El método de máxima verosimilitud y la divergencia de Kullback-Leibler

Divergencia de Kullback-Leibler

Si f y g son FDP's denotamos por

$$D_{\mathrm{KL}}(f, g) := \int f(x) \log \frac{f(x)}{g(x)} \, dx$$

a la divergencia de Kullback-Leibler entre f y g.

Divergencia de Kullback-Leibler

Si f y g son FDP's denotamos por

$$D_{\mathrm{KL}}(f, g) \coloneqq \int f(x) \log \frac{f(x)}{g(x)} \, dx$$

a la divergencia de Kullback-Leibler entre f y g. Recordemos que

- $D_{\mathrm{KL}}(f, g) \geq 0$ para cualesquiera f, g, y que
- $D_{\mathrm{KL}}(f, f) = 0$.

Máxima verosimilitud y máximo a posteriori

Sea $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f (con f no necesariamente en $\mathcal{F}$).

Sea $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f (con f no necesariamente en $\mathcal{F}$). Habiendo observado $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, denotamos por

$$L_n(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

a la función de verosimilitud.

Máxima verosimilitud y máximo a posteriori

Sea $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f (con f no necesariamente en $\mathcal{F}$). Habiendo observado $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, denotamos por

$$L_n(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

a la función de verosimilitud. Definimos el *estimador de máxima verosimilitud* como

$$\hat{\theta}_n^{\mathrm{EMV}} := \arg\max_{\theta \in \Theta} \{L_n(\theta)\}.$$

Máxima verosimilitud y máximo a posteriori

Sea $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f (con f no necesariamente en $\mathcal{F}$). Habiendo observado $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, denotamos por

$$L_n(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

a la función de verosimilitud. Definimos el *estimador de máxima verosimilitud* como

$$\hat{\theta}_n^{\mathrm{EMV}} \coloneqq \arg\max_{\theta \in \Theta} \{L_n(\theta)\}.$$

En el contexto bayesiano suponemos una distribución a priori $p(\theta)$. Definimos el *estimador máximo a posteriori* como

$$\hat{\theta}_n^{\mathrm{MAP}} \coloneqq \arg\max_{\theta \in \Theta} \{p(\theta | x_1, \ldots, x_n)\} = \arg\max_{\theta \in \Theta} \{p(x_1, \ldots, x_n | \theta) p(\theta)\}.$$

Sea $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f (con f no necesariamente en $\mathcal{F}$). Habiendo observado $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, denotamos por

$$L_n(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

a la función de verosimilitud. Definimos el *estimador de máxima verosimilitud* como

$$\hat{\theta}_n^{\mathrm{EMV}} := \arg\max_{\theta \in \Theta}\{L_n(\theta)\}.$$

En el contexto bayesiano suponemos una distribución a priori $p(\theta)$. Definimos el *estimador máximo a posteriori* como

$$\hat{\theta}_n^{\mathrm{MAP}} := \arg\max_{\theta \in \Theta}\{p(\theta|x_1, \ldots, x_n)\} = \arg\max_{\theta \in \Theta}\{p(x_1, \ldots, x_n|\theta)p(\theta)\}.$$

En particular, si θ se distribuye uniformemente (i.e. $p(\theta)$ es constante) y $p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta)$, entonces $\hat{\theta}_n^{\mathrm{MAP}} = \hat{\theta}_n^{\mathrm{EMV}}$.

Por otro lado, denotemos

$$\ell_n(\theta) \coloneqq \log L_n(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Por otro lado, denotemos

$$\ell_n(\theta) \coloneqq \log L_n(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Sabemos que maximizar $L_n(\theta)$ es equivalente a maximizar $\ell_n(\theta)$. Más aun, maximizar $\ell_n(\theta)$ es equivalente a maximizar

$$M_n(\theta) = M_n(\theta; x_1, \ldots, x_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}$$

pues $M_n(\theta) = \frac{1}{n}(\ell_n(\theta) - c)$ donde c no depende de θ.

Por otro lado, denotemos

$$\ell_n(\theta) \coloneqq \log L_n(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Sabemos que maximizar $L_n(\theta)$ es equivalente a maximizar $\ell_n(\theta)$. Más aun, maximizar $\ell_n(\theta)$ es equivalente a maximizar

$$M_n(\theta) = M_n(\theta; x_1, \ldots, x_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}$$

pues $M_n(\theta) = \frac{1}{n}(\ell_n(\theta) - c)$ donde c no depende de θ. Por la LGN,

$$M_n(\theta) \xrightarrow{P} \mathbb{E}_f \left(\log \frac{f(X;\theta)}{f(X)} \right) = \int f(x) \log \frac{f(x;\theta)}{f(x)}\, dx = -D_{\mathrm{KL}}(f, f_\theta). \quad (1)$$

Por otro lado, denotemos

$$\ell_n(\theta) \coloneqq \log L_n(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Sabemos que maximizar $L_n(\theta)$ es equivalente a maximizar $\ell_n(\theta)$. Más aun, maximizar $\ell_n(\theta)$ es equivalente a maximizar

$$M_n(\theta) = M_n(\theta; x_1, \ldots, x_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}$$

pues $M_n(\theta) = \frac{1}{n}(\ell_n(\theta) - c)$ donde c no depende de θ. Por la LGN,

$$M_n(\theta) \xrightarrow{P} \mathbb{E}_f \left(\log \frac{f(X;\theta)}{f(X)} \right) = \int f(x) \log \frac{f(x;\theta)}{f(x)}\, dx = -D_{\mathrm{KL}}(f, f_\theta). \quad (1)$$

Intuitivamente, esperaríamos que de esto se siguiera que

$$\underset{\theta\in\Theta}{\arg\max}\{M_n(\theta)\} \xrightarrow{P} \underset{\theta\in\Theta}{\arg\max}\{-D_{\mathrm{KL}}(f, f_\theta)\}.$$

Por otro lado, denotemos

$$\ell_n(\theta) \coloneqq \log L_n(\theta) = \log \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

Sabemos que maximizar $L_n(\theta)$ es equivalente a maximizar $\ell_n(\theta)$. Más aun, maximizar $\ell_n(\theta)$ es equivalente a maximizar

$$M_n(\theta) = M_n(\theta; x_1, \ldots, x_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}$$

pues $M_n(\theta) = \frac{1}{n}(\ell_n(\theta) - c)$ donde c no depende de θ. Por la LGN,

$$M_n(\theta) \xrightarrow{P} \mathbb{E}_f \left(\log \frac{f(X;\theta)}{f(X)} \right) = \int f(x) \log \frac{f(x;\theta)}{f(x)} \, dx = -D_{\mathrm{KL}}(f, f_\theta). \quad (1)$$

Intuitivamente, esperaríamos que de esto se siguiera que

$$\arg\max_{\theta \in \Theta} \{M_n(\theta)\} \xrightarrow{P} \arg\max_{\theta \in \Theta} \{-D_{\mathrm{KL}}(f, f_\theta)\}.$$

O equivalentemente, que

$$\hat{\theta}_n^{\mathrm{EMV}} \xrightarrow{P} \arg\min_{\theta \in \Theta} \{D_{\mathrm{KL}}(f, f_\theta)\}.$$

Sin embargo, en general esto no es cierto. Por eso pediremos dos cosas:

Sin embargo, en general esto no es cierto. Por eso pediremos dos cosas:

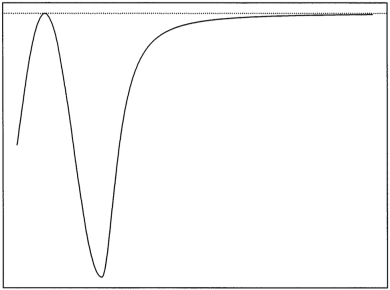
1. Pediremos que $M_n(\theta)$ converga uniformemente en probabilidad a $-D_{\mathrm{KL}}(f, f_\theta)$. (En contraste con $M_n(\theta) \xrightarrow{P} -D_{\mathrm{KL}}(f, f_\theta)$).

Sin embargo, en general esto no es cierto. Por eso pediremos dos cosas:

1. Pediremos que $M_n(\theta)$ converga uniformemente en probabilidad a $-D_{\mathrm{KL}}(f, f_\theta)$. (En contraste con $M_n(\theta) \xrightarrow{P} -D_{\mathrm{KL}}(f, f_\theta)$).
2. Pediremos que la función $\theta \mapsto D_{\mathrm{KL}}(f, f_\theta)$ se maximice en un único punto $\theta_\star$ y más aun, pediremos que solo parámetros cercanos $\theta_\star$ obtengan valores cercanos a $M(\theta_\star)$.

Sin embargo, en general esto no es cierto. Por eso pediremos dos cosas:

1. Pediremos que $M_n(\theta)$ converga uniformemente en probabilidad a $-D_{\mathrm{KL}}(f, f_\theta)$. (En contraste con $M_n(\theta) \xrightarrow{P} -D_{\mathrm{KL}}(f, f_\theta)$).
2. Pediremos que la función $\theta \mapsto D_{\mathrm{KL}}(f, f_\theta)$ se maximice en un único punto $\theta_\star$ y más aun, pediremos que solo parámetros cercanos $\theta_\star$ obtengan valores cercanos a $M(\theta_\star)$.



Teorema

Sea $\mathcal{F} = \{f(\,\cdot\,;\theta) \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f.

[1] En el caso que $f \in \mathcal{F}$, esta hipótesis es consecuencia de otra hipótesis conocida como *identificabilidad* del modelo.

Teorema

Sea $\mathcal{F} = \{f(\,\cdot\,;\theta) \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f. Denotemos

$$M_n(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}, \qquad M(\theta) = -D_{\mathrm{KL}}(f, f_\theta).$$

[1]En el caso que $f \in \mathcal{F}$, esta hipótesis es consecuencia de otra hipótesis conocida como *identificabilidad* del modelo.

Maximizar $L_n(\theta)$ es asintóticamente equivalente a minimizar D_{KL}

Teorema

Sea $\mathcal{F} = \{f(\,\cdot\,;\theta) \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f. Denotemos

$$M_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \log\frac{f(x_i;\theta)}{f(x_i)}, \qquad M(\theta) = -D_{\mathrm{KL}}(f, f_\theta).$$

Si $\theta_\star \coloneqq \arg\min_{\theta\in\Theta}\{D_{\mathrm{KL}}(f, f_\theta)\}$ esta bien definido[1],

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0, \tag{H1}$$

$$\forall\epsilon > 0, \quad \sup_{\theta : |\theta-\theta_\star| \geq \epsilon} M(\theta) < M(\theta_\star), \tag{H2}$$

[1] En el caso que $f \in \mathcal{F}$, esta hipótesis es consecuencia de otra hipótesis conocida como *identificabilidad* del modelo.

Maximizar $L_n(\theta)$ es asintóticamente equivalente a minimizar D_{KL}

Teorema

Sea $\mathcal{F} = \{f(\,\cdot\,;\theta) \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \ldots, X_n$ una muestra aleatoria de f. Denotemos

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}, \qquad M(\theta) = -D_{\mathrm{KL}}(f, f_\theta).$$

Si $\theta_\star := \arg\min_{\theta\in\Theta}\{D_{\mathrm{KL}}(f, f_\theta)\}$ esta bien definido[1],

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0, \tag{H1}$$

$$\forall \epsilon > 0, \quad \sup_{\theta : |\theta-\theta_\star| \geq \epsilon} M(\theta) < M(\theta_\star), \tag{H2}$$

entonces

$$\hat{\theta}_n^{\mathrm{EMV}} \xrightarrow{P} \theta_\star = \arg\min_{\theta\in\Theta}\{D_{\mathrm{KL}}(f, f_\theta)\}.$$

[1] En el caso que $f \in \mathcal{F}$, esta hipótesis es consecuencia de otra hipótesis conocida como *identificabilidad* del modelo.

Teorema

Sea $\mathcal{F} = \{f(\,\cdot\,;\theta) \mid \theta \in \Theta\}$ un modelo paramétrico y sea $X_1, \dots, X_n$ una muestra aleatoria de f. Denotemos

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i;\theta)}{f(x_i)}, \qquad M(\theta) = -D_{\mathrm{KL}}(f, f_\theta).$$

Si $\theta_\star := \arg\min_{\theta\in\Theta}\{D_{\mathrm{KL}}(f, f_\theta)\}$ esta bien definido[1],

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0, \tag{H1}$$

$$\forall \epsilon > 0, \quad \sup_{\theta : |\theta-\theta_\star| \geq \epsilon} M(\theta) < M(\theta_\star), \tag{H2}$$

entonces

$$\hat{\theta}_n^{\mathrm{EMV}} \xrightarrow{P} \theta_\star = \arg\min_{\theta\in\Theta}\{D_{\mathrm{KL}}(f, f_\theta)\}.$$

En particular, si $f = f_{\theta_0}$ para algun $\theta_0 \in \Theta$, entonces $\hat{\theta}_n^{\mathrm{EMV}} \xrightarrow{P} \theta_0$.

[1] En el caso que $f \in \mathcal{F}$, esta hipótesis es consecuencia de otra hipótesis conocida como *identificabilidad* del modelo.

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{EMV}}$.

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{EMV}}$. Notemos que

$$\begin{aligned}
M(\theta_\star) - M(\hat{\theta}_n) &= M(\theta_\star) - M_n(\theta_\star) + M_n(\theta_\star) - M(\hat{\theta}_n) \\
&\leq \sup_{\theta \in \Theta} |M(\theta) - M_n(\theta)| + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\
&\qquad\qquad\qquad\qquad\qquad\qquad (\hat{\theta}_n \text{ maximiza } M_n) \\
&\xrightarrow{P} 0. \qquad\qquad\qquad\qquad\qquad (\text{por } (1) \text{ y } (\mathrm{H1}))
\end{aligned}$$

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\text{EMV}}$. Notemos que

$$\begin{aligned} M(\theta_\star) - M(\hat{\theta}_n) &= M(\theta_\star) - M_n(\theta_\star) + M_n(\theta_\star) - M(\hat{\theta}_n) \\ &\leq \sup_{\theta \in \Theta} |M(\theta) - M_n(\theta)| + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\ & \qquad\qquad\qquad\qquad\qquad\qquad (\hat{\theta}_n \text{ maximiza } M_n) \\ &\xrightarrow{P} 0. \qquad\qquad\qquad\qquad\qquad (\text{por (1) y (H1)}) \end{aligned}$$

En particular, para cualquier $\delta > 0$

$$P\left(M(\hat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \tag{2}$$

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\text{EMV}}$. Notemos que

$$\begin{aligned}
M(\theta_\star) - M(\hat{\theta}_n) &= M(\theta_\star) - M_n(\theta_\star) + M_n(\theta_\star) - M(\hat{\theta}_n) \\
&\leq \sup_{\theta \in \Theta} |M(\theta) - M_n(\theta)| + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\
& \qquad\qquad\qquad\qquad (\hat{\theta}_n \text{ maximiza } M_n) \\
&\xrightarrow{P} 0. \qquad\qquad\qquad (\text{por (1) y (H1)})
\end{aligned}$$

En particular, para cualquier $\delta > 0$

$$P\left(M(\hat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \tag{2}$$

Por otro lado, sea $\epsilon > 0$ arbitrario y sea

$$\delta = M(\theta_\star) - \sup_{\theta : |\theta - \theta_\star| \geq \epsilon} M(\theta) > 0. \qquad (\text{por (H2)})$$

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{EMV}}$. Notemos que

$$
\begin{aligned}
M(\theta_\star) - M(\hat{\theta}_n) &= M(\theta_\star) - M_n(\theta_\star) + M_n(\theta_\star) - M(\hat{\theta}_n) \\
&\leq \sup_{\theta\in\Theta} |M(\theta) - M_n(\theta)| + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\
&\qquad\qquad\qquad\qquad\qquad\qquad (\hat{\theta}_n \text{ maximiza } M_n) \\
&\xrightarrow{P} 0. \qquad\qquad\qquad\qquad \text{(por (1) y (H1))}
\end{aligned}
$$

En particular, para cualquier $\delta > 0$

$$
P\left(M(\hat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \tag{2}
$$

Por otro lado, sea $\epsilon > 0$ arbitrario y sea

$$
\delta = M(\theta_\star) - \sup_{\theta:|\theta-\theta_\star|\geq\epsilon} M(\theta) > 0. \qquad \text{(por (H2))}
$$

De esta manera, $|\theta - \theta_\star| \geq \epsilon$ implica $M(\theta) < M(\theta_\star) - \delta$.

Demostración. Denotemos $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{EMV}}$. Notemos que

$$\begin{aligned} M(\theta_\star) - M(\hat{\theta}_n) &= M(\theta_\star) - M_n(\theta_\star) + M_n(\theta_\star) - M(\hat{\theta}_n) \\ &\leq \sup_{\theta \in \Theta} |M(\theta) - M_n(\theta)| + M_n(\hat{\theta}_n) - M(\hat{\theta}_n) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad (\hat{\theta}_n \text{ maximiza } M_n) \\ &\xrightarrow{P} 0. \qquad\qquad\qquad\qquad\qquad\qquad (\text{por (1) y (H1)}) \end{aligned}$$

En particular, para cualquier $\delta > 0$

$$P\left(M(\hat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \tag{2}$$

Por otro lado, sea $\epsilon > 0$ arbitrario y sea

$$\delta = M(\theta_\star) - \sup_{\theta : |\theta - \theta_\star| \geq \epsilon} M(\theta) > 0. \qquad (\text{por (H2)})$$

De esta manera, $|\theta - \theta_\star| \geq \epsilon$ implica $M(\theta) < M(\theta_\star) - \delta$. Por lo tanto,

$$P\left(|\hat{\theta}_n - \theta_\star| > \epsilon\right) \leq P\left(M(\hat{\theta}_n) < M(\theta_\star) - \delta\right) \xrightarrow{P} 0.$$

□