caracterizadas por las siguientes propiedades estadísticas

$$\mathbf{E}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \mathbf{Cov}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$$

Supongamos que y es el vector de observaciones y x es el vector que se desea estimar a partir de las observaciones utilizando un estimador bayesiano que minimize el error cuadrático medio, entonces el estimador y su covarianza son:

$$\begin{aligned}
\hat{x} &= \mathbf{E}(x|y) \\
&= \mu_x + C_{xy}C_{yy}^{-1}(y - \mu_y) \\
\mathbf{Cov}(y|x) &= C_{xx} - C_{xy}C_{yy}^{-1}C_{yx}
\end{aligned}$$

Una propiedad importante del estimado bayesiano es que es una combinación lineal de la media del vector de estimación que proporciona la distribución de probabilidad a priori y de las observaciones. Si no hay observaciones, el estimador coincide con la media de la distribución de probabilidad a priori. Según aumenta el número de observaciones, la influencia del segundo término es mayor, de forma que cuando el número de observaciones es suficientemente grande la contribución de la esperanza a priori es despreciable.

Otra importante propiedad del estimador bayesiano de error cuadrático medio mínimo es que el error de predicción es independiente de la observación. Se define el error de predicción como:

$$\begin{aligned}
\tilde{x} &= x - \hat{x} \\
&= x - \mathbf{E}(x|y)
\end{aligned}$$

En efecto, considérese el vector de variables aleatorias $[x^T\ \tilde{y}^T]^T$, su distribución de probabilidad es gaussiana con vector de esperanza y matriz de covarianza dados por las siguientes expresiones

$$\begin{aligned}
\mathbf{E}\begin{bmatrix} \tilde{x} \\ y \end{bmatrix} &= \begin{bmatrix} \mu_x \\ 0 \end{bmatrix} \\
\mathbf{Cov}\begin{bmatrix} \tilde{x} \\ y \end{bmatrix} &= \begin{bmatrix} C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} & 0 \\ 0 & C_{yy} \end{bmatrix}
\end{aligned}$$

de otra forma

$$\begin{aligned}
\mathbf{E}(\hat{x}(y-\mu_y)^T) &= \mathbf{E}(\mathbf{E}(x|y)(y-\mu_y)^T) \\
&= \mathbf{E}(\mathbf{E}(x|y)y^T) - \mathbf{E}(\mathbf{E}(x|y)\mu_y)^T) \\
&= \int_y \int_x xp(x|y)dx\, y^T p(y)dy - \int_y \int_x xp(x|y)dx\, \mu_y^T p(y)dy \\
&= \int_y \int_x x(y-\mu_y)^T p(x,y)dx\,dy \\
&= \mathbf{E}(x(y-\mu_y)^T)
\end{aligned}$$

La condición anterior equivale a

$$0 = \mathbf{E}((\hat{x}-x)(y-\mu_y)^T) = \mathbf{Cov}(\tilde{x}, y)$$

1.8.3. Estimador lineal bayesiano

Sean x e y dos vectores de variables aleatorias de dimensiones N_x y N_y, cuyos momentos hasta el segundo orden son conocidos. A diferencia del apartado anterior, no se conoce la distribución de probabilidad conjunta.

Sea $z = [x^T\ y^T]^T$. Su esperanza y covarianza son μ y C que vienen dados por las expresiones

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix}$$

Se desea obtener el mejor estimador lineal de la variable x a partir de una observación de la variable y. Las expresiones del estimador y su covarianza son las mismas del caso gaussiano

$$\begin{aligned}
\hat{x} &= \mu_x + C_{xy}C_{yy}^{-1}(y-\mu_y) \\
\mathbf{Cov}(\hat{x}) &= C_{xx} - C_{xy}C_{yy}^{-1}C_{yx}
\end{aligned}$$

Conviene señalar que aunque las expresiones son las mismas, en este caso no equivalen a la esperanza condicionada $\mathbf{E}(x|y)$ y a la covarianza condicionada. Para calcular los momentos condicionados, debe disponerse de las función de distribución de probabilidad conjunta, que no es conocida en este caso. Sin embargo, la suposición de linealidad es clave para obtener las expresiones.

El estimador se obtiene minimizando el error cuadrático medio de estimación, utilizando como estimador una función lineal de la observación $\hat{x} = L(y) = Ay + b$.

Para calcular el estimador, se tiene en cuenta que las variables aleatorias $x - \mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y)$ e $y - \mu_y$ son independientes. En efecto,

$$\begin{aligned}
\mathbf{E}\begin{bmatrix} x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y) \\ y-\mu_y \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\mathbf{Cov}\begin{bmatrix} x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y) \\ y-\mu_y \end{bmatrix} &= \begin{bmatrix} C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} & 0 \\ 0 & C_{yy} \end{bmatrix}
\end{aligned}$$

Por tanto, $x - \mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y)$ es independiente de cualquier función lineal de y.

Entonces definiendo $\hat{x} = \mu_x + C_{xy}C_{yy}^{-1}(y-\mu_y) + \phi(y)$, siendo $\phi(y)$ una función lineal de y, el error cuadrático medio de estimación es:

$$\begin{aligned}
&\mathbf{E}[(x-\hat{x})^T(x-\hat{x})] \\
&\quad = \mathbf{E}[(x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y) - \phi(y))(x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y) - \phi(y))^T] \\
&\quad = \mathbf{E}[(x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y))^T(x-\mu_x - C_{xy}C_{yy}^{-1}(y-\mu_y))] + \mathbf{E}[\phi^T(y)\phi(y))]
\end{aligned}$$

donde los términos cruzados desaparecen por independencia de $\hat{x}$ y $\phi(y)$. Como la expresión anterior es una suma de cuadrados, el mínimo se obtiene para $\phi(y) = 0$ y por tanto se concluye que $\hat{x}$ es el estimador lineal que minimiza el error cuadrático medio.

1.8.4. Estimador bayesiano de un modelo lineal

Considérese el modelo lineal

$$x = A\theta + w$$

siendo θ el vector que se desea estimar, que se considera aleatorio con distribución de probabilidad gaussiana de media μ_θ y covarianza C_θ. El vector de variables aleatorias w es el ruido del modelo, su distribución de probabilidad es gaussiana de media cero y covarianza C_w y se considera que los vectores θ y w son independientes entre sí, es decir

$$\mathbf{E}((\theta-\mu_\theta)w^T) = 0$$

La esperanza del vector de observación es

$$\begin{aligned}
\mathbf{E}(x) &= A\mathbf{E}(\theta) + \mathbf{E}(w) \\
&= A\mu_\theta = \mu_x
\end{aligned}$$

y la matriz de covarianza conjunta de $[x^T\ \theta^T]^T$ es

$$\begin{aligned}
C &= \mathbf{E}\left[\begin{bmatrix} x-\mu_x \\ \theta-\mu_\theta \end{bmatrix}\begin{bmatrix} x-\mu_x \\ \theta-\mu_\theta \end{bmatrix}^T\right] \\
&= \mathbf{E}\left[\begin{bmatrix} A(\theta-\mu_\theta)+w \\ \theta-\mu_\theta \end{bmatrix}\begin{bmatrix} A(\theta-\mu_\theta)+w \\ \theta-\mu_\theta \end{bmatrix}^T\right] \\
&= \begin{bmatrix} A \\ I \end{bmatrix}\mathbf{E}((\theta-\mu_\theta)(\theta-\mu_\theta)^T)\begin{bmatrix} A \\ I \end{bmatrix}^T + 2\begin{bmatrix} A \\ I \end{bmatrix}\mathbf{E}((\theta-\mu_\theta)(w-\mu_w)^T)\begin{bmatrix} I \\ 0 \end{bmatrix} \\
&\quad + \begin{bmatrix} I \\ 0 \end{bmatrix}\mathbf{E}((w-\mu_w)(w-\mu_w)^T)\begin{bmatrix} I \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} A \\ I \end{bmatrix} C_{\theta\theta}\begin{bmatrix} A \\ I \end{bmatrix}^T + \begin{bmatrix} I \\ 0 \end{bmatrix} C_{ww}\begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} AC_{\theta\theta}A^T + C_{ww} & AC_{\theta\theta} \\ C_{\theta\theta}A^T & C_{\theta\theta} \end{bmatrix}
\end{aligned}$$

El estimador que minimiza el error cuadrático medio y su covarianza vienen dados por la siguientes expresiones:

$$\begin{aligned}
\hat{\theta} &= \mu_\theta + AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1}(x - A\mu_\theta) \\
\mathbf{Cov}(\theta|x) &= C_{\theta\theta} - AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1}C_{\theta\theta}A^T
\end{aligned}$$

Ejemplo: Distribución a priori gaussiana Sea $\{x_k\}_{k=1}^N$ una secuencia de observaciones generada por el modelo $x_k = a + w_k$, siendo w_k un ruido blanco gaussiano de esperanza cero y varianza $\sigma^2 = 1$.

La función de densidad de probabilidad marginal del parámetro θ es gaussiana de media μ_0 y varianza σ_0^2, es decir

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right]$$

La función de densidad de probabibilidad a posteriori se obtiene de la expresión

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

siendo

$$p(x|\theta) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left[-\frac{1}{2\sigma^2}(x - \mathbf{1}\theta)^T(x - \mathbf{1}\theta)\right]$$

Definiendo las variables aleatorias $x = \mathbf{1}\theta + w$ e $y = \theta$, el estimador y su covarianza tienen las expresiones:

$$\begin{aligned} \hat{\theta} &= \mu_y + C_{yx}C_{xx}^{-1}(x - \mu_x) \\ \mathbf{Cov}(\theta) &= C_{yy} - C_{yx}C_{xx}^{-1}C_{xy} \end{aligned}$$

Teniendo en cuenta que

$$\begin{aligned} \mu_x &= \mathbf{1}m_0 \\ \mu_y &= m_0 \\ C_{xx}^{-1} &= (\sigma_0^2\mathbf{1}\mathbf{1}^T + \sigma^2 I)^{-1} = \frac{1}{\sigma^2}I - \mathbf{1}\frac{1}{\sigma^2}\frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/N}}\frac{1}{\sigma^2}\mathbf{1}^T \\ C_{yy} &= \sigma_0^2 \\ C_{xy} &= \mathbf{1}\sigma_0^2 \end{aligned}$$

y sustituyendo el las expresiones del estimador y su covarianza se obtiene:

$$\begin{aligned} \hat{\theta} &= \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/N}\bar{x} + \frac{\sigma^2/N}{\sigma_0^2 + \sigma^2/N}m_0 \\ \mathbf{Cov}(\theta|x) &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/N}} \end{aligned}$$

La distribución de probabilidad del parámetro θ condicionada a las observaciones es una gaussiana de media $\hat{\theta}$ y covarianza $\mathbf{Cov}(\theta|x)$.

Como ya se ha indicado al aumentar el número de observaciones la distribución de probabilidad a posteriori se concentra en torno a su esperanza y la influencia de la distribución de probabilidad marginal de θ es cada vez menor.

1.8.5. Relación entre el estimador clásico y el bayesiano para un modelo lineal

Sea el modelo lineal

$$x = A\theta + w$$

siendo

$$\mathbf{E}\begin{bmatrix} w \\ \theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{E}\begin{bmatrix} w \\ \theta \end{bmatrix}\begin{bmatrix} w \\ \theta \end{bmatrix}^T = \begin{bmatrix} C_{ww} & 0 \\ 0 & C_{\theta\theta} \end{bmatrix}$$

Las expresiones del estimador bayesiano son:

$$\hat{\theta} = AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1}x$$
$$\mathbf{Cov}(\theta|x) = C_{\theta\theta} - AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1}C_{\theta\theta}A^T$$

Teniendo en cuenta las igualdades

$$AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1} = (C_{\theta\theta}^{-1} + A^T C_{ww}^{-1}A)^{-1}A^T C_{ww}^{-1}$$
$$C_{\theta\theta} - AC_{\theta\theta}(AC_{\theta\theta}A^T + C_{ww})^{-1}C_{\theta\theta}A^T = (C_{\theta\theta}^{-1} + A^T C_{ww}^{-1}A)^{-1}$$

las expresiones del estimador bayesiano equivalen a

$$\hat{\theta} = (C_{\theta\theta}^{-1} + A^T C_{ww}^{-1}A)^{-1}A^T C_{ww}^{-1}x$$
$$\mathbf{Cov}(\theta|x) = (C_{\theta\theta}^{-1} + A^T C_{ww}^{-1}A)^{-1}$$

que son aproximadamente iguales a las expresiones del estimador de mínimos cuadrados si $C_{\theta\theta}$ es muy grande.

Por tanto la expresión del estimador bayesiano es aproximadamente igual a la del estimador de mínima varianza para un modelo lineal donde la esperanza y la inversa de la matriz de covarianza a priori del estimador tienden a cero.

1.9. El filtro de Kalman

El filtro de Kalman es una de las aplicaciones más importantes de la estimación bayesiana al campo de los sistemas dinámicos lineales. Dado un sistema lineal de parámetros variables con perturbaciones de proceso y en la observación, modeladas como procesos estocásticos, el filtro de Kalman proporciona el estimador bayesiano de error cuadrático medio mínimo del vector de estados del sistema x_k en el instante de tiempo k a partir de observaciones y_ℓ para $\ell \leq k$.

Aplicando teoría bayesiana de estimación, el filtro de Kalman es la esperanza del estado x_k condicionada a las observaciones y_ℓ para $\ell \leq k$.

Si los procesos aleatorios que caracterizan las perturbaciones son gaussianas, entonces la distribución de probabilidad del estado condicionada a las observaciones es gaussiana y queda completamente caracterizada por su esperanza y su matriz de covarianza, por tanto el filtro de Kalman estaría proporcionando la distribución de probabilidad a posteriori del estado. En otro caso, el filtro de Kalman solo proporciona los momentos hasta el segundo orden.

El filtro de Kalman queda expresado mediante un algoritmo recursivo en dos pasos. En el primer paso, llamado de predicción, se obtiene la esperanza y la covarianza del estado x_k condicionadas a las observaciones hasta el instante $k-1$. En el segundo paso, llamado de corrección, se modifica el valor de la predicción una vez que se dispone de la observación y_k. Comencemos obteniendo el predictor de Kalman de horizonte uno.

El filtro de Kalman se obtendrá suponiendo que la distribución de probabilidad conjunta de las perturbaciones y de las condiciones iniciales es gaussiana y conocida. Sin embargo, como se ha visto en el apartado 1.8.3, los resultados de la teoría de estimación bayesiana general para distribuciones gaussianas son válidos para estimación lineal con los momentos hasta el segundo orden conocidos.

1.9.1. El predictor de Kalman

Sea el modelo dinámico lineal

$$\begin{aligned} \hat{x}_{k+1} &= A_k x_k + w_k \\ y_k &= C_k x_k + v_k \end{aligned}$$

siendo $[w_k^T \; v_k^T]^T$ un ruido blanco gaussiano de esperanza cero y covarianza

$$\mathbf{Cov} \begin{bmatrix} w_k \\ v_k \end{bmatrix} = \begin{bmatrix} Q_k & S_k \\ S_k^T & R_k \end{bmatrix}$$

El estado inicial x_0 es una variable gaussiana de esperanza cero y covarianza Π_0 e independiente de w_k y de v_k para todo k.

Se denota como Y_k el vector columna que contiene las observaciones desde el instante inicial hasta el instante de tiempo k.

$$Y_k = \mathbf{col}(y_0, y_1, \ldots, y_k)$$

Sea $\hat{x}_{k|j} = \mathbf{E}(x_k|Y_j)$ la esperanza matemática condicionada del vector de estado x_k a las observaciones hasta el instante de tiempo j, además se define $\hat{y}_{k|j} = \mathbf{E}(y_k|Y_j) = C_k \hat{x}_{k|j}$, $\tilde{x}_{k|j} = x_k - \hat{x}_{k|j}$, $\tilde{y}_{k|j} = y_k - \hat{y}_{k|j}$, y $P_{k|j} = \mathbf{Cov}[x_k|Y_j] = \mathbf{E}[\tilde{x}_{k|j}\tilde{x}_{k|j}^T|Y_j]$.

Supongamos que en el instante k se dispone del estimador del vector de estado $\hat{x}_{k|k-1}$ y se desea obtener $\hat{x}_{k+1|k}$ utilizando estimación bayesiana. Se comienza calculando las esperanzas y covarianza del vector $[x_{k+1} \; y_k]^T$ condicionadas a las observaciones Y_{k-1}.

$$\begin{aligned} \mathbf{E}\left[\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix}|Y_{k-1}\right] &= \begin{bmatrix} A_k \\ C_k \end{bmatrix} \mathbf{E}[x_k|Y_{k-1}] + \mathbf{E}\left[\begin{bmatrix} w_k \\ v_k \end{bmatrix}|Y_{k-1}\right] \\ &= \begin{bmatrix} A_k \\ C_k \end{bmatrix} \hat{x}_{k|k-1} \\ \mathbf{Cov}\left[\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix}|Y_{k-1}\right] &= \begin{bmatrix} A_k \\ C_k \end{bmatrix} \mathbf{Cov}(x_k|Y_{k-1}] \begin{bmatrix} A_k \\ C_k \end{bmatrix}^T + \mathbf{Cov}\left[\begin{bmatrix} w_k \\ v_k \end{bmatrix}|Y_{k-1}\right] \\ &= \begin{bmatrix} A_k \\ C_k \end{bmatrix} P_{k|k-1} \begin{bmatrix} A_k \\ C_k \end{bmatrix}^T + \begin{bmatrix} Q_k & S_k \\ S_k^T & R_k \end{bmatrix} \end{aligned}$$

Una vez disponible la observación y_k, las esperanzas y covarianzas condicionadas a las observaciones Y_k se obtienen a partir de la distribución de probabilidad a posteriori utilizando las fórmulas:

$$\begin{aligned} \mathbf{E}(x_{k+1}|Y_k) &= \mathbf{E}(x_{k+1}|Y_{k-1}) + \mathbf{E}(\tilde{x}_{k+1|k-1}\tilde{y}_{k|k-1}^T|Y_{k-1})(\mathbf{Cov}(y_k|Y_{k-1}))^{-1}(y_k - \mathbf{E}(y_k|Y_{k-1})) \\ \mathbf{Cov}(x_{k+1}|Y_k) &= \mathbf{Cov}(x_{k+1}|Y_{k-1}) - \mathbf{E}(\tilde{x}_{k+1|k-1}\tilde{y}_{k|k-1}^T|Y_{k-1})(\mathbf{Cov}(y_k|Y_{k-1}))^{-1}(y_k - \mathbf{E}(y_k|Y_{k-1})) \end{aligned}$$

Sustituyendo, se obtienen las siguientes expresiones:

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_{k|k-1} + (A_k P_{k|k-1} C_k^T + S_k)(C_k P_{k|k-1} C_k^T + R_k)^{-1}(y_k - C_k \hat{x}_{k|k-1}) \\ P_{k+1|k} &= A_k P_{k|k-1} A_k^T + Q_k - (A_k P_{k|k-1} C_k^T + S_k)(C_k P_{k|k-1} C_k^T + R_k)^{-1}(A_k P_{k|k-1} C_k^T + S_k)^T\end{aligned}$$

Condiciones iniciales:

$$\begin{aligned}\hat{x}_{0|-1} &= \mathbf{E}(x_0) \\ P_{0|-1} &= \mathbf{Cov}(x_0) = \Pi_0\end{aligned}$$

1.9.2. El proceso de innovación

La innovación es el proceso aleatorio $e_k = y_k - \hat{y}_{k|k-1}$ que representa la parte no explicada de la observación. La innovación e_k es independiente de las observaciones contenidas en Y_{k-1}, es decir para y_i para $i < k$. Esta es una de las propiedades del estimador bayesiano de error cuadrático medio mínimo. Más aún, como estas observaciones pueden ponerse como $y_i = \hat{y}_{i|i-1} + e_i$, la innovación e_k es independiente de las observaciones e_i para $i < k$. Esto indica que el proceso aleatorio de la innovaciones es un ruido blanco.

Es decir la innovación satisface

$$\begin{aligned}\mathbf{E}(e_k) &= 0 \\ \mathbf{Cov}((y_i - \hat{y}_{i|i-1}), e_k) &= 0, \quad i < k \\ \mathbf{Cov}(e_i, e_k) &= 0, \quad i < k\end{aligned}$$

Las anteriores propiedades de la innovación permiten obtener la siguiente expresión: La observación es igual a la suma de la estimación obtenida con el predictor de Kalman más un proceso de ruido blanco

$$y_k = \mathbf{E}(y_k|Y_{k-1}) + e_k$$

Desarrollando la expresión anterior para $i < k$ se obtiene:

$$\begin{aligned}y_0 &= e_0 \in \mathscr{L}(e_0) \\ y_1 &= \mathbf{E}(y_1|Y_0) + e_1 = \mathbf{E}(y_1|e_0) + e_1 \in \mathscr{L}(e_0, e_1) \\ y_2 &= \mathbf{E}(y_2|Y_1) + e_2 = \mathbf{E}(y_2|e_0, e_1) + e_2 \in \mathscr{L}(e_0, e_1, e_2) \\ &\vdots \\ y_k &= \mathbf{E}(y_k|Y_{k-1}) + e_k = \mathbf{E}(y_2|e_0, e_1, \ldots, e_{k-1}) + e_k \in \mathscr{L}(e_0, e_1, \ldots, e_{k-1}, e_k)\end{aligned}$$

de donde se deduce que el espacio lineal generado por las observaciones $(y_0, y_1, \ldots, y_k)$ coincide con el espacio lineal generado por las innovaciones $(e_0, e_1, \ldots, e_k)$,

$$\mathscr{L}(y_0, y_1, \ldots, y_k) = \mathscr{L}(e_0, e_1, \ldots, e_k)$$

además, la estimación bayesiana de error cuadrático medio mínimo $\mathbf{E}(y_k|Y_{k-1})$ pertenece al espacio lineal generado por las innovaciones $(e_0, e_1, \ldots, e_{k-1})$. Debido a que las innovaciones son independientes entre sí, la expresión del estimador es:

$$\begin{aligned}\mathbf{E}(y_k|Y_{k-1}) &= \mathbf{E}(y_k|e_0, e_1, \ldots, e_{k-1}) \\ &= \sum_{i=1}^{k-1} \mathbf{Cov}(y_k, e_i)(\mathbf{Cov}(e_i, e_i))^{-1} e_i \\ &= \sum_{i=1}^{k-1} \pi(y_k, e_i) e_i\end{aligned}$$

siendo

$$\pi(y_k, e_i) = \mathbf{Cov}(y_k, e_i)(\mathbf{Cov}(e_i, e_i))^{-1}$$

Entonces la relación entre las observaciones y las innovaciones viene dada por la expresión

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ \pi(y_1, e_0) & I & 0 & \cdots & 0 \\ \pi(y_2, e_0) & \pi(y_2, e_1) & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi(y_k, e_0) & \pi(y_k, e_1) & \pi(y_k, e_2) & \cdots & I \end{bmatrix} \begin{bmatrix} e_0 \\ e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}$$

La innovación juega un papel fundamental en la estimación bayesiana. En general el estimador bayesiano de error cuadrático medio de un proceso aleatorio s_{k+1} a partir de las observaciones $Y_k = \mathbf{col}(y_0, \ldots, y_k)$ se puede expresar mediante la siguiente expresión en función de las innovaciones

$$\begin{aligned} \hat{s}_{k+1|k} &= \mathbf{E}(s_{k+1}|Y_k) \\ &= \sum_{i=1}^{k} \mathbf{Cov}(s_{k+1}, e_k)(\mathbf{Cov}(e_k, e_k))^{-1} e_k \end{aligned}$$

Los resultados obtenidos tienen una interesante interpretación geométrica. La operador $\pi(\cdot, e_k)e_k$ puede interpretarse como la proyección ortogonal sobre la innovación e_k. Por tanto, el estimador bayesiano de error cuadrático medio mínimo de una variable aleatoria se obtiene proyectando dicha variable sobre el espacio lineal generado por las innovaciones que son forman una base ortogonal del espacio lineal generado por las observaciones y que se pueden calcular de forma recursiva según van llegando las observaciones. Por ejemplo, el conocido método de ortogonalización de Gram-Schmidt podría ser utilizado para ir generando las innovaciones en tiempo real según van llegando las observaciones.

Algoritmo de Gram-Schmidt El algoritmo actúa de la siguiente forma:

0. Inicialización:

$$e_0 = y_0$$

1. Para $1 < k$,

   1.1. Cálculo de la innovación e_k

$$\begin{aligned} \hat{y}_{k|k-1} &= \sum_{i=1}^{k-1} \pi(y_k, e_i)e_i \\ e_k &= y_k - \hat{y}_{k|k-1} \end{aligned}$$

   1.2. Cálculo de la predicción $\hat{s}_{k|k-1}$:

$$\hat{s}_{k+1|k} = \sum_{i=1}^{k} \pi(s_{k+1}, e_i)e_i$$

Para finalizar, es interesante notar que la expresión

$$\hat{x}_{k+1|k} = \sum_{i=0}^{k} \pi(x_{i+1}, e_i)e_i$$

permite derivar fácilmente las expresiones del predictor de Kalman.

1.9.3. El filtro de Kalman

El filtro de Kalman calcula el estimador bayesiano de error cuadrático medio del estado en el instante k utilizando las observaciones hasta ese mismo instante Y_k. Es decir, el filtro de Kalman calcula

$$\hat{x}_{k+1|k+1} = \mathbf{E}[x_{k+1}|Y_{k+1}]$$

y la covarianza

$$P_{k+1|k+1} = \mathbf{Cov}[x_{k+1}|Y_{k+1}]$$

La expresión del estimador del estado que proporciona el filtro de Kalman puede obtenerse corrigiendo el valor obtenido mediante el predictor de Kalman una vez que la observación y_k está disponible

Las esperanzas y covarianzas a priori se obtienen directamente del predictor de Kalman y tienen las expresiones:

$$\begin{aligned}
\mathbf{E}\left[\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} \Big| Y_k\right] &= \begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{y}_{k+1|k} \end{bmatrix} \\
\mathbf{Cov}\left[\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} \Big| Y_k\right] &= \begin{bmatrix} \mathbf{Cov}(x_{k+1}|Y_k) & C_{k+1}\mathbf{Cov}(x_k|Y_k) \\ \mathbf{Cov}(x_{k+1}|Y_k)C_{k+1}^T & C_{k+1}\mathbf{Cov}(x_{k+1}|Y_{k-1})C_{k+1}^T + \mathbf{Cov}(v_{k+1}|Y_k) \end{bmatrix} \\
&= \begin{bmatrix} P_{k+1|k} & C_{k+1}P_{k+1|k} \\ P_{k+1|k}C_{k+1}^T & C_{k+1}P_{k+1|k}C_{k+1}^T + R_{k+1} \end{bmatrix}
\end{aligned}$$

en el cálculo de las covarianzas se ha tenido en cuenta que

$$\begin{aligned}
\mathbf{E}[\tilde{x}_{k+1|k}v_{k+1}^T|Y_k] &= \mathbf{E}[(A_k\tilde{x}_{k|k} + w_k)v_{k+1}^T|Y_k] \\
&= A_k\mathbf{E}(\tilde{x}_{k|k}v_{k+1}^T|Y_k) + \mathbf{E}(w_k v_{k+1}^T|Y_k) \\
&= 0
\end{aligned}$$

ya que v_{k+1} es independiente de $\tilde{x}_{i|i}$ y de w_i para $i \leq k$.

El estimador a posteriori del estado $\hat{x}_{k+1|k+1}$ y su covarianza $P_{k+1|k+1}$ se obtienen aplicando las fórmulas de la esperanza y covarianza condicionadas

$$\begin{aligned}
\hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + (P_{k+1|k}C_{k+1}^T)(C_{k+1}P_{k+1|k}C_{k+1}^T + R_{k+1})^{-1}(\hat{y}_{k+1} - \hat{y}_{k+1|k}) \\
P_{k+1|k+1} &= P_{k+1|k} - (P_{k+1|k}C_{k+1}^T)(C_{k+1}P_{k+1|k}C_{k+1}^T + R_{k+1})^{-1}(C_{k+1}P_{k+1|k})
\end{aligned}$$

Por tanto el filtro de Kalman consta de dos pasos, un paso de predicción y otro de corrección que se expresan de la siguiente forma:

- Condiciones iniciales:

$$\begin{aligned}
\hat{x}_{0|-1} &= \mathbf{E}(x_0) = 0 \\
P_{0|-1} &= \mathbf{Cov}(x_0) = \Pi_0
\end{aligned}$$

- Paso de predicción: Se dispone de $\hat{x}_{k|k-1}$, $\hat{y}_{k|k-1}$, $P_{k|k-1}$ y la observaciones $Y_k = \mathbf{col}(y_1 \dots y_k)$, por tanto se puede llevar a cabo justo después de observar y_k.

$$\begin{aligned} L_{k+1|k} &= (A_k P_{k|k-1} C_k^T + S_k)(C_k P_{k|k-1} C_k^T + R_k)^{-1} \\ \hat{x}_{k+1|k} &= A_k \hat{x}_{k|k-1} + L_{k+1|k}(y_k - \hat{y}_{k|k-1}) \\ \hat{y}_{k+1|k} &= C_{k+1} \hat{x}_{k+1|k} \\ P_{k+1|k} &= A_k P_{k|k-1} A_k^T + Q_k - L_{k+1|k}(C_k P_{k|k-1} C_k^T + R_k) L_{k+1|k}^T \end{aligned}$$

- Paso de corrección: Se realiza en el momento en que llega la observación y_{k+1}.

$$\begin{aligned} L_{k+1|k+1} &= (P_{k+1|k} C_{k+1}^T)(C_{k+1} P_{k+1|k} C_{k+1}^T + R_{k+1})^{-1} \\ \hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + L_{k+1|k+1}(y_{k+1} - \hat{y}_{k+1|k}) \\ \hat{y}_{k+1|k+1} &= C_{k+1} \hat{x}_{k+1|k+1} \\ P_{k+1|k+1} &= P_{k+1|k} - L_{k+1|k+1}(C_{k+1} P_{k+1|k} C_{k+1}^T + R_{k+1}) L_{k+1|k+1}^T \end{aligned}$$

En el caso de que w_k y v_k sean independientes, es decir si $S_k = \mathbf{E}(w_k v_k^T) = 0$, la expresión del filtro de Kalman queda simplificada ya que

$$\begin{aligned} L_{k+1|k} &= A_k P_{k|k-1} C_k^T (C_k P_{k|k-1} C_k^T + R_k)^{-1} = A_k L_{k|k} \\ \hat{x}_{k+1|k} &= A_k \hat{x}_{k|k-1} + A_k L_{k|k}(y_k - \hat{y}_{k|k-1}) = A_k \hat{x}_{k|k} \\ P_{k+1|k} &= A_k P_{k|k-1} A_k^T + Q_k - A_k L_{k|k}(C_k P_{k|k-1} C_k^T + R_k) L_{k|k}^T A_k^T = A_k P_{k|k} A_k^T + Q_k \end{aligned}$$

En este caso, las ecuaciones del filtro de Kalman son:

- Condiciones iniciales:

$$\begin{aligned} \hat{x}_{0|0} &= \mathbf{E}(x_0) = 0 \\ P_{0|0} &= \mathbf{Cov}(x_0) = \Pi_0 \end{aligned}$$

- Paso de predicción: Se dispone de $\hat{x}_{k|k}$, $\hat{y}_{k|k}$, $P_{k|k}$ y la observaciones $Y_k = \mathbf{col}(y_1 \dots y_k)$ por tanto se puede llevar a cabo en el momento en que se dispone de la observación y_k.

$$\begin{aligned} L_{k+1|k} &= A_k L_{k|k} \\ \hat{x}_{k+1|k} &= A_k \hat{x}_{k|k} \\ \hat{y}_{k+1|k} &= C_{k+1} \hat{x}_{k+1|k} \\ P_{k+1|k} &= A_k P_{k|k} A_k^T + Q_k \end{aligned}$$

- Paso de corrección: Se realiza en el momento en que llega la observación y_{k+1}.

$$\begin{aligned} L_{k+1|k+1} &= (P_{k+1|k} C_{k+1}^T)(C_{k+1} P_{k+1|k} C_{k+1}^T + R_{k+1})^{-1} \\ \hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + L_{k+1|k+1}(y_{k+1} - \hat{y}_{k+1|k}) \\ \hat{y}_{k+1|k+1} &= C_{k+1} \hat{x}_{k+1|k+1} \\ P_{k+1|k+1} &= P_{k+1|k} - L_{k+1|k+1}(C_{k+1} P_{k+1|k} C_{k+1}^T + R_{k+1}) L_{k+1|k+1}^T \end{aligned}$$