Elección de λ: el riesgo de predicción

Javier Olaya Ochoa

Programa Académico de Estadística
Escuela de Estadística
Universidad del Valle
Cali - Colombia

19 de noviembre de 2020

Contenido

Contenido

Más sobre el estimador *UBRE*

- El estimador *UBRE* es una opción para elegir λ que podría considerarse como el primer intento sustentado de selección.
- Su principal limitación es su dependencia del conocimiento de la varianza σ^2, por lo que sería interesante indagar si es posible proponer métodos de selección de λ que no dependan de ese conocimiento.

Más sobre el estimador *UBRE*

- El estimador *UBRE* es una opción para elegir λ que podría considerarse como el primer intento sustentado de selección.
- Su principal limitación es su dependencia del conocimiento de la varianza σ^2, por lo que sería interesante indagar si es posible proponer métodos de selección de λ que no dependan de ese conocimiento.

Otro criterio de desempeño

- Definamos primero otra medida de calidad del estimador μ_λ.
- Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales.
- Llamaremos $\boldsymbol{y}_N$ al vector de nuevas observaciones.
- Asumamos que se satisface que:

$$y_{Ni} = \mu(x_i) + \epsilon_{Ni},\ i = 1, \ldots, n$$

- Sea μ la misma función de regresión del modelo básico y los ϵ_{Ni} variables aleatorias no correlacionadas entre sí ni con los ϵ_i, con varianza común σ^2.

Otro criterio de desempeño

- Definamos primero otra medida de calidad del estimador μ_λ.
- Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales.
- Llamaremos $\boldsymbol{y}_N$ al vector de nuevas observaciones.
- Asumamos que se satisface que:

$$y_{Ni} = \mu(x_i) + \epsilon_{Ni},\ i = 1, \ldots, n$$

- Sea μ la misma función de regresión del modelo básico y los ϵ_{Ni} variables aleatorias no correlacionadas entre sí ni con los ϵ_i, con varianza común σ^2.

Otro criterio de desempeño

- Definamos primero otra medida de calidad del estimador μ_λ.
- Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales.
- Llamaremos $\boldsymbol{y_N}$ al vector de nuevas observaciones.
- Asumamos que se satisface que:

$$y_{Ni} = \mu(x_i) + \epsilon_{Ni},\ i = 1, \ldots, n$$

- Sea μ la misma función de regresión del modelo básico y los ϵ_{Ni} variables aleatorias no correlacionadas entre sí ni con los ϵ_i, con varianza común σ^2.

Otro criterio de desempeño

- Definamos primero otra medida de calidad del estimador μ_λ.
- Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales.
- Llamaremos $\boldsymbol{y_N}$ al vector de nuevas observaciones.
- Asumamos que se satisface que:

$$y_{Ni} = \mu(x_i) + \epsilon_{Ni},\ i = 1, \ldots, n$$

- Sea μ la misma función de regresión del modelo básico y los ϵ_{Ni} variables aleatorias no correlacionadas entre sí ni con los ϵ_i, con varianza común σ^2.

Otro criterio de desempeño

- Definamos primero otra medida de calidad del estimador μ_λ.
- Supongamos que nos proponemos obtener n nuevas observaciones de Y que se supone pueden ser modelados de la misma forma que los datos originales.
- Llamaremos $\boldsymbol{y_N}$ al vector de nuevas observaciones.
- Asumamos que se satisface que:

$$y_{Ni} = \mu(x_i) + \epsilon_{Ni},\ i = 1, \ldots, n$$

- Sea μ la misma función de regresión del modelo básico y los ϵ_{Ni} variables aleatorias no correlacionadas entre sí ni con los ϵ_i, con varianza común σ^2.

Riesgo de Predicción

- Digamos que queremos usar nuestro estimador μ_λ para predecir los y_{Ni}.
- Definamos entonces el *riesgo de predicción* $P(\lambda)$ como sigue:

$$P(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_{Ni} - \mu_{\lambda i})^2$$

Riesgo de Predicción

- Digamos que queremos usar nuestro estimador μ_λ para predecir los y_{Ni}.
- Definamos entonces el *riesgo de predicción* $P(\lambda)$ como sigue:

$$P(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_{Ni} - \mu_{\lambda i})^2$$

Riesgo de Predicción en términos del Riesgo

- El riesgo de predicción $P(\lambda)$ y el riesgo $R(\lambda)$ se relacionan de la siguiente manera:

$$P(\lambda) = \sigma^2 + R(\lambda)$$

- Esto permitiría definir un estimador no insesgado de $P(\lambda)$ basado en el estimador *UBRE*.
- Pero los resultados serían los mismos ya que estos dos métodos son equivalentes en el sentido que el λ que hace mínimo $P(\lambda)$ es el mismo que minimiza $R(\lambda)$.

Riesgo de Predicción en términos del Riesgo

- El riesgo de predicción $P(\lambda)$ y el riesgo $R(\lambda)$ se relacionan de la siguiente manera:

$$P(\lambda) = \sigma^2 + R(\lambda)$$

- Esto permitiría definir un estimador no insesgado de $P(\lambda)$ basado en el estimador *UBRE*.
- Pero los resultados serían los mismos ya que estos dos métodos son equivalentes en el sentido que el λ que hace mínimo $P(\lambda)$ es el mismo que minimiza $R(\lambda)$.

Riesgo de Predicción en términos del Riesgo

- El riesgo de predicción $P(\lambda)$ y el riesgo $R(\lambda)$ se relacionan de la siguiente manera:

$$P(\lambda) = \sigma^2 + R(\lambda)$$

- Esto permitiría definir un estimador no insesgado de $P(\lambda)$ basado en el estimador *UBRE*.
- Pero los resultados serían los mismos ya que estos dos métodos son equivalentes en el sentido que el λ que hace mínimo $P(\lambda)$ es el mismo que minimiza $R(\lambda)$.

Un segundo conjunto de datos

- Si uno dispusiera de un segundo conjunto de n datos, estimaría μ_λ con los primeros n datos y buscaría el λ que minimice $P(\lambda)$ usando el segundo conjunto de datos.
- En la práctica, sin embargo, sería mejor estimar μ utilizando los $2n$ datos, lo que nos lleva a continuar nuestra búsqueda de un nuevo procedimiento para elegir λ que no dependa de σ^2.

Un segundo conjunto de datos

- Si uno dispusiera de un segundo conjunto de n datos, estimaría μ_λ con los primeros n datos y buscaría el λ que minimice $P(\lambda)$ usando el segundo conjunto de datos.
- En la práctica, sin embargo, sería mejor estimar μ utilizando los $2n$ datos, lo que nos lleva a continuar nuestra búsqueda de un nuevo procedimiento para elegir λ que no dependa de σ^2.

Contenido

Una opción

- Una opción es dividir el conjunto de n observaciones en n sub-muestras de tamaño $n-1$ mediante el mecanismo de dejar por fuera una observación diferente cada vez.
- Si denotamos $\mu_{\lambda(i)}$ a la estimación de μ_i obtenida al suprimir de la muestra la observación i, entonces la observación y_i sería una observación adicional que podríamos utilizar para construir un estimador de $P(\lambda)$ que denotaremos $CV(\lambda)$ y que llamaremos criterio de *validación cruzada*

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \mu_{\lambda(i)}]^2$$

Una opción

- Una opción es dividir el conjunto de n observaciones en n sub-muestras de tamaño $n-1$ mediante el mecanismo de dejar por fuera una observación diferente cada vez.
- Si denotamos $\mu_{\lambda(i)}$ a la estimación de μ_i obtenida al suprimir de la muestra la observación i, entonces la observación y_i sería una observación adicional que podríamos utilizar para construir un estimador de $P(\lambda)$ que denotaremos $CV(\lambda)$ y que llamaremos criterio de *validación cruzada*

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \mu_{\lambda(i)}]^2$$

Una alternativa de cálculo

- Calcular el criterio *CV* por el método propuesto es complejo, porque requiere una gran carga de procesamiento computacional
- Green y Silverman (2000) sugieren simplificar este cálculo utilizando en cambio:

$$CV_{GS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mu_{\lambda i}}{1 - S_{\lambda ii}}\right)^2$$

- Donde $S_{\lambda ii}$ es el elemento i de la matriz S_λ.

Una alternativa de cálculo

- Calcular el criterio CV por el método propuesto es complejo, porque requiere una gran carga de procesamiento computacional
- Green y Silverman (2000) sugieren simplificar este cálculo utilizando en cambio:

$$CV_{GS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mu_{\lambda i}}{1 - S_{\lambda ii}}\right)^2$$

- Donde $S_{\lambda ii}$ es el elemento i de la matriz S_λ.

Una alternativa de cálculo

- Calcular el criterio CV por el método propuesto es complejo, porque requiere una gran carga de procesamiento computacional
- Green y Silverman (2000) sugieren simplificar este cálculo utilizando en cambio:

$$CV_{GS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mu_{\lambda i}}{1 - S_{\lambda ii}}\right)^2$$

- Donde $S_{\lambda ii}$ es el elemento i de la matriz S_λ.

GCV_{GS}

- Otra posible solución se debe a Wahba (1990), quien sugiere utilizar otro criterio llamado *validación cruzada generalizada* $GCV(\lambda)$.
- Green y Silverman (2000) lo definen como

$$GCV_{GS}(\lambda) = n^{-1} \frac{\sum_{i=1}^{n}(y_i - \mu_{\lambda i})^2}{(1 - n^{-1}\text{tr}[S_\lambda])^2}$$

- Una expresión más sencilla, sugerida por Eubank (1999) es

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{(n^{-1}\text{tr}[I - S_\lambda])^2}$$

GCV_{GS}

- Otra posible solución se debe a Wahba (1990), quien sugiere utilizar otro criterio llamado *validación cruzada generalizada* $GCV(\lambda)$.
- Green y Silverman (2000) lo definen como

$$GCV_{GS}(\lambda) = n^{-1} \frac{\sum_{i=1}^{n}(y_i - \mu_{\lambda i})^2}{(1 - n^{-1}\text{tr}[S_\lambda])^2}$$

- Una expresión más sencilla, sugerida por Eubank (1999) es

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{(n^{-1}\text{tr}[I - S_\lambda])^2}$$

GCV_{GS}

- Otra posible solución se debe a Wahba (1990), quien sugiere utilizar otro criterio llamado *validación cruzada generalizada* $GCV(\lambda)$.
- Green y Silverman (2000) lo definen como

$$GCV_{GS}(\lambda) = n^{-1} \frac{\sum_{i=1}^{n}(y_i - \mu_{\lambda i})^2}{(1 - n^{-1}\text{tr}[S_\lambda])^2}$$

- Una expresión más sencilla, sugerida por Eubank (1999) es

$$GCV(\lambda) = \frac{n^{-1} RSS(\lambda)}{(n^{-1}\text{tr}[I - S_\lambda])^2}$$

¿Generalizada?

- Aunque la palabra “generalizada” deja la impresión de que el segundo criterio generaliza el primero, esto no es en general cierto y se trata de criterios diferentes que permiten estimar el riesgo de predicción $P(\lambda)$.
- Wahba (1990) justifica el uso de este criterio como un buen método de selección de λ demostrando que si $n^{-1}\text{tr}[S_\lambda] < 1$, entonces la diferencia entre $E[GCV(\lambda)]$ y $P(\lambda)$ relativa al tamaño de $R(\lambda)$ será pequeña, especialmente para tamaños de muestra grande.

¿Generalizada?

- Aunque la palabra “generalizada” deja la impresión de que el segundo criterio generaliza el primero, esto no es en general cierto y se trata de criterios diferentes que permiten estimar el riesgo de predicción $P(\lambda)$.
- Wahba (1990) justifica el uso de este criterio como un buen método de selección de λ demostrando que si $n^{-1}\text{tr}[S_\lambda] < 1$, entonces la diferencia entre $E[GCV(\lambda)]$ y $P(\lambda)$ relativa al tamaño de $R(\lambda)$ será pequeña, especialmente para tamaños de muestra grande.

Contenido

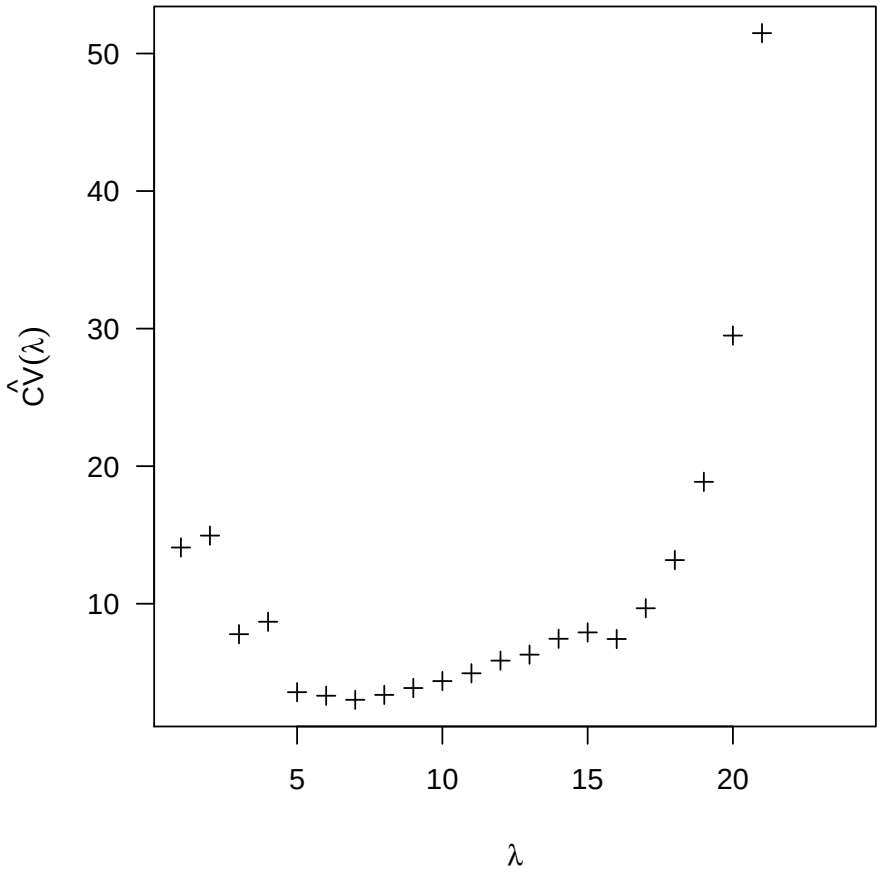
Los criterios $CV(\lambda)$ y $GCV(\lambda)$

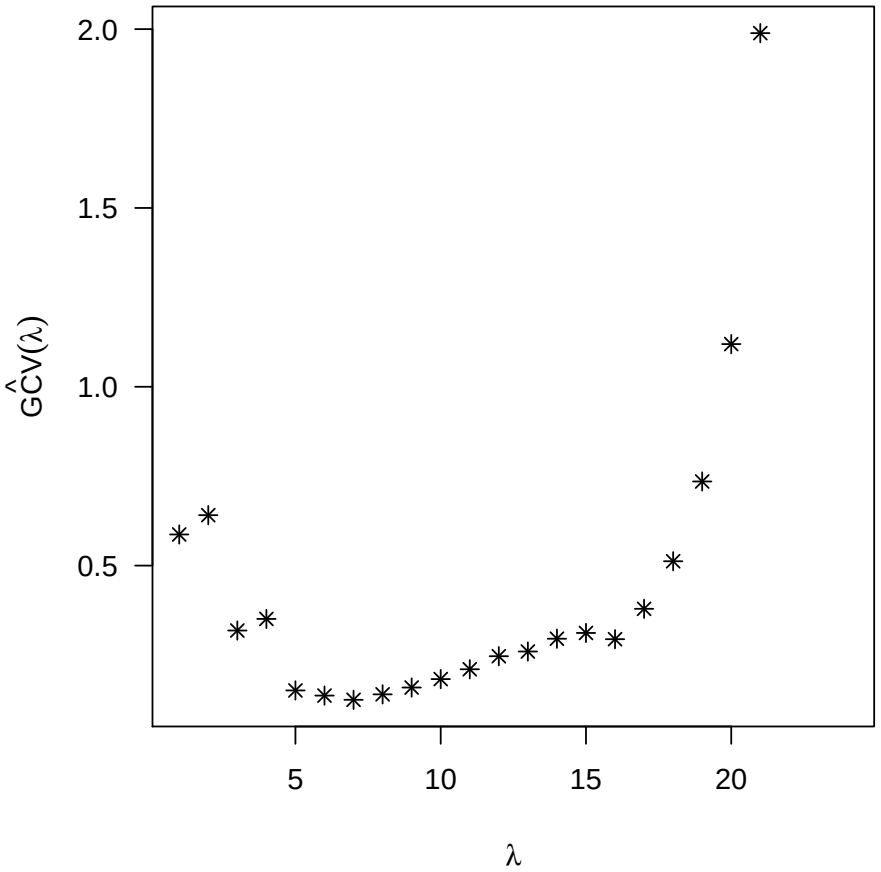
- Los criterios $CV(\lambda)$ y $GCV(\lambda)$ son estimadores del riesgo de predicción.
- Veamos cómo operan estos criterios en nuestro ejemplo de la contaminación del aire debida al CO en la Calle 15 de Cali.

Los criterios $CV(\lambda)$ y $GCV(\lambda)$

- Los criterios $CV(\lambda)$ y $GCV(\lambda)$ son estimadores del riesgo de predicción.
- Veamos cómo operan estos criterios en nuestro ejemplo de la contaminación del aire debida al CO en la Calle 15 de Cali.

Selección de λ con los estimadores $\hat{CV}(\lambda)$ y $\hat{GCV}(\lambda)$





Los criterios $CV(\lambda)$ y $GCV(\lambda)$

- Ambos criterios conducen a elegir el valor $\lambda = 7$ como el más adecuado para estimar μ en el ejemplo de la estimación de la función de regresión con los datos de contaminación del aire por *CO* en la calle 15 de Cali, cuando se usa una serie de cosenos.
- Este es el mismo valor que encontramos para λ usando el estimador *UBRE*

Los criterios $CV(\lambda)$ y $GCV(\lambda)$

- Ambos criterios conducen a elegir el valor $\lambda = 7$ como el más adecuado para estimar μ en el ejemplo de la estimación de la función de regresión con los datos de contaminación del aire por *CO* en la calle 15 de Cali, cuando se usa una serie de cosenos.
- Este es el mismo valor que encontramos para λ usando el estimador *UBRE*

Contenido

Actividad 1: 60 vinos

- Encuentre, con los criterios CV y GCV, el valor óptimo de λ en su estimación de la función de regresión de acidez fija a partir del ph, usando la muestra de 60 vinos que eligió en la Tarea 3.
- Compare estas selecciones de λ entre ellas y con la selección del *UBRE*.

Actividad 1: 60 vinos

- Encuentre, con los criterios CV y GCV, el valor óptimo de λ en su estimación de la función de regresión de acidez fija a partir del ph, usando la muestra de 60 vinos que eligió en la Tarea 3.
- Compare estas selecciones de λ entre ellas y con la selección del *UBRE*.

Actividad 2: Ozono

- Considere la base de datos *OzonoCompartir2019* (ver Classroom del curso). Esta información es del año 2019, de la estación Compartir (agradecimientos al DAGMA)
- Seleccione cinco días consecutivos (de lunes a viernes) y construya la curva de regresión para cada día, usando seis funciones de la base de cosenos
- Asuma que cada curva es la observación del día correspondiente. Así que ahora usted tiene cinco datos, uno para cada día. A este tipo de datos se les llama *Datos Funcionales* (Ramsay y Silverman 2005, Ferraty y Vieu 2006, Ramsay, Spencer y Hooker 2010, Ramsay, Wickham, Graves y Hooker 2011)
- Represente sus cinco datos funcionales en un solo gráfico. Luego, calcule e interprete la media y la desviación estándar funcionales de estos cinco datos. Represente esta curvas en el mismo gráfico

Actividad 2: Ozono

- Considere la base de datos *OzonoCompartir2019* (ver Classroom del curso). Esta información es del año 2019, de la estación Compartir (agradecimientos al DAGMA)
- Seleccione cinco días consecutivos (de lunes a viernes) y construya la curva de regresión para cada día, usando seis funciones de la base de cosenos
- Asuma que cada curva es la observación del día correspondiente. Así que ahora usted tiene cinco datos, uno para cada día. A este tipo de datos se les llama *Datos Funcionales* (Ramsay y Silverman 2005, Ferraty y Vieu 2006, Ramsay, Spencer y Hooker 2010, Ramsay, Wickham, Graves y Hooker 2011)
- Represente sus cinco datos funcionales en un solo gráfico. Luego, calcule e interprete la media y la desviación estándar funcionales de estos cinco datos. Represente esta curvas en el mismo gráfico

Actividad 2: Ozono

- Considere la base de datos *OzonoCompartir2019* (ver Classroom del curso). Esta información es del año 2019, de la estación Compartir (agradecimientos al DAGMA)
- Seleccione cinco días consecutivos (de lunes a viernes) y construya la curva de regresión para cada día, usando seis funciones de la base de cosenos
- Asuma que cada curva es la observación del día correspondiente. Así que ahora usted tiene cinco datos, uno para cada día. A este tipo de datos se les llama *Datos Funcionales* (Ramsay y Silverman 2005, Ferraty y Vieu 2006, Ramsay, Spencer y Hooker 2010, Ramsay, Wickham, Graves y Hooker 2011)
- Represente sus cinco datos funcionales en un solo gráfico. Luego, calcule e interprete la media y la desviación estándar funcionales de estos cinco datos. Represente esta curvas en el mismo gráfico

Actividad 2: Ozono

- Considere la base de datos *OzonoCompartir2019* (ver Classroom del curso). Esta información es del año 2019, de la estación Compartir (agradecimientos al DAGMA)
- Seleccione cinco días consecutivos (de lunes a viernes) y construya la curva de regresión para cada día, usando seis funciones de la base de cosenos
- Asuma que cada curva es la observación del día correspondiente. Así que ahora usted tiene cinco datos, uno para cada día. A este tipo de datos se les llama *Datos Funcionales* (Ramsay y Silverman 2005, Ferraty y Vieu 2006, Ramsay, Spencer y Hooker 2010, Ramsay, Wickham, Graves y Hooker 2011)
- Represente sus cinco datos funcionales en un solo gráfico. Luego, calcule e interprete la media y la desviación estándar funcionales de estos cinco datos. Represente esta curvas en el mismo gráfico

Contenido

Penalización

- Para estimar μ usando series de Fourier, hemos utilizado la suma

$$\sum_{j=1}^{\lambda} \beta_j f_j(x)$$

- Una posibilidad de mejoramiento de esta estimación consiste en *penalizar* la suma anterior, con alguna medida del grado de *rugosidad* de la curva
- Por ejemplo, si se usa la segunda derivada de μ como medida de rugosidad, el estimador sería

$$\sum_{j=1}^{\lambda} \beta_j f_j(x) + \delta \int_0^1 [\mu''(x)]^2 dx$$

- Por lo que ahora debemos elegir dos medidas óptimas: λ y δ

Penalización

- Para estimar μ usando series de Fourier, hemos utilizado la suma

$$\sum_{j=1}^{\lambda} \beta_j f_j(x)$$

- Una posibilidad de mejoramiento de esta estimación consiste en *penalizar* la suma anterior, con alguna medida del grado de *rugosidad* de la curva
- Por ejemplo, si se usa la segunda derivada de μ como medida de rugosidad, el estimador sería

$$\sum_{j=1}^{\lambda} \beta_j f_j(x) + \delta \int_0^1 [\mu''(x)]^2 dx$$

- Por lo que ahora debemos elegir dos medidas óptimas: λ y δ

Penalización

- Para estimar μ usando series de Fourier, hemos utilizado la suma

$$\sum_{j=1}^{\lambda} \beta_j f_j(x)$$

- Una posibilidad de mejoramiento de esta estimación consiste en *penalizar* la suma anterior, con alguna medida del grado de *rugosidad* de la curva
- Por ejemplo, si se usa la segunda derivada de μ como medida de rugosidad, el estimador sería

$$\sum_{j=1}^{\lambda} \beta_j f_j(x) + \delta \int_0^1 [\mu''(x)]^2 dx$$

- Por lo que ahora debemos elegir dos medidas óptimas: λ y δ

Penalización

- Para estimar μ usando series de Fourier, hemos utilizado la suma

$$\sum_{j=1}^{\lambda} \beta_j f_j(x)$$

- Una posibilidad de mejoramiento de esta estimación consiste en *penalizar* la suma anterior, con alguna medida del grado de *rugosidad* de la curva
- Por ejemplo, si se usa la segunda derivada de μ como medida de rugosidad, el estimador sería

$$\sum_{j=1}^{\lambda} \beta_j f_j(x) + \delta \int_0^1 [\mu''(x)]^2 dx$$

- Por lo que ahora debemos elegir dos medidas óptimas: λ y δ

Contenido

Bibliografía

Eubank (1999), *Nonparametric Regression and Spline Smoothing*, second edn, Marcel Dekker, New York, NY.

Ferraty y Vieu (2006), *Nonparametric Functional Data Analysis Theory and Practice*, Springer.

Green y Silverman (2000), *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*, Chapman & Hall/CRC, Boca Raton, FL.

Ramsay, Spencer y Hooker (2010), *Functional Data Analysis with **R** and **MATLAB***, Springer.

Ramsay, Wickham, Graves y Hooker (2011), *fda: Functional Data Analysis*. R package version 2.2.6.

Ramsay y Silverman (2005), *Functional Data Analysis*, 2nd. edn, Springer.

Wahba (1990), *Spline Models for Observational data*, CBMS-NSF Series, SIAM.