Familias Exponenciales e Inferencia Bayesiana

Manuel Mendoza Ramírez[2] y Eduardo Gutiérrez-Peña[1]

[1] Departamento de Estadística, ITAM

[2] Departamento de Probabilidad y Estadística, IIMAS-UNAM

II Encuentro Conjunto RSME - SMM
Málaga, España. Enero 17-20, 2012

Contenido

Famila Exponencial Natural (FEN)

- Densidad

$$p_\theta(x|\theta) = b(x)\exp\{\theta x - M(\theta)\} \qquad (\theta \in \Xi)$$

$$\text{con } M(\theta) = \log \int b(x)\ \exp\{\theta x\}\, \eta(dx)$$

- Espacio parametral (canónico)

$$\Xi = \{\theta \in \mathbb{R} : M(\theta) < \infty\}$$

- Parámetro medio

$$\mu = \mu(\theta) = E[\,X|\theta\,] = \mathrm{d}M(\theta)/\mathrm{d}\theta$$

espacio del parámetro medio: $\Omega = \mu(\Xi)$

- Función de varianza

$$V(\mu) = \mathrm{Var}[\,X|\theta(\mu)\,] = \mathrm{d}^2 M(\theta(\mu))/\mathrm{d}\theta^2$$

Famila Exponencial Natural (FEN)

- Sean $X_1, \ldots, X_n \sim p_\theta(x|\theta)$ y $S = \sum_{i=1}^{n} X_i$. Entonces

$$p_\theta(s|\theta, n) = b(s, n) \exp\{\theta s - n\, M(\theta)\}$$

- Verosimilitud

$$L_\theta(\theta|s, n) \propto \exp\{\theta s - n\, M(\theta)\}$$

- Información de Fisher

$$i_\theta(\theta) = \mathrm{d}^2 M(\theta)/\mathrm{d}\theta^2$$

  En términos del parámetro medio

$$i_\mu(\mu) = V(\mu)^{-1}$$

Familia Conjugada

- Familia conjugada natural

$$\pi_\theta(\theta|s_0, n_0) \propto L_\theta(\theta|s_0, n_0) \qquad (s_0 \in \mathbb{R}, n_0 \in \mathbb{R})$$

- Forma normalizada

$$\pi_\theta(\theta|s_0, n_0) = \exp\{[s_0\theta - n_0\, M(\theta)] - W(s_0, n_0)\}$$

con $W(s_0, n_0) = \log \int \exp\{s_0\theta - n_0\, M(\theta)\}\mathrm{d}\theta$ (Diaconis-Ylvisaker)

- Inicial de Jeffreys

$$\pi_J(\theta) \propto i_\theta(\theta)^{1/2}$$

- La posterior de Jeffreys es miembro de la familia conjugada de DY si la FEN tiene función de varianza cuadrática (FVC)

- Entropía relativa

$$D_{KL}(f(\cdot)||g(\cdot)) = \int f(x) \ln \left\{ \frac{f(x)}{g(x)} \right\} \mathrm{d}x$$

también se conoce como la divergencia de Kullback-Liebler entre f y g.

- Información Mutua

$$\begin{aligned} I(X, \Theta) &= \int \int p(x, \theta) \ln \left\{ \frac{p(x, \theta)}{p(x)p(\theta)} \right\} \mathrm{d}x\mathrm{d}\theta \\ &= \int p(x) \left\{ \int p(\theta|x) \ln \left\{ \frac{p(\theta|x)}{p(\theta)} \right\} \mathrm{d}\theta \right\} \mathrm{d}x \\ &= \int p(x) \left\{ D_{KL}(p(\theta|x)||p(\theta)) \right\} \mathrm{d}x. \end{aligned}$$

Optimización

Las familias exponenciales de distribuciones se pueden obtener como soluciones a problemas de optimización de la entropía.

- Mínima entropía relativa: para una g fija, se minimiza $D_{KL}(f(\cdot)||g(\cdot))$ respecto a f sujeto a las restricciones

$$\int h_i(x) f(x) dx = m_i; \quad i = 1, \ldots, k$$

- Si existe una solución f^*, entonces pertenece a una familia exponencial general

Inferencia Bayesiana Paramétrica

- X: observable que describe el fenómeno de interés

$$p(x|\theta) \quad \text{y} \quad p(\theta)$$

- Elementos

$$\begin{aligned} p(x,\theta) &= p(\theta)\, p(x|\theta) \\ &= p(x)\, p(\theta|x) \end{aligned}$$

Función de pérdida: $L(\hat{\theta}, \theta)$

Inferencia Bayesiana Paramétrica

- X: observable que describe el fenómeno de interés

$$p(x|\theta) \quad \text{y} \quad p(\theta)$$

- Elementos

$$\begin{aligned} p(x,\theta) &= p(\theta)\, p(x|\theta) \\ &= p(x)\, p(\theta|x) \end{aligned}$$

Función de pérdida: $L(\hat{\theta},\theta)$

- I. Selección del modelo de muestreo $p(x|\theta)$

$$\int h_i(x)p(x|\theta)dx = m_i; \quad i = 1, \ldots, k$$

con $p(x|\theta)$ "cercana" a la medida base $b(x)$

Entropía Relativa Mínima (ERM, Kullback 1959):

La solución a este problema de optimización pertenece a una familia exponencial

$$p(x|\theta) = b(x)\exp\{\theta x - M(\theta)\}$$

Inferencia Bayesiana Paramétrica

- X: observable que describe el fenómeno de interés

$$p(x|\theta) \quad \text{and} \quad p(\theta)$$

- Elementos

$$\begin{aligned} p(x,\theta) &= p(\theta)\, p(x|\theta) \\ &= p(x)\, p(\theta|x) \end{aligned}$$

Función de pérdida: $L(\hat{\theta}, \theta)$

- II. Selección de la inicial $p(\theta)$

  De nuevo, las restricciones $+$ "cercanía" a la medida $B(\theta)$ $+$ ERM

  La solución de este problema
  es un miembro de otra familia exponencial

  Si las restricciones se formulan en términos de θ y $M(\theta)$,

  $p(\theta)$ es conjugada para $p(x|\theta)$

$$p(\theta) \propto B(\theta) \exp\{x_0\theta - n_0 M(\theta)\}$$

- III. Proceso de aprendizaje

  – Una muestra

$$X_{(n)} = (X_1, \ldots, X_n) \text{ muestra aleatoria de } p(x|\theta)$$

$p(\theta|x_{(n)})$ pertenece a la misma familia a la que pertenece $p(\theta)$ y el cálculo se simplifica

  – Varias muestras

$$X_{(n_i)} \text{ muestra aleatoria de } p(x|\theta_i); \qquad \text{i=1,...,k}$$

$p(\theta_i|x_{(n_i)})$ pertenece a la misma familia a la que pertenece $p(\theta_i)$

– Varias muestras

Si $\theta_1, \ldots, \theta_k$ son intercambiables, entonces pueden considerarse una m.a. de

$$p(\theta|x_0, n_0) \propto B(\theta) \exp\{x_0\theta - n_0 M(\theta)\}$$

Si los hiperparámetros (x_0, n_0) de la inicial común son desconocidos, se puede utilizar el argumento previo para producir una inicial $p(x_0, n_0)$

Esta inicial resuelve otro problema de optimización; pertenece a una familia exponencial y puede ser conjugada para $p(\theta|x_0, n_0)$

(George, Makov & Smith; 1993)

Iniciales Máximo Informativas

- X: observable que describe el fenómeno de interés

$$p(x|\theta) \quad \text{y} \quad p(\theta)$$

- Elementos

$$\begin{aligned} p(x,\theta) &= p(\theta)\, p(x|\theta) \\ &= p(x)\, p(\theta|x) \end{aligned}$$

$$L(\hat{\theta},\theta) = D_{KL}(p(x|\hat{\theta})||p(x|\theta))$$

Iniciales Máximo Informativas

- Información Mutua: $I(X, \Theta)$ es una funcional de $p(x, \theta)$

  - Dada $p(x)$, la minimización de $I(X, \Theta)$ con respecto a $p(\theta|x)$ sujeta a la restricción

$$\int \int p(x)p(\theta|x)L(\hat{\theta}(x), \theta)\mathrm{d}x\mathrm{d}\theta \leq I$$

    conduce a una posterior conjugada (GP & Muliere; 2004)

$$p(\theta|x) \propto \exp\{(x + x_0)\theta - (n + n_0)M(\theta)\}$$

  - En otro sentido, si se fija $p(x|\theta)$ y se *maximiza* (asintóticamente) $I(X, \Theta)$ con respecto a $p(\theta)$, se obtiene la inicial de Jeffreys (Clarke & Barron; 1994)

* Grünwald and Dawid (2004) exploran la relación entre máxima entropía y mínima pérdida esperada en el caso menos favorable. Consideran problemas de decisión y funciones de pérdida arbitrarios.

  Definen versiones generalizadas de entropía, divergencia y familias exponenciales. Como caso particular, la información mutua puede interpretarse como una entropía generalizada.

Medida Base

Volviendo a la selección de la inicial en el caso no jerárquico...

$$p(\theta) \propto B(\theta) \exp\{x_0\theta - n_0 M(\theta)\}$$

- $B(\theta) = \pi_L(\theta) \propto 1 \quad \Rightarrow \quad p(\theta) \propto \exp\{x_0\theta - n_0 M(\theta)\}$

$$\begin{aligned} E(\mu|x_{(n)}) &= \frac{\sum_i x_i + x_0}{n + n_0} \\ &= \bar{x} \qquad (\text{ si } x_0 = 0, n_0 = 0) \end{aligned}$$

(Diaconis & Ylvisaker; 1979)

También,

$$\hat{\mu}_{KL} = E(\mu|x_{(n)}) \qquad \text{(GP; 1992)}$$

Medida Base

- $B(\theta) \propto \pi_J(\theta) \quad \Rightarrow \quad p(\theta) \propto [i_\theta(\theta)]^{1/2} \exp\{x_0\theta - n_0 M(\theta)\}$

Esta es la inicial de referencia con restricciones (Bernardo & Smith; 1994)

$$\text{FEN-FVC} \quad \Rightarrow \quad \theta \smile \mu \quad \Leftrightarrow \quad i_\theta(\theta) \propto \exp\{k_1\theta - k_2 M(\theta)\} \quad \text{para algunos } k_1, k_2.$$

$$\begin{aligned} p(\theta) &\propto \exp\left\{\left(x_0 + \frac{k_1}{2}\right)\theta - \left(n_0 + \frac{k_2}{2}\right) M(\theta)\right\} \\ &= \exp\{\tilde{x}_0\theta - \tilde{n}_0 M(\theta)\} \end{aligned}$$

De esta forma, la inicial de referencia (Jeffreys) puede interpretarse como un caso límite de la familia conjugada DY a medida que $\tilde{x}_0 \to k_1/2$ y $\tilde{n}_0 \to k_2/2$

(GP & Smith; 1995, 1997)

Medida Base

- Considere una FEN-FVC, como en el caso previo, y sea $\lambda = \lambda(\theta)$ una reparametrización
- Suponga que: (i) $\lambda \smile \theta$, y (ii) existe un estimador $\hat{\lambda} = \hat{\lambda}(x_{(n)})$ tal que $E(\hat{\lambda}|\theta) = \lambda(\theta)$

Medida Base

- Considere una FEN-FVC, como en el caso previo, y sea $\lambda = \lambda(\theta)$ una reparametrización
- Suponga que: (i) $\lambda \smile \theta$, y (ii) existe un estimador $\hat{\lambda} = \hat{\lambda}(x_{(n)})$ tal que $E(\hat{\lambda}|\theta) = \lambda(\theta)$

Pregunta: ¿Qué medida base $B(\theta)$ produce $E(\lambda|x_{(n)}) = \hat{\lambda}(x_{(n)})$?

Medida Base

- Considere una FEN-FVC, como en el caso previo, y sea $\lambda = \lambda(\theta)$ una reparametrización
- Suponga que: (i) $\lambda \smile \theta$, y (ii) existe un estimador $\hat{\lambda} = \hat{\lambda}(x_{(n)})$ tal que $E(\hat{\lambda}|\theta) = \lambda(\theta)$

Pregunta: ¿Qué medida base $B(\theta)$ produce $E(\lambda|x_{(n)}) = \hat{\lambda}(x_{(n)})$?

La respuesta es $B(\theta) \propto i_\theta(\theta)|J_\lambda(\theta)|^{-1}$. En este caso,

$$\begin{aligned} p(\theta) &\propto \exp\{(x_0 + k_1 - r_1)\,\theta - (n_0 + k_2 - r_2)\,M(\theta)\} \\ &= \exp\{\check{x}_0\theta - \check{n}_0 M(\theta)\} \end{aligned}$$

con r_1 y r_2 determinados por $\lambda(\cdot)$

$B(\theta)$ también puede interpretarse como un caso límite de la familia conjugada DY como $\check{x}_0 \to (k_1 - r_1)$ y $\check{n}_0 \to (k_2 - r_2)$

(Hartigan, 1965; GP & Mendoza, 1999)

Medida Base

- Así, $\pi_H(\theta) \propto i_\theta(\theta)|J_\lambda(\theta)|^{-1}$ es una inicial "no informativa" para λ. En términos de λ, resulta

$$\pi_\lambda(\lambda) \propto i_\lambda(\lambda)$$

Medida Base

- Así, $\pi_H(\theta) \propto i_\theta(\theta)|J_\lambda(\theta)|^{-1}$ es una inicial "no informativa" para λ. En términos de λ, resulta

$$\pi_\lambda(\lambda) \propto i_\lambda(\lambda)$$

Observe que

$$\pi_J(\lambda) \propto [\, i_\lambda(\lambda)\,]^{1/2}$$

Medida Base

- Así, $\pi_H(\theta) \propto i_\theta(\theta)|J_\lambda(\theta)|^{-1}$ es una inicial "no informativa" para λ. En términos de λ, resulta

$$\pi_\lambda(\lambda) \propto i_\lambda(\lambda)$$

Observe que

$$\pi_J(\lambda) \propto [\, i_\lambda(\lambda)\,]^{1/2}$$

y

$$\pi_L(\lambda) \propto 1$$

Observaciones Finales

- Las familias Exponenciales y la maximización de la entropía con restricciones, están muy relacionadas. Si las restricciones se eligen apropiadamente, pueden producir conjugamiento.

- Existe una *dualidad* entre maximización de la entropía y minimización de la pérdida esperada en el caso menos favorable.

- Se pueden obtener distintas clases de iniciales "no informativas" como casos límite de las familias conjugadas DY. Un caso interesante se presenta cuando el valor esperado final del parámetro λ coincide con el correspondiente estimador *insesgado* $\hat{\lambda}$.

Observaciones Finales

- Otra *dualidad* es la que existe entre la noción clásica de *insegamiento* y la minimización de la pérdida cuadrática esperada. Los resultados sugieren que esta relación se podría extender utilizando una definición más general de insesgamiento (para funciones de pérdida arbitrarias) como la que introducen Noorbaloochi & Meeden (1983).

- Otras generalizaciones interesantes podrían examinarse a partir de la teoría desarrollada por Grünwald & Dawid.

References

- Bernardo JM & Smith AFM (1994). *Bayesian Theory*. Wiley.
- Clarke BS & Barron AR (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Jounal of Statistical Planning and Inference 41*, 37–60.
- Diaconis P & Ylvisaker D (1979). Conjugate priors for exponential exponential families. *Annals of Statistics 7*, 269–281.
- George E, Makov U & Smith AFM (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics 26*, 509-517.
- Gutiérrez-Peña E & Mendoza M (1999). A note on Bayes estimates for exponential families. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales (España) 93*, 351–356.
- Gutiérrez-Peña E & Muliere P (2004). Conjugate priors represent strog pre-experimental assumptions. *Scandinavian Journal of Statistics 31*, 235–246.
- Gutiérrez-Peña E & Smith AFM (1995). Conjugate parametrizations for natural exponential families. *Journal of the American Statistical Association 90*, 1347–1356.
- Gutiérrez-Peña E & Smith AFM (1997). Exponential and Bayesian conjugate families: Review and Extensions. *Test 6*, 1–90.
- Grünwald, PD '& Dawid, AP (2004). Game theory, maximum extropy, minumum discrepancy and robust Bayesian decision theory. *Annals of Statistics 32*, 1367–1433.
- Hartigan J (1965). The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics 36*, 1137–1152.
- Kullback S (1959). *Information Theory and Statistics*. Wiley.
- Noorbaloochi S & Meeden G (1983). Unbiasedness as the dual of being Bayes. *Journal of the American Statistical Association 78*, 619–623.