Inferencia estadística desde una perspectiva Bayesiana no paramétrica

E. Gutiérrez-Peña[1] S.G. Walker[2]

[1]Departamento de Probabilidad y Estadística
IIMAS, UNAM

[2]Instituto de Matemáticas, Estadística y Ciencia Actuarial
Universidad de Kent, Inglaterra

Seminario Aleatorio, ITAM - 11 de mayo de 2007

Temario

Motivación

- Teoría Bayesiana de la Decisión: se basa en una serie de postulados (*Axiomas de Coherencia*) que describen qué se entiende por comportamiento racional.
- Establece cómo deben tomarse las decisiones si uno desea evitar un comportamiento contradictorio.
- Ni esta teoría ni la Estadística Bayesiana están exentas de críticas, pero para mucha gente constituyen el enfoque más convincente para hacer inferencia estadística.
- Concepto fundamental: coherencia – Se utiliza con frecuencia como un argumento contra otros enfoques de la estadística tales como el frecuentista o el fiducial.

Motivación

- Teoría Bayesiana de la Decisión: se basa en una serie de postulados (*Axiomas de Coherencia*) que describen qué se entiende por comportamiento racional.
- Establece cómo deben tomarse las decisiones si uno desea evitar un comportamiento contradictorio.
- Ni esta teoría ni la Estadística Bayesiana están exentas de críticas, pero para mucha gente constituyen el enfoque más convincente para hacer inferencia estadística.
- Concepto fundamental: coherencia – Se utiliza con frecuencia como un argumento contra otros enfoques de la estadística tales como el frecuentista o el fiducial.

Motivación

- Teoría Bayesiana de la Decisión: se basa en una serie de postulados (*Axiomas de Coherencia*) que describen qué se entiende por comportamiento racional.
- Establece cómo deben tomarse las decisiones si uno desea evitar un comportamiento contradictorio.
- Ni esta teoría ni la Estadística Bayesiana están exentas de críticas, pero para mucha gente constituyen el enfoque más convincente para hacer inferencia estadística.
- Concepto fundamental: coherencia – Se utiliza con frecuencia como un argumento contra otros enfoques de la estadística tales como el frecuentista o el fiducial.

Motivación

- Teoría Bayesiana de la Decisión: se basa en una serie de postulados (*Axiomas de Coherencia*) que describen qué se entiende por comportamiento racional.
- Establece cómo deben tomarse las decisiones si uno desea evitar un comportamiento contradictorio.
- Ni esta teoría ni la Estadística Bayesiana están exentas de críticas, pero para mucha gente constituyen el enfoque más convincente para hacer inferencia estadística.
- Concepto fundamental: coherencia – Se utiliza con frecuencia como un argumento contra otros enfoques de la estadística tales como el frecuentista o el fiducial.

Motivación

- Aspectos principales de la Teoría Bayesiana de la Decisión (e.g. Bernardo & Smith, 1994; Hirshleifer & Riley, 1992):

  Dado un conjunto de decisiones o acciones potenciales:

  1. Toda la incertidumbre debe ser descrita en términos de una medida de probabilidad (subjetiva).
  2. Todas las preferencias deben cuantificarse en términos de una función de utilidad.
  3. La mejor decisión es aquella que maximiza la utilidad esperada (respecto a la medida de probabilidad del Punto 1).

Motivación

- Aspectos principales de la Teoría Bayesiana de la Decisión (e.g. Bernardo & Smith, 1994; Hirshleifer & Riley, 1992):

  Dado un conjunto de decisiones o acciones potenciales:

  1. Toda la incertidumbre debe ser descrita en términos de una medida de probabilidad (subjetiva).
  2. Todas las preferencias deben cuantificarse en términos de una función de utilidad.
  3. La mejor decisión es aquella que maximiza la utilidad esperada (respecto a la medida de probabilidad del Punto 1).

Motivación

- Aspectos principales de la Teoría Bayesiana de la Decisión (e.g. Bernardo & Smith, 1994; Hirshleifer & Riley, 1992):

  Dado un conjunto de decisiones o acciones potenciales:

  1. Toda la incertidumbre debe ser descrita en términos de una medida de probabilidad (subjetiva).
  2. Todas las preferencias deben cuantificarse en términos de una función de utilidad.
  3. La mejor decisión es aquella que maximiza la utilidad esperada (respecto a la medida de probabilidad del Punto 1).

Motivación

- Aspectos principales de la Teoría Bayesiana de la Decisión (e.g. Bernardo & Smith, 1994; Hirshleifer & Riley, 1992):

  Dado un conjunto de decisiones o acciones potenciales:

  1. Toda la incertidumbre debe ser descrita en términos de una medida de probabilidad (subjetiva).
  2. Todas las preferencias deben cuantificarse en términos de una función de utilidad.
  3. La mejor decisión es aquella que maximiza la utilidad esperada (respecto a la medida de probabilidad del Punto 1).

Motivación

- Los Axioma de Coherencia implican que la probabilidad debe de interpretarse como el grado de creencia sobre la ocurrencia de sucesos inciertos que pueden afectar nuestras decisiones.
- Tales creencias dependen del estado de información disponible al momento de tomar la decisión.
- Cuando aparece nueva información (ocurrencia de un evento, observación de una muestra), el tomador de decisiones debe actualizar sus creencias acerca de los sucesos inciertos.
- Otra consecuencia de los axiomas es que esta actualización debe hacerse a través del Teorema de Bayes.
- Más aún, el Teorema de Bayes es *la única forma coherente de actualizar nuestras creencias*.

Motivación

- Los Axioma de Coherencia implican que la probabilidad debe de interpretarse como el grado de creencia sobre la ocurrencia de sucesos inciertos que pueden afectar nuestras decisiones.
- Tales creencias dependen del estado de información disponible al momento de tomar la decisión.
- Cuando aparece nueva información (ocurrencia de un evento, observación de una muestra), el tomador de decisiones debe actualizar sus creencias acerca de los sucesos inciertos.
- Otra consecuencia de los axiomas es que esta actualización debe hacerse a través del Teorema de Bayes.
- Más aún, el Teorema de Bayes es *la única forma coherente de actualizar nuestras creencias*.

Motivación

- Los Axioma de Coherencia implican que la probabilidad debe de interpretarse como el grado de creencia sobre la ocurrencia de sucesos inciertos que pueden afectar nuestras decisiones.
- Tales creencias dependen del estado de información disponible al momento de tomar la decisión.
- Cuando aparece nueva información (ocurrencia de un evento, observación de una muestra), el tomador de decisiones debe actualizar sus creencias acerca de los sucesos inciertos.
- Otra consecuencia de los axiomas es que esta actualización debe hacerse a través del Teorema de Bayes.
- Más aún, el Teorema de Bayes es *la única forma coherente de actualizar nuestras creencias*.

Motivación

- Los Axioma de Coherencia implican que la probabilidad debe de interpretarse como el grado de creencia sobre la ocurrencia de sucesos inciertos que pueden afectar nuestras decisiones.
- Tales creencias dependen del estado de información disponible al momento de tomar la decisión.
- Cuando aparece nueva información (ocurrencia de un evento, observación de una muestra), el tomador de decisiones debe actualizar sus creencias acerca de los sucesos inciertos.
- Otra consecuencia de los axiomas es que esta actualización debe hacerse a través del Teorema de Bayes.
- Más aún, el Teorema de Bayes es *la única forma coherente de actualizar nuestras creencias*.

Motivación

- Los Axioma de Coherencia implican que la probabilidad debe de interpretarse como el grado de creencia sobre la ocurrencia de sucesos inciertos que pueden afectar nuestras decisiones.
- Tales creencias dependen del estado de información disponible al momento de tomar la decisión.
- Cuando aparece nueva información (ocurrencia de un evento, observación de una muestra), el tomador de decisiones debe actualizar sus creencias acerca de los sucesos inciertos.
- Otra consecuencia de los axiomas es que esta actualización debe hacerse a través del Teorema de Bayes.
- Más aún, el Teorema de Bayes es *la única forma coherente de actualizar nuestras creencias*.

Motivación

- Las aplicaciones de la Teoría Bayesiana de la Decisión a problemas de inferencia estadística se han concentrado casi exclusivamente en problemas paramétricos.
- Problemas tradicionales como estimación puntual y contraste de hipótesis se han resuelto de manera coherente usando esta teoría.
- Pero en lo que se refiere al problema de selección de modelos, la maquinaria Bayesiana ha mostrado ser incoherente.
- Box (1980) considera a la modelación estadística como un proceso dinámico y sugiere que tal vez uno debería ser Bayesiano al hacer inferencias sobre un modelo (paramétrico) dado... ¡y frecuentista al determinar su ajuste y el de otros modelos alternativos!

Motivación

- Las aplicaciones de la Teoría Bayesiana de la Decisión a problemas de inferencia estadística se han concentrado casi exclusivamente en problemas paramétricos.
- Problemas tradicionales como estimación puntual y contraste de hipótesis se han resuelto de manera coherente usando esta teoría.
- Pero en lo que se refiere al problema de selección de modelos, la maquinaria Bayesiana ha mostrado ser incoherente.
- Box (1980) considera a la modelación estadística como un proceso dinámico y sugiere que tal vez uno debería ser Bayesiano al hacer inferencias sobre un modelo (paramétrico) dado... ¡y frecuentista al determinar su ajuste y el de otros modelos alternativos!

Motivación

- Las aplicaciones de la Teoría Bayesiana de la Decisión a problemas de inferencia estadística se han concentrado casi exclusivamente en problemas paramétricos.
- Problemas tradicionales como estimación puntual y contraste de hipótesis se han resuelto de manera coherente usando esta teoría.
- Pero en lo que se refiere al problema de selección de modelos, la maquinaria Bayesiana ha mostrado ser incoherente.
- Box (1980) considera a la modelación estadística como un proceso dinámico y sugiere que tal vez uno debería ser Bayesiano al hacer inferencias sobre un modelo (paramétrico) dado... ¡y frecuentista al determinar su ajuste y el de otros modelos alternativos!

Motivación

- Las aplicaciones de la Teoría Bayesiana de la Decisión a problemas de inferencia estadística se han concentrado casi exclusivamente en problemas paramétricos.
- Problemas tradicionales como estimación puntual y contraste de hipótesis se han resuelto de manera coherente usando esta teoría.
- Pero en lo que se refiere al problema de selección de modelos, la maquinaria Bayesiana ha mostrado ser incoherente.
- Box (1980) considera a la modelación estadística como un proceso dinámico y sugiere que tal vez uno debería ser Bayesiano al hacer inferencias sobre un modelo (paramétrico) dado... ¡y frecuentista al determinar su ajuste y el de otros modelos alternativos!

Motivación - ¿Bayesianos incoherentes?

- Consideremos el problema de asignar una distribución inicial sobre el conjunto de funciones de densidad

$$\Omega = \{f_1, f_2, \ldots\}.$$

- El Estadístico A considera que Ω contiene a la densidad verdadera y por lo tanto asigna una distribución inicial $\Pi(\cdot)$ tal que $\Pi(\Omega) = 1$.
- El Estadístico B no está convencido de construir una medida de probabilidad sobre todo Ω y, para alguna M finita, asigna toda la masa de probabilidad a $\Omega_M = \{f_1, \ldots, f_M\}$, de manera que $\Pi_M(\Omega_M) = 1$.
- Supongamos ahora que, al observar los datos, es evidente que la asignación de probabilidad 1 a Ω_M no era apropiada...

Motivación - ¿Bayesianos incoherentes?

- Consideremos el problema de asignar una distribución inicial sobre el conjunto de funciones de densidad

$$\Omega = \{f_1, f_2, \ldots\}.$$

- El Estadístico A considera que Ω contiene a la densidad verdadera y por lo tanto asigna una distribución inicial $\Pi(\cdot)$ tal que $\Pi(\Omega) = 1$.
- El Estadístico B no está convencido de construir una medida de probabilidad sobre todo Ω y, para alguna M finita, asigna toda la masa de probabilidad a $\Omega_M = \{f_1, \ldots, f_M\}$, de manera que $\Pi_M(\Omega_M) = 1$.
- Supongamos ahora que, al observar los datos, es evidente que la asignación de probabilidad 1 a Ω_M no era apropiada...

Motivación - ¿Bayesianos incoherentes?

- Consideremos el problema de asignar una distribución inicial sobre el conjunto de funciones de densidad

$$\Omega = \{f_1, f_2, \ldots\}.$$

- El Estadístico A considera que Ω contiene a la densidad verdadera y por lo tanto asigna una distribución inicial $\Pi(\cdot)$ tal que $\Pi(\Omega) = 1$.
- El Estadístico B no está convencido de construir una medida de probabilidad sobre todo Ω y, para alguna M finita, asigna toda la masa de probabilidad a $\Omega_M = \{f_1, \ldots, f_M\}$, de manera que $\Pi_M(\Omega_M) = 1$.
- Supongamos ahora que, al observar los datos, es evidente que la asignación de probabilidad 1 a Ω_M no era apropiada...

Motivación - ¿Bayesianos incoherentes?

- Consideremos el problema de asignar una distribución inicial sobre el conjunto de funciones de densidad

$$\Omega = \{f_1, f_2, \ldots\}.$$

- El Estadístico A considera que Ω contiene a la densidad verdadera y por lo tanto asigna una distribución inicial $\Pi(\cdot)$ tal que $\Pi(\Omega) = 1$.
- El Estadístico B no está convencido de construir una medida de probabilidad sobre todo Ω y, para alguna M finita, asigna toda la masa de probabilidad a $\Omega_M = \{f_1, \ldots, f_M\}$, de manera que $\Pi_M(\Omega_M) = 1$.
- Supongamos ahora que, al observar los datos, es evidente que la asignación de probabilidad 1 a Ω_M no era apropiada...

Motivación - ¿Bayesianos incoherentes?

- El Estadístico A simplemente actualiza su inicial vía el Teorema de Bayes para obtener la distribución final $\Pi(\cdot|\mathbf{D})$.
- El Estadístico B no puede hacer lo mismo, así que cambia su distribución inicial sobre Ω, digamos a $\Pi_{M^*}(\cdot)$, y entonces... actualiza esta nueva inicial vía el Teorema de Bayes para obtener la distribución final $\Pi_{M^*}(\cdot|\mathbf{D})$
- !Un momento! El Estadístico B ha *actualizado* su inicial $\Pi_M(\cdot)$ a la luz de los datos, obteniendo una nueva distribución inicial $\Pi_{M^*}(\cdot)$, pero no ha usado el Teorema de Bayes. Esto es una muestra de comportamiento incoherente.
- *El argumento es el mismo si ahora tomamos a Ω como la familia de todas las densidades y a Ω_M como una familia paramétrica de dimensión finita.*

Motivación - ¿Bayesianos incoherentes?

- El Estadístico A simplemente actualiza su inicial vía el Teorema de Bayes para obtener la distribución final $\Pi(\cdot|\mathbf{D})$.
- El Estadístico B no puede hacer lo mismo, así que cambia su distribución inicial sobre Ω, digamos a $\Pi_{M^*}(\cdot)$, y entonces... actualiza esta nueva inicial vía el Teorema de Bayes para obtener la distribución final $\Pi_{M^*}(\cdot|\mathbf{D})$
- !Un momento! El Estadístico B ha *actualizado* su inicial $\Pi_M(\cdot)$ a la luz de los datos, obteniendo una nueva distribución inicial $\Pi_{M^*}(\cdot)$, pero no ha usado el Teorema de Bayes. Esto es una muestra de comportamiento incoherente.
- *El argumento es el mismo si ahora tomamos a Ω como la familia de todas las densidades y a Ω_M como una familia paramétrica de dimensión finita.*

Motivación - ¿Bayesianos incoherentes?

- El Estadístico A simplemente actualiza su inicial vía el Teorema de Bayes para obtener la distribución final $\Pi(\cdot|\mathbf{D})$.
- El Estadístico B no puede hacer lo mismo, así que cambia su distribución inicial sobre Ω, digamos a $\Pi_{M^*}(\cdot)$, y entonces... actualiza esta nueva inicial vía el Teorema de Bayes para obtener la distribución final $\Pi_{M^*}(\cdot|\mathbf{D})$
- !Un momento! El Estadístico B ha *actualizado* su inicial $\Pi_M(\cdot)$ a la luz de los datos, obteniendo una nueva distribución inicial $\Pi_{M^*}(\cdot)$, pero no ha usado el Teorema de Bayes. Esto es una muestra de comportamiento incoherente.
- *El argumento es el mismo si ahora tomamos a Ω como la familia de todas las densidades y a Ω_M como una familia paramétrica de dimensión finita.*

Motivación - ¿Bayesianos incoherentes?

- El Estadístico A simplemente actualiza su inicial vía el Teorema de Bayes para obtener la distribución final $\Pi(\cdot|\mathbf{D})$.
- El Estadístico B no puede hacer lo mismo, así que cambia su distribución inicial sobre Ω, digamos a $\Pi_{M^*}(\cdot)$, y entonces... actualiza esta nueva inicial vía el Teorema de Bayes para obtener la distribución final $\Pi_{M^*}(\cdot|\mathbf{D})$
- !Un momento! El Estadístico B ha *actualizado* su inicial $\Pi_M(\cdot)$ a la luz de los datos, obteniendo una nueva distribución inicial $\Pi_{M^*}(\cdot)$, pero no ha usado el Teorema de Bayes. Esto es una muestra de comportamiento incoherente.
- *El argumento es el mismo si ahora tomamos a Ω como la familia de todas las densidades y a Ω_M como una familia paramétrica de dimensión finita.*

Motivación - ¿Bayesianos incoherentes?

- El Estadístico A simplemente actualiza su inicial vía el Teorema de Bayes para obtener la distribución final $\Pi(\cdot|\mathbf{D})$.
- El Estadístico B no puede hacer lo mismo, así que cambia su distribución inicial sobre Ω, digamos a $\Pi_{M^*}(\cdot)$, y entonces... actualiza esta nueva inicial vía el Teorema de Bayes para obtener la distribución final $\Pi_{M^*}(\cdot|\mathbf{D})$
- !Un momento! El Estadístico B ha *actualizado* su inicial $\Pi_M(\cdot)$ a la luz de los datos, obteniendo una nueva distribución inicial $\Pi_{M^*}(\cdot)$, pero no ha usado el Teorema de Bayes. Esto es una muestra de comportamiento incoherente.
- *El argumento es el mismo si ahora tomamos a Ω como la familia de todas las densidades y a Ω_M como una familia paramétrica de dimensión finita.*

Motivación - ¿Bayesianos incoherentes?

- Incluso un modelo paramétrico

$$\Omega_0 = \{f \in \Omega : f(\cdot) \equiv f(\cdot;\theta), \theta \in \Theta\},$$

con distribución inicial $\pi(\theta)$ definida sobre Θ, determina una medida de probabilidad sobre Ω

$$\Pr(f \in B) \equiv \Pi_0(B) = \int_{\{\theta : f(\cdot;\theta) \in B\}} \pi(\theta)\, d\theta$$

pero es tal que $\Pi_0(\Omega_0) = 1$.

- Si una vez que se observan los datos se descubre que la familia paramétrica es deficiente en algún sentido, la distribución inicial sobre Ω se cambia sin tomar en cuenta la incoherencia.

Motivación - ¿Bayesianos incoherentes?

- Incluso un modelo paramétrico

$$\Omega_0 = \{f \in \Omega : f(\cdot) \equiv f(\cdot;\theta), \theta \in \Theta\},$$

con distribución inicial $\pi(\theta)$ definida sobre Θ, determina una medida de probabilidad sobre Ω

$$\Pr(f \in B) \equiv \Pi_0(B) = \int_{\{\theta : f(\cdot;\theta) \in B\}} \pi(\theta)\, d\theta$$

pero es tal que $\Pi_0(\Omega_0) = 1$.

- Si una vez que se observan los datos se descubre que la familia paramétrica es deficiente en algún sentido, la distribución inicial sobre Ω se cambia sin tomar en cuenta la incoherencia.

Motivación - ¿Bayesianos incoherentes?

- Sí. Desde un punto de vista práctico es necesario actualizar la distribución inicial a la luz de los datos cuando la inicial original es claramente inadecuada. Este es el mensaje de Box (1980).
- Pero es incoherente hacerlo así. El problema radica en la impaciencia del Estadístico B por hacerse la vida más fácil... ¡recordemos la ley de Cromwell!
- Lindsey (1999) y Draper (1999) ven esto como un serio problema para la inferencia Bayesiana.
- [Una posibilidad para evitar esta incoherencia es “aferrarse” a la distribución inicial original, aunque ahora se considere absurda.]

Motivación - ¿Bayesianos incoherentes?

- Sí. Desde un punto de vista práctico es necesario actualizar la distribución inicial a la luz de los datos cuando la inicial original es claramente inadecuada. Este es el mensaje de Box (1980).
- Pero es incoherente hacerlo así. El problema radica en la impaciencia del Estadístico B por hacerse la vida más fácil... ¡recordemos la ley de Cromwell!
- Lindsey (1999) y Draper (1999) ven esto como un serio problema para la inferencia Bayesiana.
- [Una posibilidad para evitar esta incoherencia es “aferrarse” a la distribución inicial original, aunque ahora se considere absurda.]

Motivación - ¿Bayesianos incoherentes?

- Sí. Desde un punto de vista práctico es necesario actualizar la distribución inicial a la luz de los datos cuando la inicial original es claramente inadecuada. Este es el mensaje de Box (1980).
- Pero es incoherente hacerlo así. El problema radica en la impaciencia del Estadístico B por hacerse la vida más fácil... ¡recordemos la ley de Cromwell!
- Lindsey (1999) y Draper (1999) ven esto como un serio problema para la inferencia Bayesiana.
- [Una posibilidad para evitar esta incoherencia es “aferrarse” a la distribución inicial original, aunque ahora se considere absurda.]

Motivación - ¿Bayesianos incoherentes?

- Sí. Desde un punto de vista práctico es necesario actualizar la distribución inicial a la luz de los datos cuando la inicial original es claramente inadecuada. Este es el mensaje de Box (1980).
- Pero es incoherente hacerlo así. El problema radica en la impaciencia del Estadístico B por hacerse la vida más fácil... ¡recordemos la ley de Cromwell!
- Lindsey (1999) y Draper (1999) ven esto como un serio problema para la inferencia Bayesiana.
- [Una posibilidad para evitar esta incoherencia es “aferrarse” a la distribución inicial original, aunque ahora se considere absurda.]

Motivación - ¿Bayesianos incoherentes?

- Sí. Desde un punto de vista práctico es necesario actualizar la distribución inicial a la luz de los datos cuando la inicial original es claramente inadecuada. Este es el mensaje de Box (1980).
- Pero es incoherente hacerlo así. El problema radica en la impaciencia del Estadístico B por hacerse la vida más fácil... ¡recordemos la ley de Cromwell!
- Lindsey (1999) y Draper (1999) ven esto como un serio problema para la inferencia Bayesiana.
- [Una posibilidad para evitar esta incoherencia es “aferrarse” a la distribución inicial original, aunque ahora se considere absurda.]

Motivación - ¿Bayesianos incoherentes?

- Una manera obvia de darle la vuelta a este problema es asignándole probabilidad positiva a todos los subconjuntos relevantes del espacio de densidades Ω. Esto se puede lograr con una distribución inicial no paramétrica.
- Existen varias distribuciones iniciales que pueden utilizarse en este contexto, incluyendo el Proceso Dirichlet (Ferguson, 1973), Mezclas basadas en Procesos Dirichlet (Lo, 1984) y Árboles de Pólya (Lavine, 1992). Para más detalles, pueden consultarse los trabajos de Walker et al. (1999) y Dey et al. (1998).
- [Bayesiano Paramétrico Necio $<$ Bayesiano Paramétrico Incoherente]
- [Bayesiano Paramétrico Incoherente $<<$ Bayesiano No Paramétrico]

Motivación - ¿Bayesianos incoherentes?

- Una manera obvia de darle la vuelta a este problema es asignándole probabilidad positiva a todos los subconjuntos relevantes del espacio de densidades Ω. Esto se puede lograr con una distribución inicial no paramétrica.
- Existen varias distribuciones iniciales que pueden utilizarse en este contexto, incluyendo el Proceso Dirichlet (Ferguson, 1973), Mezclas basadas en Procesos Dirichlet (Lo, 1984) y Árboles de Pólya (Lavine, 1992). Para más detalles, pueden consultarse los trabajos de Walker et al. (1999) y Dey et al. (1998).
- [Bayesiano Paramétrico Necio $<$ Bayesiano Paramétrico Incoherente]
- [Bayesiano Paramétrico Incoherente $<<$ Bayesiano No Paramétrico]

Motivación - ¿Bayesianos incoherentes?

- Una manera obvia de darle la vuelta a este problema es asignándole probabilidad positiva a todos los subconjuntos relevantes del espacio de densidades Ω. Esto se puede lograr con una distribución inicial no paramétrica.
- Existen varias distribuciones iniciales que pueden utilizarse en este contexto, incluyendo el Proceso Dirichlet (Ferguson, 1973), Mezclas basadas en Procesos Dirichlet (Lo, 1984) y Árboles de Pólya (Lavine, 1992). Para más detalles, pueden consultarse los trabajos de Walker et al. (1999) y Dey et al. (1998).
- [Bayesiano Paramétrico Necio $<$ Bayesiano Paramétrico Incoherente]
- [Bayesiano Paramétrico Incoherente $<<$ Bayesiano No Paramétrico]

Motivación - ¿Bayesianos incoherentes?

- Una manera obvia de darle la vuelta a este problema es asignándole probabilidad positiva a todos los subconjuntos relevantes del espacio de densidades Ω. Esto se puede lograr con una distribución inicial no paramétrica.
- Existen varias distribuciones iniciales que pueden utilizarse en este contexto, incluyendo el Proceso Dirichlet (Ferguson, 1973), Mezclas basadas en Procesos Dirichlet (Lo, 1984) y Árboles de Pólya (Lavine, 1992). Para más detalles, pueden consultarse los trabajos de Walker et al. (1999) y Dey et al. (1998).

- [Bayesiano Paramétrico Necio $<$ Bayesiano Paramétrico Incoherente]
- [Bayesiano Paramétrico Incoherente $<<$ Bayesiano No Paramétrico]

Motivación - Los modelos paramétricos son útiles

- ¡Pero los procedimientos paramétricos son útiles!
- Muchos estadísticos prefieren trabajar con modelos paramétricos.
- Los motivos incluyen la relativa simplicidad del análisis, de la interpretación y de la comunicación de resultados.
- La posible incoherencia inherente a los procedimientos de selección de modelos paramétricos generalmente se ignora por pragmatismo.

Motivación - Los modelos paramétricos son útiles

- ¡Pero los procedimientos paramétricos son útiles!
- Muchos estadísticos prefieren trabajar con modelos paramétricos.
- Los motivos incluyen la relativa simplicidad del análisis, de la interpretación y de la comunicación de resultados.
- La posible incoherencia inherente a los procedimientos de selección de modelos paramétricos generalmente se ignora por pragmatismo.

Motivación - Los modelos paramétricos son útiles

- ¡Pero los procedimientos paramétricos son útiles!
- Muchos estadísticos prefieren trabajar con modelos paramétricos.
- Los motivos incluyen la relativa simplicidad del análisis, de la interpretación y de la comunicación de resultados.
- La posible incoherencia inherente a los procedimientos de selección de modelos paramétricos generalmente se ignora por pragmatismo.

Motivación - Los modelos paramétricos son útiles

- ¡Pero los procedimientos paramétricos son útiles!
- Muchos estadísticos prefieren trabajar con modelos paramétricos.
- Los motivos incluyen la relativa simplicidad del análisis, de la interpretación y de la comunicación de resultados.
- La posible incoherencia inherente a los procedimientos de selección de modelos paramétricos generalmente se ignora por pragmatismo.

Motivación - Los modelos paramétricos son útiles

- En este trabajo se muestra cómo es posible realizar inferencias paramétricas evitando comportamientos incoherentes.
- El marco es Bayesiano no paramétrico y las inferencias se realizan vía la Teoría de la Decisión.
- Ninguno de los ingredientes discutidos aquí es nuevo, pero el argumento sólo se hace evidente cuando las distribuciones iniciales son vistas como medidas sobre familias generales de densidades en lugar de medidas sobre espacios parametrales de dimensión finita, a pesar del deseo de hacer inferencias paramétricas.
- [Decisiones/Preferencias vs Sucesos/Creencias; c.f. Hipótesis simples]

Motivación - Los modelos paramétricos son útiles

- En este trabajo se muestra cómo es posible realizar inferencias paramétricas evitando comportamientos incoherentes.
- El marco es Bayesiano no paramétrico y las inferencias se realizan vía la Teoría de la Decisión.
- Ninguno de los ingredientes discutidos aquí es nuevo, pero el argumento sólo se hace evidente cuando las distribuciones iniciales son vistas como medidas sobre familias generales de densidades en lugar de medidas sobre espacios parametrales de dimensión finita, a pesar del deseo de hacer inferencias paramétricas.
- [Decisiones/Preferencias vs Sucesos/Creencias; c.f. Hipótesis simples]

Motivación - Los modelos paramétricos son útiles

- En este trabajo se muestra cómo es posible realizar inferencias paramétricas evitando comportamientos incoherentes.
- El marco es Bayesiano no paramétrico y las inferencias se realizan vía la Teoría de la Decisión.
- Ninguno de los ingredientes discutidos aquí es nuevo, pero el argumento sólo se hace evidente cuando las distribuciones iniciales son vistas como medidas sobre familias generales de densidades en lugar de medidas sobre espacios parametrales de dimensión finita, a pesar del deseo de hacer inferencias paramétricas.
- [Decisiones/Preferencias vs Sucesos/Creencias; c.f. Hipótesis simples]

Motivación - Los modelos paramétricos son útiles

- En este trabajo se muestra cómo es posible realizar inferencias paramétricas evitando comportamientos incoherentes.
- El marco es Bayesiano no paramétrico y las inferencias se realizan vía la Teoría de la Decisión.
- Ninguno de los ingredientes discutidos aquí es nuevo, pero el argumento sólo se hace evidente cuando las distribuciones iniciales son vistas como medidas sobre familias generales de densidades en lugar de medidas sobre espacios parametrales de dimensión finita, a pesar del deseo de hacer inferencias paramétricas.
- [Decisiones/Preferencias vs Sucesos/Creencias; c.f. Hipótesis simples]

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*: $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*: $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Decisiones

- Sea $\mathbf{X}_n = (X_1, \ldots, X_n)$ una muestra de observaciones i.i.d. de una densidad desconocida f definida sobre $\mathcal{X}$. Consideremos el siguiente problema de decisión estadístico, con elementos:
  - *Espacio de decisiones*:
    $\mathcal{D}$, el conjunto de todas las decisiones relevantes.
  - *Estados de la naturaleza*:
    $\Omega = \{f : f \text{ es una función de densidad sobre } \mathcal{X}\}$.
  - *Distribución inicial*: $\Pi(\cdot)$, una distribución no paramétrica sobre Ω.
  - *Función de utilidad*: $U(a, f)$, con $a \in \mathcal{D}$ y $f \in \Omega$.
- La mejor decisión es elegir $a^* \in \mathcal{D}$ tal que maximice a

$$U_n(a) = \int_{\Omega} U(a, f)\, \Pi(df|\mathbf{x}_n).$$

- *La propuesta es hacer inferencias eligiendo apropiadamente a $\mathcal{D}$ y a $U(a, f)$ en cada problema.*

Inferencia estadística como un problema de decisión

- Bernardo & Smith (1994) argumentan que el problema de hacer inferencias debe verse como un problema de decisión estadístico en el que el espacio de decisiones es el conjunto de densidades de probabilidad para la cantidad de interés.
- Más aún, proponen el uso de la función de puntaje logarítmica como una función de utilidad. Su contexto es paramétrico, pero el argumento es análogo en el caso no paramétrico aquí descrito.
- Desde un punto de vista predictivo, la *cantidad de interés* es el valor desconocido de una observación futura, X.
- El *espacio de decisiones* es por lo tanto $\mathcal{D} \equiv \Omega$, el conjunto de todas las densidades definidas sobre $\mathcal{X}$.

Inferencia estadística como un problema de decisión

- Bernardo & Smith (1994) argumentan que el problema de hacer inferencias debe verse como un problema de decisión estadístico en el que el espacio de decisiones es el conjunto de densidades de probabilidad para la cantidad de interés.
- Más aún, proponen el uso de la función de puntaje logarítmica como una función de utilidad. Su contexto es paramétrico, pero el argumento es análogo en el caso no paramétrico aquí descrito.
- Desde un punto de vista predictivo, la *cantidad de interés* es el valor desconocido de una observación futura, X.
- El *espacio de decisiones* es por lo tanto $\mathcal{D} \equiv \Omega$, el conjunto de todas las densidades definidas sobre $\mathcal{X}$.

Inferencia estadística como un problema de decisión

- Bernardo & Smith (1994) argumentan que el problema de hacer inferencias debe verse como un problema de decisión estadístico en el que el espacio de decisiones es el conjunto de densidades de probabilidad para la cantidad de interés.
- Más aún, proponen el uso de la función de puntaje logarítmica como una función de utilidad. Su contexto es paramétrico, pero el argumento es análogo en el caso no paramétrico aquí descrito.
- Desde un punto de vista predictivo, la *cantidad de interés* es el valor desconocido de una observación futura, X.
- El *espacio de decisiones* es por lo tanto $\mathcal{D} \equiv \Omega$, el conjunto de todas las densidades definidas sobre $\mathcal{X}$.

Inferencia estadística como un problema de decisión

- Bernardo & Smith (1994) argumentan que el problema de hacer inferencias debe verse como un problema de decisión estadístico en el que el espacio de decisiones es el conjunto de densidades de probabilidad para la cantidad de interés.
- Más aún, proponen el uso de la función de puntaje logarítmica como una función de utilidad. Su contexto es paramétrico, pero el argumento es análogo en el caso no paramétrico aquí descrito.
- Desde un punto de vista predictivo, la *cantidad de interés* es el valor desconocido de una observación futura, X.
- El *espacio de decisiones* es por lo tanto $\mathcal{D} \equiv \Omega$, el conjunto de todas las densidades definidas sobre $\mathcal{X}$.

Inferencia estadística como un problema de decisión

- Aquí utilizaremos la siguiente función de utilidad:

$$U(\hat{f}, f) = \int \log\{\hat{f}(x)\} f(x)\, dx.$$

- Maximizarla equivale a minimizar $d_K(\hat{f}|f)$, la divergencia de Kullback-Leibler entre las densidades $\hat{f}$ y f.
- La utilidad esperada correspondiente es:

$$U_n(\hat{f}) = \int \log\{\hat{f}(x)\} f_n(x)\, dx,$$

donde $f_n \equiv E[f|\mathbf{x}_n]$ y el valor esperado es respecto a la distribución final no paramétrica.

- La solución a este problema de maximización es $\hat{f}^* = f_n$, i.e. la densidad predictiva final (no paramétrica).

Inferencia estadística como un problema de decisión

- Aquí utilizaremos la siguiente función de utilidad:

$$U(\hat{f}, f) = \int \log\{\hat{f}(x)\} f(x)\, dx.$$

- Maximizarla equivale a minimizar $d_K(\hat{f}|f)$, la divergencia de Kullback-Leibler entre las densidades $\hat{f}$ y f.
- La utilidad esperada correspondiente es:

$$U_n(\hat{f}) = \int \log\{\hat{f}(x)\} f_n(x)\, dx,$$

donde $f_n \equiv E[f|\mathbf{x}_n]$ y el valor esperado es respecto a la distribución final no paramétrica.

- La solución a este problema de maximización es $\hat{f}^* = f_n$, i.e. la densidad predictiva final (no paramétrica).

Inferencia estadística como un problema de decisión

- Aquí utilizaremos la siguiente función de utilidad:

$$U(\hat{f}, f) = \int \log\{\hat{f}(x)\} f(x)\, dx.$$

- Maximizarla equivale a minimizar $d_K(\hat{f}|f)$, la divergencia de Kullback-Leibler entre las densidades $\hat{f}$ y f.
- La utilidad esperada correspondiente es:

$$U_n(\hat{f}) = \int \log\{\hat{f}(x)\} f_n(x)\, dx,$$

donde $f_n \equiv E[f|\mathbf{x}_n]$ y el valor esperado es respecto a la distribución final no paramétrica.

- La solución a este problema de maximización es $\hat{f}^* = f_n$, i.e. la densidad predictiva final (no paramétrica).

Inferencia estadística como un problema de decisión

- Aquí utilizaremos la siguiente función de utilidad:

$$U(\hat{f}, f) = \int \log\{\hat{f}(x)\} f(x)\, dx.$$

- Maximizarla equivale a minimizar $d_K(\hat{f}|f)$, la divergencia de Kullback-Leibler entre las densidades $\hat{f}$ y f.
- La utilidad esperada correspondiente es:

$$U_n(\hat{f}) = \int \log\{\hat{f}(x)\} f_n(x)\, dx,$$

donde $f_n \equiv E[f|\mathbf{x}_n]$ y el valor esperado es respecto a la distribución final no paramétrica.

- La solución a este problema de maximización es $\hat{f}^* = f_n$, i.e. la densidad predictiva final (no paramétrica).

Distribuciones predictivas sustitutas

- Sin embargo, un "estadístico paramétrico" puede optar por concentrar su atención en una clase

$$\mathbf{F}_\Lambda = \{f(\cdot;\lambda) : \lambda \in \Lambda\}$$

de modelos más simples y más suaves.

- El problema es entonces encontrar una densidad predictiva *sustituta*, que se usará como una alternativa simple de la densidad predictiva no paramétrica (f_n) o como un estimador de la densidad desconocida f.

Distribuciones predictivas sustitutas

- Sin embargo, un "estadístico paramétrico" puede optar por concentrar su atención en una clase

$$\mathbf{F}_\Lambda = \{f(\cdot;\lambda) : \lambda \in \Lambda\}$$

de modelos más simples y más suaves.

- El problema es entonces encontrar una densidad predictiva *sustituta*, que se usará como una alternativa simple de la densidad predictiva no paramétrica (f_n) o como un estimador de la densidad desconocida f.

Distribuciones predictivas sustitutas

- La inferencias requieren de la elección de un $\lambda \in \Lambda$, donde Λ y $f_\lambda(\cdot) \equiv f(\cdot;\lambda)$ deben especificarse en cada problema.
- En este caso el espacio de decisiones puede identificarse con el conjunto de índices Λ, de manera que $\mathcal{D} \equiv \Lambda$.
- Esta es la formulación más general. Se selecciona la densidad predictiva paramétrica (es decir, se selecciona $\lambda^* \in \Lambda$) tal que maximiza:

$$U_n(\lambda) = \int \log\{f_\lambda(x)\}\, f_n(x)\, dx.$$

Distribuciones predictivas sustitutas

- La inferencias requieren de la elección de un $\lambda \in \Lambda$, donde Λ y $f_\lambda(\cdot) \equiv f(\cdot; \lambda)$ deben especificarse en cada problema.
- En este caso el espacio de decisiones puede identificarse con el conjunto de índices Λ, de manera que $\mathcal{D} \equiv \Lambda$.
- Esta es la formulación más general. Se selecciona la densidad predictiva paramétrica (es decir, se selecciona $\lambda^* \in \Lambda$) tal que maximiza:

$$U_n(\lambda) = \int \log\{f_\lambda(x)\}\, f_n(x)\, dx.$$

Distribuciones predictivas sustitutas

- La inferencias requieren de la elección de un $\lambda \in \Lambda$, donde Λ y $f_\lambda(\cdot) \equiv f(\cdot;\lambda)$ deben especificarse en cada problema.
- En este caso el espacio de decisiones puede identificarse con el conjunto de índices Λ, de manera que $\mathcal{D} \equiv \Lambda$.
- Esta es la formulación más general. Se selecciona la densidad predictiva paramétrica (es decir, se selecciona $\lambda^* \in \Lambda$) tal que maximiza:

$$U_n(\lambda) = \int \log\{f_\lambda(x)\}\, f_n(x)\, dx.$$

Distribuciones predictivas sustitutas

- En resumen, el procedimiento consiste en:

  (1) establecer la forma adecuada de la inferencia deseada;
  (2) asociar ésta con una familia paramétrica de densidades indexada por λ; y
  (3) seleccionar el valor de λ que maximiza $U_n(\lambda)$.

- Un caso particular importante se obtiene con el *bootstrap* Bayesiano:

$$U_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log f_\lambda(x_i).$$

- Maximizar esta función es equivalente a maximizar:

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i).$$

Distribuciones predictivas sustitutas

- En resumen, el procedimiento consiste en:

  (1) establecer la forma adecuada de la inferencia deseada;
  (2) asociar ésta con una familia paramétrica de densidades indexada por λ; y
  (3) seleccionar el valor de λ que maximiza $U_n(\lambda)$.

- Un caso particular importante se obtiene con el *bootstrap* Bayesiano:

$$U_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log f_\lambda(x_i).$$

- Maximizar esta función es equivalente a maximizar:

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i).$$

Distribuciones predictivas sustitutas

- En resumen, el procedimiento consiste en:

  (1) establecer la forma adecuada de la inferencia deseada;
  (2) asociar ésta con una familia paramétrica de densidades indexada por λ; y
  (3) seleccionar el valor de λ que maximiza $U_n(\lambda)$.

- Un caso particular importante se obtiene con el *bootstrap* Bayesiano:

$$U_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log f_\lambda(x_i).$$

- Maximizar esta función es equivalente a maximizar:

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i).$$

Distribuciones predictivas sustitutas

- *¿Cómo elegir familias de densidades predictivas apropiadas*?
- Eso depende del analista y del problema en cuestión.
- Sin embargo, la práctica usual sugiere una forma simple y estructurada de escoger $\mathbf{F}_\Lambda$.
- La idea es derivar las densidades predictivas sustitutas a partir de la colección de modelos paramétricos que se están considerando para el problema en cuestión:

$$\mathcal{M} = \{M_j : j = 1, \ldots, M\},$$

donde

$$M_j = \{p_j(\cdot; \theta_j), \pi_j(\theta_j) : \theta_j \in \Theta_j\}.$$

Distribuciones predictivas sustitutas

- *¿Cómo elegir familias de densidades predictivas apropiadas*?
- Eso depende del analista y del problema en cuestión.
- Sin embargo, la práctica usual sugiere una forma simple y estructurada de escoger $\mathbf{F}_\Lambda$.
- La idea es derivar las densidades predictivas sustitutas a partir de la colección de modelos paramétricos que se están considerando para el problema en cuestión:

$$\mathcal{M} = \{M_j : j = 1, \ldots, M\},$$

donde

$$M_j = \{p_j(\cdot; \theta_j), \pi_j(\theta_j) : \theta_j \in \Theta_j\}.$$

Distribuciones predictivas sustitutas

- *¿Cómo elegir familias de densidades predictivas apropiadas*?
- Eso depende del analista y del problema en cuestión.
- Sin embargo, la práctica usual sugiere una forma simple y estructurada de escoger $\mathbf{F}_\Lambda$.
- La idea es derivar las densidades predictivas sustitutas a partir de la colección de modelos paramétricos que se están considerando para el problema en cuestión:

$$\mathcal{M} = \{M_j : j = 1, \ldots, M\},$$

donde

$$M_j = \{p_j(\cdot; \theta_j), \pi_j(\theta_j) : \theta_j \in \Theta_j\}.$$

Distribuciones predictivas sustitutas

- *¿Cómo elegir familias de densidades predictivas apropiadas*?
- Eso depende del analista y del problema en cuestión.
- Sin embargo, la práctica usual sugiere una forma simple y estructurada de escoger $\mathbf{F}_\Lambda$.
- La idea es derivar las densidades predictivas sustitutas a partir de la colección de modelos paramétricos que se están considerando para el problema en cuestión:

$$\mathcal{M} = \{M_j : j = 1, \ldots, M\},$$

donde

$$M_j = \{p_j(\cdot; \theta_j),\ \pi_j(\theta_j) : \theta_j \in \Theta_j\}.$$

Distribuciones predictivas sustitutas

- En la mayoría de los casos $\pi_j(\cdot)$ también tiene una estructura paramétrica:

$$\pi_j(\cdot) \equiv \pi_j(\cdot; \phi_j), \qquad \phi_i \in \Phi_j,\ j = 1, \ldots, M,$$

de manera que cada $\pi_j(\cdot)$ sólo depende de un parámetro ϕ_j de dimensión finita.

- Las densidades $\pi_j(\cdot)$ deben considerarse simplemente como "artificios matemáticos", convenientes para construir densidades predictivas paramétricas, no como *distribuciones iniciales reales*.
- El interés se centra en las densidades predictivas, y por lo tanto la única distribución inicial que debe reconocerse es Π, la inicial (no paramétrica) sobre el espacio de todas las densidades.
- Este enfoque cubre todos los problemas de inferencia tradicionales, de selección de modelos a estimación puntual.

Distribuciones predictivas sustitutas

- En la mayoría de los casos $\pi_j(\cdot)$ también tiene una estructura paramétrica:

$$\pi_j(\cdot) \equiv \pi_j(\cdot; \phi_j), \qquad \phi_i \in \Phi_j,\ j = 1, \ldots, M,$$

de manera que cada $\pi_j(\cdot)$ sólo depende de un parámetro ϕ_j de dimensión finita.

- Las densidades $\pi_j(\cdot)$ deben considerarse simplemente como "artificios matemáticos", convenientes para construir densidades predictivas paramétricas, no como *distribuciones iniciales reales*.

- El interés se centra en las densidades predictivas, y por lo tanto la única distribución inicial que debe reconocerse es Π, la inicial (no paramétrica) sobre el espacio de todas las densidades.

- Este enfoque cubre todos los problemas de inferencia tradicionales, de selección de modelos a estimación puntual.

Distribuciones predictivas sustitutas

- En la mayoría de los casos $\pi_j(\cdot)$ también tiene una estructura paramétrica:

$$\pi_j(\cdot) \equiv \pi_j(\cdot; \phi_j), \qquad \phi_i \in \Phi_j,\ j = 1, \ldots, M,$$

de manera que cada $\pi_j(\cdot)$ sólo depende de un parámetro ϕ_j de dimensión finita.

- Las densidades $\pi_j(\cdot)$ deben considerarse simplemente como "artificios matemáticos", convenientes para construir densidades predictivas paramétricas, no como *distribuciones iniciales reales*.

- El interés se centra en las densidades predictivas, y por lo tanto la única distribución inicial que debe reconocerse es Π, la inicial (no paramétrica) sobre el espacio de todas las densidades.

- Este enfoque cubre todos los problemas de inferencia tradicionales, de selección de modelos a estimación puntual.

Distribuciones predictivas sustitutas

- En la mayoría de los casos $\pi_j(\cdot)$ también tiene una estructura paramétrica:

$$\pi_j(\cdot) \equiv \pi_j(\cdot; \phi_j), \qquad \phi_i \in \Phi_j,\ j = 1, \ldots, M,$$

de manera que cada $\pi_j(\cdot)$ sólo depende de un parámetro ϕ_j de dimensión finita.

- Las densidades $\pi_j(\cdot)$ deben considerarse simplemente como "artificios matemáticos", convenientes para construir densidades predictivas paramétricas, no como *distribuciones iniciales reales*.

- El interés se centra en las densidades predictivas, y por lo tanto la única distribución inicial que debe reconocerse es Π, la inicial (no paramétrica) sobre el espacio de todas las densidades.

- Este enfoque cubre todos los problemas de inferencia tradicionales, de selección de modelos a estimación puntual.

Problemas tradicionales - Modelos ponderados

- Aquí $\lambda = \{\boldsymbol{\mu}, \pi_1(\cdot), \ldots, \pi_M(\cdot)\}$ y $\Lambda = W \times \mathbf{P}_1 \times \cdots \times \mathbf{P}_M$, donde

$$W = \left\{ \boldsymbol{\mu} \in \mathrm{IR}^M : \mu_j > 0, j = 1, \ldots, M; \sum_{j=1}^{M} \mu_j = 1 \right\}$$

y $\mathbf{P}_j = \mathbf{P}(\Theta_j)$ para cada $j = 1, \ldots, M$, con

$$\mathbf{P}(\Theta) \equiv \{\pi(\cdot) : \pi(\cdot) \text{ es una f.d.p. sobre } \Theta\}.$$

- En este caso la densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \sum_{j=1}^{M} \mu_j \, f_j(\cdot; \pi_j),$$

donde

$$f_j(\cdot; \pi_j) = \int p_j(\cdot; \theta_j) \, \pi_j(\theta_j) \, d\theta_j.$$

- [Estimación m.v. para los parámetros de una mezcla...]

Problemas tradicionales - Modelos ponderados

- Aquí $\lambda = \{\boldsymbol{\mu}, \pi_1(\cdot), \ldots, \pi_M(\cdot)\}$ y $\Lambda = W \times \mathbf{P}_1 \times \cdots \times \mathbf{P}_M$, donde

$$W = \left\{ \boldsymbol{\mu} \in \mathbb{R}^M : \mu_j > 0, j = 1, \ldots, M; \sum_{j=1}^{M} \mu_j = 1 \right\}$$

y $\mathbf{P}_j = \mathbf{P}(\Theta_j)$ para cada $j = 1, \ldots, M$, con

$$\mathbf{P}(\Theta) \equiv \{\pi(\cdot) : \pi(\cdot) \text{ es una f.d.p. sobre } \Theta\}.$$

- En este caso la densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \sum_{j=1}^{M} \mu_j \, f_j(\cdot; \pi_j),$$

donde

$$f_j(\cdot; \pi_j) = \int p_j(\cdot; \theta_j) \, \pi_j(\theta_j) \, d\theta_j.$$

- [Estimación m.v. para los parámetros de una mezcla...]

Problemas tradicionales - Modelos ponderados

- Aquí $\lambda = \{\boldsymbol{\mu}, \pi_1(\cdot), \ldots, \pi_M(\cdot)\}$ y $\Lambda = W \times \mathbf{P}_1 \times \cdots \times \mathbf{P}_M$, donde

$$W = \left\{ \boldsymbol{\mu} \in \mathbb{R}^M : \mu_j > 0, j = 1, \ldots, M; \sum_{j=1}^{M} \mu_j = 1 \right\}$$

y $\mathbf{P}_j = \mathbf{P}(\Theta_j)$ para cada $j = 1, \ldots, M$, con

$$\mathbf{P}(\Theta) \equiv \{\pi(\cdot) : \pi(\cdot) \text{ es una f.d.p. sobre } \Theta\}.$$

- En este caso la densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \sum_{j=1}^{M} \mu_j \, f_j(\cdot; \pi_j),$$

donde

$$f_j(\cdot; \pi_j) = \int p_j(\cdot; \theta_j) \, \pi_j(\theta_j) \, d\theta_j.$$

- [Estimación m.v. para los parámetros de una mezcla...]

Problemas tradicionales - Selección de modelos

- Este es un caso particular del de modelos ponderados, donde la medida de probabilidad μ se degenera en uno de los M modelos, i.e.

$$W = \left\{ \mu \in \mathbb{R}^M : \mu_k = \mathrm{I}(k = j);\ j, k = 1, \ldots, M \right\}.$$

- Por lo tanto $\lambda = \{j, \pi_j(\cdot)\}$ y $\Lambda = \bigcup_{j=1}^{M} (\{j\} \times \mathbf{P}_j)$.
- La densidad predictiva sustituta está dada en este caso por

$$f_\lambda(\cdot) = f_j(\cdot; \pi_j).$$

- Al ser un caso particular, *seleccionar un modelo nunca es preferible a ponderar los modelos* (en el sentido de que nunca produce una utilidad esperada mayor).

Problemas tradicionales - Selección de modelos

- Este es un caso particular del de modelos ponderados, donde la medida de probabilidad μ se degenera en uno de los M modelos, i.e.

$$W = \left\{ \mu \in \mathrm{I\!R}^M : \mu_k = \mathrm{I}(k = j);\ j, k = 1, \ldots, M \right\}.$$

- Por lo tanto $\lambda = \{j, \pi_j(\cdot)\}$ y $\Lambda = \bigcup_{j=1}^{M} (\{j\} \times \mathbf{P}_j)$.
- La densidad predictiva sustituta está dada en este caso por

$$f_\lambda(\cdot) = f_j(\cdot; \pi_j).$$

- Al ser un caso particular, *seleccionar un modelo nunca es preferible a ponderar los modelos* (en el sentido de que nunca produce una utilidad esperada mayor).

Problemas tradicionales - Selección de modelos

- Este es un caso particular del de modelos ponderados, donde la medida de probabilidad μ se degenera en uno de los M modelos, i.e.

$$W = \left\{ \mu \in \mathrm{I\!R}^M : \mu_k = \mathrm{I}(k = j);\ j, k = 1, \ldots, M \right\}.$$

- Por lo tanto $\lambda = \{j, \pi_j(\cdot)\}$ y $\Lambda = \bigcup_{j=1}^{M} (\{j\} \times \mathbf{P}_j)$.
- La densidad predictiva sustituta está dada en este caso por

$$f_\lambda(\cdot) = f_j(\cdot; \pi_j).$$

- Al ser un caso particular, *seleccionar un modelo nunca es preferible a ponderar los modelos* (en el sentido de que nunca produce una utilidad esperada mayor).

Problemas tradicionales - Selección de modelos

- Este es un caso particular del de modelos ponderados, donde la medida de probabilidad μ se degenera en uno de los M modelos, i.e.

$$W = \left\{ \mu \in \mathrm{I\!R}^M : \mu_k = \mathrm{I}(k = j);\ j, k = 1, \ldots, M \right\}.$$

- Por lo tanto $\lambda = \{j, \pi_j(\cdot)\}$ y $\Lambda = \bigcup_{j=1}^{M} (\{j\} \times \mathbf{P}_j)$.
- La densidad predictiva sustituta está dada en este caso por

$$f_\lambda(\cdot) = f_j(\cdot; \pi_j).$$

- Al ser un caso particular, *seleccionar un modelo nunca es preferible a ponderar los modelos* (en el sentido de que nunca produce una utilidad esperada mayor).

Problemas tradicionales - Contraste de hipótesis

- Contrastar hipótesis acerca del valor de un parámetro θ que indexa a una familia de densidades $\Omega_0 = \{p(\cdot;\theta) : \theta \in \Theta\}$ puede considerarse como un problema de selección de modelos.
- Sean $\Theta_1 \subset \Theta$ y $\Theta_2 \subset \Theta$ tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y supongamos que se desea contratar $H_1 : \theta \in \Theta_1$ vs $H_2 : \theta \in \Theta_2$. Supongamos también que θ tiene una "distribución inicial" $\pi(\theta)$.
- En este caso $M = 2$ y

$$M_j = \left\{ p_j(\cdot;\theta) \equiv p(\cdot;\theta),\ \pi_j(\theta) \equiv \frac{\pi(\theta)\ I_{\Theta_j}(\theta)}{\int_{\Theta_j} \pi(\theta)\, d\theta} \right\}, \qquad j = 1, 2,$$

donde $I_\Theta(\cdot)$ denota a la función indicadora del conjunto Θ.

- Aquí $\lambda = \{j, \pi(\cdot)\}$ y $\Lambda = \{1, 2\} \times \mathbf{P}(\Theta)$. La densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \int p(\cdot;\theta)\, \pi_j(\theta)\, d\theta.$$

Problemas tradicionales - Contraste de hipótesis

- Contrastar hipótesis acerca del valor de un parámetro θ que indexa a una familia de densidades $\Omega_0 = \{p(\cdot;\theta) : \theta \in \Theta\}$ puede considerarse como un problema de selección de modelos.
- Sean $\Theta_1 \subset \Theta$ y $\Theta_2 \subset \Theta$ tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y supongamos que se desea contratar $H_1 : \theta \in \Theta_1$ vs $H_2 : \theta \in \Theta_2$. Supongamos también que θ tiene una "distribución inicial" $\pi(\theta)$.
- En este caso $M = 2$ y

$$M_j = \left\{ p_j(\cdot;\theta) \equiv p(\cdot;\theta),\ \pi_j(\theta) \equiv \frac{\pi(\theta)\ I_{\Theta_j}(\theta)}{\int_{\Theta_j} \pi(\theta)\, d\theta} \right\}, \qquad j = 1, 2,$$

donde $I_\Theta(\cdot)$ denota a la función indicadora del conjunto Θ.
- Aquí $\lambda = \{j, \pi(\cdot)\}$ y $\Lambda = \{1, 2\} \times \mathbf{P}(\Theta)$. La densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \int p(\cdot;\theta)\, \pi_j(\theta)\, d\theta.$$

Problemas tradicionales - Contraste de hipótesis

- Contrastar hipótesis acerca del valor de un parámetro θ que indexa a una familia de densidades $\Omega_0 = \{p(\cdot;\theta) : \theta \in \Theta\}$ puede considerarse como un problema de selección de modelos.
- Sean $\Theta_1 \subset \Theta$ y $\Theta_2 \subset \Theta$ tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y supongamos que se desea contratar $H_1 : \theta \in \Theta_1$ vs $H_2 : \theta \in \Theta_2$. Supongamos también que θ tiene una "distribución inicial" $\pi(\theta)$.
- En este caso $M = 2$ y

$$M_j = \left\{ p_j(\cdot;\theta) \equiv p(\cdot;\theta),\ \pi_j(\theta) \equiv \frac{\pi(\theta)\ I_{\Theta_j}(\theta)}{\int_{\Theta_j} \pi(\theta)\, d\theta} \right\}, \qquad j = 1, 2,$$

donde $I_\Theta(\cdot)$ denota a la función indicadora del conjunto Θ.

- Aquí $\lambda = \{j, \pi(\cdot)\}$ y $\Lambda = \{1, 2\} \times \mathbf{P}(\Theta)$. La densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \int p(\cdot;\theta)\, \pi_j(\theta)\, d\theta.$$

Problemas tradicionales - Contraste de hipótesis

- Contrastar hipótesis acerca del valor de un parámetro θ que indexa a una familia de densidades $\Omega_0 = \{p(\cdot;\theta) : \theta \in \Theta\}$ puede considerarse como un problema de selección de modelos.
- Sean $\Theta_1 \subset \Theta$ y $\Theta_2 \subset \Theta$ tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y supongamos que se desea contratar $H_1 : \theta \in \Theta_1$ vs $H_2 : \theta \in \Theta_2$. Supongamos también que θ tiene una "distribución inicial" $\pi(\theta)$.
- En este caso $M = 2$ y

$$M_j = \left\{ p_j(\cdot;\theta) \equiv p(\cdot;\theta),\ \pi_j(\theta) \equiv \frac{\pi(\theta)\ I_{\Theta_j}(\theta)}{\int_{\Theta_j} \pi(\theta)\, d\theta} \right\}, \qquad j = 1, 2,$$

donde $I_\Theta(\cdot)$ denota a la función indicadora del conjunto Θ.

- Aquí $\lambda = \{j, \pi(\cdot)\}$ y $\Lambda = \{1, 2\} \times \mathbf{P}(\Theta)$. La densidad predictiva sustituta está dada por

$$f_\lambda(\cdot) = \int p(\cdot;\theta)\, \pi_j(\theta)\, d\theta.$$

Problemas tradicionales - Estimación por intervalos

- Como en el caso anterior, consideremos la familia paramétrica Ω_0 y supongamos que θ tiene una "inicial" $\pi(\theta)$ definida sobre Θ.
- Supongamos que estamos interesados en un intervalo de máxima densidad con "probabilidad" $(1-\alpha)$ c.r.a. π. Sea $\Theta(\pi,\alpha) = \{\theta \in \Theta : \pi(\theta) \geq p_\alpha\}$ tal intervalo, donde p_α es la constante más grande tal que

$$\int_{\Theta(\pi,\alpha)} \pi(\theta)\, d\theta = 1-\alpha.$$

- En este caso podemos tomar $\lambda = \{\pi(\cdot)\}$ con $\Lambda = \mathbf{P}(\Theta)$, y una densidad predictiva sustituta de la forma $f_\lambda(\cdot) = f(\cdot;\pi_\alpha)$, donde

$$f(\cdot;\pi_\alpha) = \int p(\cdot;\theta)\pi_\alpha(\theta)\, d\theta$$

y $\pi_\alpha(\theta) = (1-\alpha)^{-1}\,\pi(\theta)\,\mathsf{I}_{\Theta(\pi,\alpha)}(\theta)$.

Problemas tradicionales - Estimación por intervalos

- Como en el caso anterior, consideremos la familia paramétrica Ω_0 y supongamos que θ tiene una "inicial" $\pi(\theta)$ definida sobre Θ.
- Supongamos que estamos interesados en un intervalo de máxima densidad con "probabilidad" $(1-\alpha)$ c.r.a. π. Sea $\Theta(\pi,\alpha) = \{\theta \in \Theta : \pi(\theta) \geq p_\alpha\}$ tal intervalo, donde p_α es la constante más grande tal que

$$\int_{\Theta(\pi,\alpha)} \pi(\theta)\, d\theta = 1-\alpha.$$

- En este caso podemos tomar $\lambda = \{\pi(\cdot)\}$ con $\Lambda = \mathbf{P}(\Theta)$, y una densidad predictiva sustituta de la forma $f_\lambda(\cdot) = f(\cdot;\pi_\alpha)$, donde

$$f(\cdot;\pi_\alpha) = \int p(\cdot;\theta)\pi_\alpha(\theta)\, d\theta$$

y $\pi_\alpha(\theta) = (1-\alpha)^{-1}\,\pi(\theta)\,\mathrm{I}_{\Theta(\pi,\alpha)}(\theta)$.

Problemas tradicionales - Estimación por intervalos

- Como en el caso anterior, consideremos la familia paramétrica Ω_0 y supongamos que θ tiene una "inicial" $\pi(\theta)$ definida sobre Θ.
- Supongamos que estamos interesados en un intervalo de máxima densidad con "probabilidad" $(1-\alpha)$ c.r.a. π. Sea $\Theta(\pi,\alpha) = \{\theta \in \Theta : \pi(\theta) \geq p_\alpha\}$ tal intervalo, donde p_α es la constante más grande tal que

$$\int_{\Theta(\pi,\alpha)} \pi(\theta)\, d\theta = 1-\alpha.$$

- En este caso podemos tomar $\lambda = \{\pi(\cdot)\}$ con $\Lambda = \mathbf{P}(\Theta)$, y una densidad predictiva sustituta de la forma $f_\lambda(\cdot) = f(\cdot;\pi_\alpha)$, donde

$$f(\cdot;\pi_\alpha) = \int p(\cdot;\theta)\pi_\alpha(\theta)\, d\theta$$

y $\pi_\alpha(\theta) = (1-\alpha)^{-1}\, \pi(\theta)\, I_{\Theta(\pi,\alpha)}(\theta)$.

Problemas tradicionales - Estimación puntual

- Este también es un caso particular de selección de modelos donde $M = 1$ y $\pi(\cdot) = \delta_\theta(\cdot)$, i.e. $\pi(\cdot)$ se degenera en $\theta \in \Theta$.
- Aquí $\lambda = \theta$ y $\Lambda = \Theta$, y la densidad predictiva sustituta toma la forma

$$f_\lambda(\cdot) = p(\cdot; \theta).$$

- Escoger un valor de λ es equivalente a encontrar un estimador puntual para θ.
- [Máxima verosimilitud...]

Problemas tradicionales - Estimación puntual

- Este también es un caso particular de selección de modelos donde $M = 1$ y $\pi(\cdot) = \delta_\theta(\cdot)$, i.e. $\pi(\cdot)$ se degenera en $\theta \in \Theta$.
- Aquí $\lambda = \theta$ y $\Lambda = \Theta$, y la densidad predictiva sustituta toma la forma

$$f_\lambda(\cdot) = p(\cdot; \theta).$$

- Escoger un valor de λ es equivalente a encontrar un estimador puntual para θ.
- [Máxima verosimilitud...]

Problemas tradicionales - Estimación puntual

- Este también es un caso particular de selección de modelos donde $M = 1$ y $\pi(\cdot) = \delta_\theta(\cdot)$, i.e. $\pi(\cdot)$ se degenera en $\theta \in \Theta$.
- Aquí $\lambda = \theta$ y $\Lambda = \Theta$, y la densidad predictiva sustituta toma la forma

$$f_\lambda(\cdot) = p(\cdot; \theta).$$

- Escoger un valor de λ es equivalente a encontrar un estimador puntual para θ.
- [Máxima verosimilitud...]

Problemas tradicionales - Estimación puntual

- Este también es un caso particular de selección de modelos donde $M = 1$ y $\pi(\cdot) = \delta_\theta(\cdot)$, i.e. $\pi(\cdot)$ se degenera en $\theta \in \Theta$.
- Aquí $\lambda = \theta$ y $\Lambda = \Theta$, y la densidad predictiva sustituta toma la forma

$$f_\lambda(\cdot) = p(\cdot; \theta).$$

- Escoger un valor de λ es equivalente a encontrar un estimador puntual para θ.
- [Máxima verosimilitud...]

Problemas tradicionales - Distribución final

- En este caso $M = 1$, de manera que $\lambda = \{\pi(\cdot)\}$ y $\Lambda = \mathbf{P}(\Theta)$. Entonces

$$f_\lambda(\cdot) \equiv f(\cdot;\pi) = \int p(\cdot;\theta)\,\pi(\theta)\,d\theta.$$

- Seleccionar λ es equivalente a escoger una "densidad final" $\pi(\cdot)$ para θ.

- [¿Distribución inicial? - Procedimientos empíricos Bayesianos]

Problemas tradicionales - Distribución final

- En este caso $M = 1$, de manera que $\lambda = \{\pi(\cdot)\}$ y $\Lambda = \mathbf{P}(\Theta)$. Entonces

$$f_\lambda(\cdot) \equiv f(\cdot;\pi) = \int p(\cdot;\theta)\,\pi(\theta)\,d\theta.$$

- Seleccionar λ es equivalente a escoger una "densidad final" $\pi(\cdot)$ para θ.

- [¿Distribución inicial? - Procedimientos empíricos Bayesianos]

Problemas tradicionales - Distribución final

- En este caso $M = 1$, de manera que $\lambda = \{\pi(\cdot)\}$ y $\Lambda = \mathbf{P}(\Theta)$. Entonces

$$f_\lambda(\cdot) \equiv f(\cdot;\pi) = \int p(\cdot;\theta)\,\pi(\theta)\,d\theta.$$

- Seleccionar λ es equivalente a escoger una "densidad final" $\pi(\cdot)$ para θ.

- [¿Distribución inicial? - Procedimientos empíricos Bayesianos]

Comentarios finales

- Los enfoques Bayesianos tradicionales de selección de modelos pueden ser incoherentes.
- El enfoque predictivo de inferencia paramétrica discutido aquí produce un procedimiento de selección que evita este problema.
- La idea es usar un modelo que sea lo suficientemente flexible como para que no sea sujeto de verificación, sin importar la naturaleza de los datos observados.
  - Caso i.i.d.: modelos Bayesianos no paramétricos.
  - Otros casos (e.g. regresión): ¿modelos Bayesianos semiparamétricos?

- ¿Es inevitable que aparezca algún grado de incoherencia en un análisis Bayesiano?

Comentarios finales

- Los enfoques Bayesianos tradicionales de selección de modelos pueden ser incoherentes.
- El enfoque predictivo de inferencia paramétrica discutido aquí produce un procedimiento de selección que evita este problema.
- La idea es usar un modelo que sea lo suficientemente flexible como para que no sea sujeto de verificación, sin importar la naturaleza de los datos observados.
  - Caso i.i.d.: modelos Bayesianos no paramétricos.
  - Otros casos (e.g. regresión): ¿modelos Bayesianos semiparamétricos?
- ¿Es inevitable que aparezca algún grado de incoherencia en un análisis Bayesiano?

Referencias

- Bernardo, J.M. y Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: Wiley.
- Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society A* 143, 383–430.
- Dey, D., Sinha, D. y Müller, P. (eds.) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Lecture Notes in Statistics. New York: Springer.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* 57, 45–97.
- Draper, D. (1999). Discussion of the paper “Bayesian nonparametric inference for random distributions and related functions”, by Walker, et al. *Journal of the Royal Statistical Society B* 61, 485–527.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1, 209–230.
- Hirshleifer, J. y Riley, J.G. (1992). *The Analysis of Uncertainty and Information*. Cambridge: Cambridge University Press.
- **Gutiérrez-Peña, E. y Walker, S.G. (2005). Statistical Decision Problems and Bayesian Nonparametric Methods.** *International Statistical Review* 73, 309–330.
- Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Annals of Statistics* 20, 1203-1221.
- Lindsey, J.K. (1999). Some statistical heresies. *The Statistician* 48, 1–40.
- Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Annals of Statistics* 12, 351-357.
- Walker, S.G., Damien P., Laud, P.W. y Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B* 61, 485-527.

Ejemplo - Distribución final

- Seam Ω_0 la familia de las densidades normales con varianza 1, y consideremos la distribución inicial no paramétrica

$$X_i|\theta_i \sim \mathrm{N}(\theta_i, 1)$$
$$\theta_i|F \sim F$$
$$F \sim \mathcal{D}(c, G).$$

  donde $\mathcal{D}(c, G)$ es un proceso Dirichlet con $c = 1$ y $G = \mathrm{N}(0, 10^2)$.
- Denotemos por $\{Z_{n1}, \ldots, Z_{nN}\}$ a una muestra simulada de tamaño N de f_n.
- Entonces la utilidad esperada final puede aproximarse a través de

$$U_n(\lambda) \approx \frac{1}{N} \sum_{j=1}^{N} \log f(z_{nj}; \pi).$$

- Maximizar $U_n(\lambda)$ con respecto a π es lo mismo que maximizar

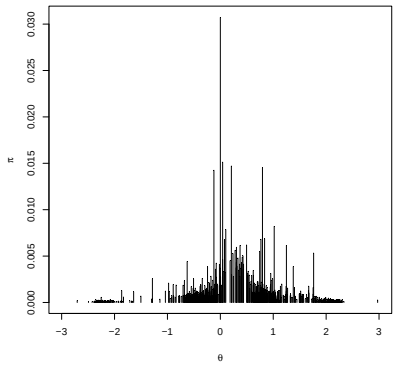
$$\prod_{j=1}^{N} f(z_{nj}; \pi),$$

  para lo cual existen diversos algoritmos. La solución es una distribución discreta $\hat{\pi}$ con soporte en a lo más N puntos.

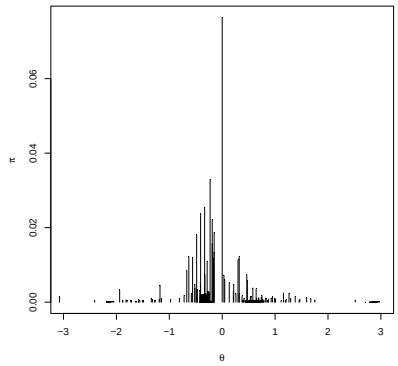
Ejemplo - Distribución final

- Se generaron cuatro muestras, de tamaños 10, 100, 1 000 y 10 000, de una densidad $\mathrm{N}(\theta_0, 1)$ con $\theta_0 = 0$.
- Para cada una de estas muestras, se encontró $\hat{\pi}$ con base en una muestra de Monte Carlo de tamaño $N = 10\,000$ de la correspondiente distribución predictiva no paramétrica.
- Las siguientes figuras muestran la solución $\hat{\pi}$ para cada uno de los cuatro casos.
- Se incluyen las varianzas paramétricas finales (respecto a distribuciones "iniciales" no informativas) y las correspondientes varianzas de $\hat{\pi}$.
- Como era de esperarse, las distribuciones resultantes tienden a concentrarse alrededor del valor $\theta_0 = 0$ conforme el tamaño de muestra crece.
- Las distribuciones finales paramétrica usuales son menos dispersas que la "distribuciones finales" óptimas $\hat{\pi}$.

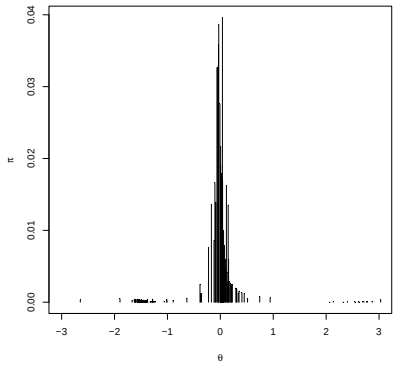
Ejemplo - Distribución final



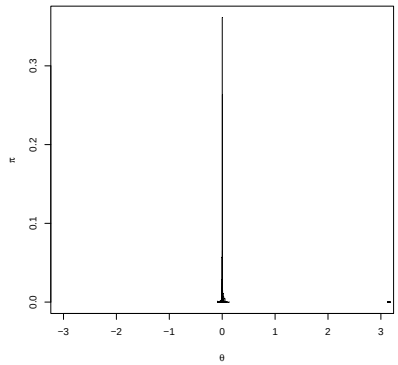
[$n = 10$] Varianza final: P=0.1; NP=0.66



[$n = 100$] Varianza final: P=0.01; NP=0.30



[$n = 1\,000$] Varianza final: P=0.001; NP=0.06



[$n = 10\,000$] Varianza final: P$\approx$0; NP$\approx$0