

Tarea 2

Fecha de entrega: 25 de octubre de 2010

(El trabajo se puede elaborar en equipo de hasta 2 personas)

- A) Genere una muestra aleatoria de $n = 100$ datos, $\mathbf{Y}^{(100)} := \{Y_i\}_{i=1}^{100}$, de la siguiente manera $\{Y_i\}_{i=1}^{33} \stackrel{\text{iid}}{\sim} N(-8, 1)$, $\{Y_i\}_{i=34}^{77} \stackrel{\text{iid}}{\sim} N(0, 1)$ y $\{Y_i\}_{i=78}^{100} \stackrel{\text{iid}}{\sim} N(8, 1)$

Modela los datos $\mathbf{Y}^{(100)}$ mediante un *modelo de mezclas Bayesiano no-paramétrico* dado de la siguiente forma

$$\begin{aligned} Y_i | X_i &\stackrel{\text{ind}}{\sim} f(\cdot | X_i) \\ X_i | \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P} \\ \tilde{P} &\sim \mathcal{P} \end{aligned} \quad (1)$$

1. Sean $f(Y_i | X_i) = N(Y_i; X_i, 1)$, $\mathcal{P} = \mathcal{D}ir(a P_0)$ y $P_0 = N(\bar{Y}, t^3)$ con t el rango de los datos $\mathbf{Y}^{(100)}$.

a) Demostrar que, bajo estos supuestos, se tiene

$$P\left(X_i \in \cdot \mid X_{-i}^{(n)}, \theta, Y^{(n)}\right) = q_{i,0}^* N\left(X_i \in \cdot \mid \frac{t^3 Y_i + \bar{X}}{1 + t^3}, \frac{t^3}{t^3 + 1}\right) + \sum_{j=1}^{k_i} q_{i,j}^* \delta_{X_j^*}(\cdot), \quad (2)$$

donde k_i indica el número de valores distintos, X_j^* 's, en $X_{-i}^{(n)}$ y

$$q_{i,0}^* \propto \frac{a}{a + n - 1} N(\cdot | \bar{Y}, t^3 + 1) \quad \text{and} \quad q_{i,j}^* \propto \left\{ \frac{n_j}{a + n - 1} \right\} N(\cdot | X_j^*, 1),$$

- b) Utilizando el esquema de urnas de Pólya y el sistema de actualización obtenido en el inciso anterior elabora un programa de computo (o utiliza el dado en clase!) para dar un estimador de f mediante

$$\hat{f}(y) = \frac{1}{N} \sum_{t=1}^N \int_{\mathbb{X}} f(y|x) E\left[\tilde{P}(dx) \mid X_{1,t}^*, \dots, X_{k_t,t}^*\right]$$

Asociado a una muestra de tamaño 100 de P , donde $P \stackrel{\text{iid}}{\sim} \mathcal{D}ir(a P_0)$ (c.s. discreta), existe una distribución sobre K_n (el número de valores distintos en una muestra de tamaño n) dada por

$$Pr(K_n = k) = \frac{a^k}{(a)_n} |s(n, k)|, \quad k = 1, \dots, n$$

donde $s_{n,k}$ denotan los números de Stirling del primer tipo. Con $a = 0.2, 0.5, 1, 3$ esta distribución tiene moda en $k = 2, 3, 5, 11$ respectivamente.

Así pues, para los valores de $a = 0.2, 0.5, 1, 3$ obtén los estimadores para f y para la distribución posterior de $K_n \mid \mathbf{Y}^{(n)}$ dada por

$$P\left[K_n = k \mid \mathbf{Y}^{(n)}\right] \approx \frac{1}{N} \sum_{t=1}^N I\{k^{(t)} = k\}$$

Exhibe los resultados de dichas simulaciones para los siguientes escenarios: (i) $t = 5, t = 10, t = 100, t = 1000, t = 5000$ y burn = 2000 + $t = 10000$ iteraciones del muestreo por Gibbs resultante del sistema de urnas arriba mencionado (t denota el número de iteraciones y “burn” el periodo de calentamiento). Los resultados de las estimaciones de f los deberás de reportar en gráficas y los valores de la distribución final (posterior) de K_n en tablas.

Para el caso burn = 2000 + $t = 10000$ con $a = 1$ repite el ejercicio de simulación con varianzas 100, 500, 1000, 5000, 20000 en vez de t^3 .

Menciona que observas de este ejercicio, e.g. precisión del estimador de f , del número de grupos en los datos, que tan “rápido” (núm. de iteraciones) converge a los valores reales (y como depende esta velocidad de la elección de a / γ), que tanto depende de la dispersión en P_0 (del soporte!), etc.

- c) Repite el ejercicio anterior cuando $\mathcal{P}$ resulta de normalizar un proceso estable (denotado por $\mathcal{E}(\gamma, P_0)$), donde γ es el parámetro de dicha medida de probabilidad aleatoria y $\gamma \in (0, 1)$.

En este caso

$$Pr(K_n = k) = \frac{\Gamma(k)}{\Gamma(n)} \frac{\mathcal{E}(n, k; \gamma)}{\gamma}, \quad k = 1, \dots, n$$

donde

$$\mathcal{E}(n, k; \gamma) := \frac{1}{k!} \sum_{j=0}^k (-1)^j \binom{k}{j} (-j\gamma)_n$$

denota el coeficiente factorial generalizado.

Esta distribución es mucho menos leptocúrtica que la correspondiente al proceso de Dirichlet. En vez de usar los valores de a del inciso anterior utiliza $\gamma = 0.1, 0.3, 0.5, 0.9$.

B) Ejercicio opcional (+1 punto calificación final).

Repita el ejercicio en (A) pero con $\{Y_i\}_{i=1}^{50} \stackrel{\text{iid}}{\sim} \text{Exp}(0.5)$ y $\{Y_i\}_{i=51}^{100} \stackrel{\text{iid}}{\sim} \text{Exp}(2)$, donde $\text{Exp}(\lambda)$ denota una distribución exponencial con densidad $\text{Exp}(x, \lambda) = \lambda e^{-x\lambda} I(\lambda > 0)$. Además supóngase que $f(y | x) = \text{Exp}(y; x)$ y $P_0 = \text{Exp}(\lambda_0)$. **Nota:** Esto también cambia la estructura de (2) y sus correspondientes q_i^* 's (por lo tanto el programa hecho en clase), sin embargo las distribuciones iniciales de K_n no cambian! y por lo tanto valores a (que corresponden a ciertas modas) tampoco!.

Considérese que (dependiendo del programa/computadora) los resultados pueden tomar algo de tiempo.

SUERTE!