UNIVERSITÉ TOULOUSE III PAUL SABATIER

Université Fédérale Toulouse Midi-Pyrénées

Introducción a la Selección de Modelos

Ciudad de México - Diciembre 2017

Xavier Gendre

Introducción a la Selección de Modelos

Ciudad de México - Diciembre 2017

Xavier Gendre

Índice general

Prefacio

Estas notas son relacionadas con el curso “Introducción a la Selección de Modelos” dado en la Universidad Nacional Autónoma de México los días 13 y 14 de diciembre de 2017 como parte de la escuela temática *Data Analysis*. Este evento fue organizado con el apoyo del Instituto de Matemáticas de la UNAM y de la Embajada de Francia en México. Agradezco a Eric Bonnetier, Annie Marchegay, Itzel Lara, Gerónimo Uribe Bravo y Jean-Joinville Vacher por toda la asistencia prestada.

El curso fue impartido por Xavier Gendre del Institut de Mathématiques de Toulouse. Estas notas y los datos para la parte práctica son disponibles en la página web del autor.

Para cualquier solicitud o comentario, contáctese con el autor a xavier.gendre@math.univ-toulouse.fr.

1 — Selección de modelos: el caso lineal

1.1. Motivaciones

Los problemas considerados en estadística inferencial pueden generalmente formularse en función de la distribución desconocida de una variable aleatoria ξ. Más concretamente, nos interesamos por un objeto $s \in \mathscr{S}$ vinculado con esta distribución. Los objetivos pueden ser diversos como la estimación de s, la construcción de una región de confianza para s, ... Por eso, tenemos observaciones de ξ que llamamos los *datos* y nuestros procedimientos estadísticos pueden sólo apoyarse en estos datos. En particular, no podemos utilizar la distribución desconocida de ξ para ese propósito. Aquí hay algunos ejemplos de marcos estadísticos clásicos:

- *regresión funcional*: teniendo un espacio $\mathscr{X}$ y n pares de variables aleatorias $\xi_1 = (X_1, Y_1), \ldots, \xi_n = (X_n, Y_n) \in \mathscr{X} \times \mathbb{R}$ con la misma distribución que un par (X, Y), consideramos los datos $\xi = (\xi_1, \ldots, \xi_n)$ y la *función de regresión* desconocida $s : \mathscr{X} \to \mathbb{R}$ dada por, para todo $x \in \mathscr{X}$,
$$s(x) = \mathbb{E}[Y \mid X = x].$$
En este marco, $\mathscr{S}$ es el espacio de las funciones de $\mathscr{X}$ en $\mathbb{R}$ y se puede poner los datos en la forma siguiente
$$Y_i = s(X_i) + \varepsilon_i,\ i \in \{1, \ldots, n\},$$
donde las variables $\varepsilon_i = Y_i - \mathbb{E}[Y_i \mid X_i]$ son centradas.

- *aprendizaje estadístico*: teniendo un espacio $\mathscr{X}$ y n pares de variables aleatorias $\xi_1 = (X_1, Y_1), \ldots, \xi_n = (X_n, Y_n) \in \mathscr{X} \times \{0, 1\}$ con la misma distribución que un par (X, Y), consideramos los datos $\xi = (\xi_1, \ldots, \xi_n)$ y el *clasificador bayesiano* $s : \mathscr{X} \to \{0, 1\}$ dado por, para todo $x \in \mathscr{X}$,
$$s(x) = \begin{cases} 1 & \text{si } \eta(x) \geqslant 1/2, \\ 0 & \text{en caso contrario,} \end{cases}$$
donde $\eta(x) = \mathbb{E}[Y \mid X = x]$. El espacio $\mathscr{S}$ es el de todos los clasificadores binarios sobre $\mathscr{X}$.

- *densidad de probabilidad*: si los datos $\xi = (\xi_1, \ldots, \xi_n) \in E^n$ son n observaciones independientes con la misma distribución p absolutamente continua con respecto a una medida de probabilidad μ sobre el espacio medible $(E, \mathscr{E})$, podemos considerar la función medible $s : E \to \mathbb{R}$ dada por la derivada de Radon-Nykodym

$$s = \frac{dp}{d\mu}.$$

  Aquí, el espacio $\mathscr{S}$ es el de las densidades de probabilidad sobre $(E, \mathscr{E}, \mu)$.

Las herramientas estadísticas desarrolladas más adelante en este curso se pueden adaptar a estos diferentes marcos estadísticos. Sin embargo, en lo que sigue, desarrollaremos principalmente el marco de la regresión estadística.

Sin suposiciones adicionales, el espacio $\mathscr{S}$ en el que se encuentra el objeto de interés s suele ser muy grande, o incluso de dimensión infinita. En la práctica, es común (o necesario) de tener hipótesis sobre s (regularidad de una función, estructura geométrica, ...) o restricciones externas (dimensión finita, clase de distribuciones, ...) que pueden restringir el campo de posibilidades. Formalmente, esto puede ser posible mediante un subespacio $S \subset \mathscr{S}$ que llamamos un *modelo*. El siguiente paso es generalmente estimar s en el modelo S para desplegar nuestros procedimientos estadísticos. Antes de desarrollar este punto, debe entenderse que la elección de un modelo S no deja de tener consecuencias. En efecto, un método estadístico puede tener buenas propiedades teóricas en un modelo particular, pero adolecer de una representación pobre de s. También es posible que las suposiciones hechas sobre s sean discutibles o difíciles de verificar en la práctica. Para evitar estas dificultades, podemos considerar varios modelos al mismo tiempo e intentar elegir uno que sea lo "mejor" posible. El objetivo de la selección de modelos es proponer procedimientos estadísticos para hacer tales elecciones.

Para estimar s, consideramos un *criterio empírico* $\gamma_n : \mathscr{S} \to \mathbb{R}$ que se calcula solo a partir de los datos tal que la función

$$t \in \mathscr{S} \mapsto \gamma(t) = \mathbb{E}[\gamma_n(t)]$$

es mínima en s. Dado un modelo $S \subset \mathscr{S}$, se puede considerar un estimador $\hat{s} \in S$ como cualquier minimizador de γ_n en S,

$$\hat{s} \in \operatorname*{argmin}_{t \in S} \gamma_n(t).$$

La idea detrás de este tipo de estimador muy clásico es que, al minimizar γ_n, esperamos obtener un elemento cercano a s, al menos cuando s pertenece a S. Para medir la calidad de la representación de s por $t \in \mathscr{S}$, trabajamos con la *función de pérdida* asociada con γ_n,

$$\ell(s,t) = \gamma(t) - \gamma(s)$$

que es positiva, por definición. Aquí hay algunos ejemplos de criterios empíricos con sus funciones de pérdida:

- *regresión funcional*: podemos usar el *criterio de mínimos cuadrados*, para todo $t \in \mathscr{S}$,

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - t(X_i))^2.$$

Obtenemos directamente que

$$\gamma(t) = \mathbb{E}\left[(s(X) - t(X))^2\right] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\varepsilon_i^2\right]$$

es mínimo en s y que la función de pérdida es dada por la norma habitual de $L^2(X)$,

$$\ell(s,t) = \mathbb{E}\left[(s(X) - t(X))^2\right].$$

- *aprendizaje estadístico*: un criterio básico es dado por la *tasa de clasificación errónea empírica*, para todo clasificador $t \in \mathscr{S}$,

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq t(X_i)}.$$

La función $t \mapsto \gamma(t) = \mathbb{P}(Y \neq t(X))$ es mínimo en s y la función de pérdida es el exceso de riesgo

$$\ell(s,t) = \mathbb{P}(Y \neq t(X)) - \mathbb{P}(Y \neq s(X)).$$

- *densidad de probabilidad*: consideramos el *criterio de máxima verosimilitud*, para toda densidad $t \in \mathscr{S}$,

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(t(\xi_i)\right).$$

Sabemos que

$$\gamma(t) = -\int_E \log\left(t(x)\right) s(x) d\mu(x)$$

y, por lo tanto, la función de pérdida es dada por la *divergencia de Kullback-Leibler*,

$$\ell(s,t) = \int \log\left(\frac{s(x)}{t(x)}\right) s(x) d\mu(x).$$

Dada una colección contable de modelos $\{S_m\}_{m\in\mathscr{M}}$, tenemos minimizadores $\hat{s}_m \in S_m$ de γ_n en cada modelo. El enfoque general de la selección de modelos que proponemos desarrollar en el resto de este curso se basa en la minimización de un criterio penalizado para elegir un índice $\hat{m} \in \mathscr{M}$ como

$$\hat{m} \in \operatorname*{argmin}_{m\in\mathscr{M}} \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\}$$

donde $\mathrm{pen} : \mathscr{M} \to \mathbb{R}_+$ es una *función de penalización* determinista que se precisará ulteriormente. La relevancia de este enfoque se discutirá más adelante, pero parece importante explicar aquí el principio. Tenemos en mente la minimización del riesgo $\mathbb{E}[\ell(s,\hat{s}_m)]$ con respecto a $m \in \mathscr{M}$. Si notamos el estimador seleccionado $\tilde{s} = \hat{s}_{\hat{m}}$, podemos escribir la siguiente desigualdad por definición de $\hat{m}$, para cualquier $m \in \mathscr{M}$,

$$\begin{aligned}
\gamma(\tilde{s}) &= \gamma_n(\tilde{s}) + [\gamma(\tilde{s}) - \gamma_n(\tilde{s})] \\
&\leqslant \gamma_n(\hat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + [\gamma(\tilde{s}) - \gamma_n(\tilde{s})] \\
&\leqslant \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\} + \sum_{m'\in\mathscr{M}} \left([\gamma(\hat{s}_{m'}) - \gamma_n(\hat{s}_{m'})] - \mathrm{pen}(m')\right)_+
\end{aligned}$$

donde $x_+ = \max\{0,x\}$ es la parte positiva de $x \in \mathbb{R}$. Entonces, deducimos el siguiente límite superior del riesgo de $\tilde{s}$,

$$\mathbb{E}\left[\ell(s,\tilde{s})\right] \leqslant \inf_{m\in\mathscr{M}} \left\{\mathbb{E}\left[\gamma_n(\hat{s}_m) - \gamma(s)\right] + \mathrm{pen}(m)\right\} + \sum_{m'\in\mathscr{M}} \mathbb{E}\left[\left(\left[\gamma(\hat{s}_{m'}) - \gamma_n(\hat{s}_{m'})\right] - \mathrm{pen}(m')\right)_+\right].$$

Para obtener un límite superior del mismo orden que el riesgo mínimo $\inf_{m\in\mathscr{M}} \mathbb{E}[\ell(s,\hat{s}_m)]$, vemos que debemos considerar una penalización que

- no sea demasiado grande en comparación con el riesgo $\mathbb{E}[\ell(s,\hat{s}_m)]$,
- no sea demasiado pequeña para mantener insignificante la suma en $m' \in \mathscr{M}$.

El corazón de los métodos de selección de modelos consistirá precisamente en encontrar tales compromisos.

1.2. Modelo de regresión lineal

En esta parte, consideramos un caso especial del marco estadístico de la regresión funcional presentado en la sección anterior en el que se supone que la función de regresión es lineal. Entonces, dado enteros $n > 0$ y $p \geqslant 0$, observamos una muestra aleatoria $(X_1,Y_1),\dots,(X_n,Y_n) \in \mathbb{R}^p \times \mathbb{R}$ y notamos $X_i = (X_i^1,\dots,X_i^p)^\top \in \mathbb{R}^p$ el i-ésimo vector de las observaciones de p variables reales. La relación buscada entre las observaciones Y_i y las variables X_i se fórmula como

$$Y_i = \alpha_0 + \alpha_1 X_i^1 + \cdots + \alpha_p X_i^p + \varepsilon_i$$

donde las componentes del vector $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^\top \in \mathbb{R}^n$ representan errores en la relación lineal. Se supone que estos errores son tales que $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ y que existe $\sigma^2 > 0$ tal que, para cualquier $i,j \in \{1,\dots,n\}$,

$$\mathbb{E}[\varepsilon_i\varepsilon_j \mid X_1,\dots,X_n] = \begin{cases} \sigma^2 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases} \tag{1.1}$$

Los $p+1$ coeficientes $\alpha_0,\dots,\alpha_p \in \mathbb{R}$ son desconocidos y el objeto de interés es el vector $s = (s_1,\dots,s_n) \in \mathbb{R}^n$ definido por $s_i = \mathbb{E}[Y_i \mid X_i]$ para cualquier $i \in \{1,\dots,n\}$.

Para estimar el vector de los coeficientes $\alpha = (\alpha_0,\dots,\alpha_p) \in \mathbb{R}^{p+1}$, minimizamos el criterio de mínimos cuadrados

$$\gamma_n(\alpha) = \sum_{i=1}^n \left(Y_i - \alpha_0 - \alpha_1 X_i^1 - \cdots - \alpha_p X_i^p\right)^2.$$

Definiendo el vector $Y = (Y_1,\dots,Y_n)^\top \in \mathbb{R}^n$ y la matriz X de tamaño $n \times (p+1)$ dada por

$$X = \begin{pmatrix} 1 & X_1^1 & \dots & X_1^p \\ 1 & X_2^1 & \dots & X_2^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^1 & \dots & X_n^p \end{pmatrix},$$

el criterio γ_n puede formularse así

$$\gamma_n(\alpha) = \|Y - X\alpha\|^2$$

donde $\|x\|^2 = x_1^2 + \cdots + x_n^2$ para cualquier $x = (x_1,\dots,x_n)^\top \in \mathbb{R}^n$.

Proposición 1.1. *Si el rango de la matriz X es $p+1$, entonces el único minimizador de $\alpha \in \mathbb{R}^{p+1} \mapsto \gamma_n(\alpha)$ es*

$$\hat{\alpha} = \left(X^\top X\right)^{-1} X^\top Y.$$

El estimador $\hat{\alpha}$ es insesgado para estimar α. Además, los valores de la función de regresión asociada con $\hat{\alpha}$ son dadas por

$$\hat{s} = X\hat{\alpha} = HY$$

donde $H = X\left(X^\top X\right)^{-1} X^\top$ es la proyección ortogonal en el espacio generado por las columnas de X.

Demostración. La función $\alpha \mapsto \gamma_n(\alpha)$ es cuadrática con respecto a $\alpha_0, \ldots, \alpha_p$. Por lo tanto, la diferenciamos fácilmente para tener

$$\frac{\partial \gamma_n}{\partial \alpha}(\alpha) = -2X^\top (Y - X\alpha) \quad \text{y} \quad \frac{\partial^2 \gamma_n}{\partial \alpha \partial \alpha^\top}(\alpha) = 2X^\top X.$$

Ya que la matriz X tiene un rango columna máximo, la matriz $X^\top X$ es definida positiva y se obtiene el único minimizador de γ_n tomando el cero de la primera derivada, $X^\top (Y - X\alpha) = 0$, a saber

$$\hat{\alpha} = \left(X^\top X\right)^{-1} X^\top Y.$$

El estimador $\hat{\alpha}$ de α es sin sesgo porque

$$\mathbb{E}[\hat{\alpha}] = \mathbb{E}\left[\left(X^\top X\right)^{-1} X^\top Y\right] = \mathbb{E}\left[\left(X^\top X\right)^{-1} X^\top X\alpha\right] + \mathbb{E}\left[\left(X^\top X\right)^{-1} X^\top \varepsilon\right] = \alpha.$$

Es obvio que H es una proyección en el espacio generado por las columnas de X,

$$H^2 = X\left(X^\top X\right)^{-1} X^\top X\left(X^\top X\right)^{-1} X^\top = X\left(X^\top X\right)^{-1} X^\top = H.$$

Además, la ortogonalidad de H proviene de su simetría. □

Suponiendo que el rango de X es $p+1$ y utilizando el resultado anterior, tenemos a nuestra disposición el estimador $\hat{s} = HY$ de s en el modelo lineal generado por las columnas de X. Escribiendo $Y = s + \varepsilon$, notemos que la esperanza $\mathbb{E}[HY \mid X] = Hs$ es la proyección de s sobre este modelo lineal. La calidad de $\hat{s}$ se mide por su riesgo, a saber la esperanza de la función de pérdida asociada con el criterio de mínimos cuadrados,

$$\begin{aligned}\mathbb{E}[\ell(s,\hat{s})] &= \mathbb{E}\left[\|s - HY\|^2\right] \\ &= \mathbb{E}\left[\|s - Hs\|^2\right] + \mathbb{E}\left[\|H\varepsilon\|^2\right].\end{aligned}$$

Esta descomposición contiene una idea importante para lo que sigue. De un lado, el *término de sesgo* $\mathbb{E}\left[\|s - Hs\|^2\right]$ cuantifica la capacidad del modelo de acercarse al vector s. En el otro lado, el *término de varianza* $\mathbb{E}\left[\|H\varepsilon\|^2\right]$ mide la complejidad del modelo en el sentido de su

dimensión. En efecto, las hipótesis sobre ε conducen a

$$\begin{aligned}\mathbb{E}\left[\|H\varepsilon\|^2\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{j=1}^{n}H_{ij}\varepsilon_j\right)^2\right]\\ &= \sigma^2\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{n}H_{ij}^2\right]\\ &= \sigma^2\mathrm{tr}(H^2)\\ &= \sigma^2\mathrm{tr}(H)\\ &= \sigma^2(p+1)\end{aligned}$$

Entonces, el riesgo del estimador $\hat{s}$ de s en el modelo lineal generado por las columnas de X se fórmula como la suma

$$\mathbb{E}\left[\|s-Hs\|^2\right]+\sigma^2(p+1).$$

Para tener un "buen" modelo lineal en el sentido de un riesgo bajo, vemos que debemos encontrar un compromiso entre la capacidad de acercarse s y la dimensión del modelo.

1.3. Selección de modelos lineales

En la sección anterior, vimos cómo estimar un vector $s=\mathbb{E}[Y]\in\mathbb{R}^n$ en un modelo lineal a partir de la observación de datos $Y\in\mathbb{R}^n$. Esta estimación consiste en proyectar Y sobre el modelo. Entonces, considerar varios modelos lineales equivale a considerar una colección de proyecciones $\{H_m\}_{m\in\mathscr{M}}$. Claro, queremos elegir un modelo lineal a través de la selección de una proyección $H_{\hat{m}}$ con $\hat{m}\in\mathscr{M}$. Más adelante, el conjunto $\mathscr{M}$ siempre será contable para evitar problemas de medibilidad y, para cualquier $m\in\mathscr{M}$, notaremos la dimensión del modelo $\mathrm{tr}(H_m)=p_m+1$ y el estimador $\hat{s}_m=H_mY$.

No vamos a tener un resultado teórico en este capítulo (este será el tema del segundo capítulo), sino solo presentaremos una heurística para motivar el enfoque de los criterios penalizados. Esta heurística se debe a Mallows en los años 70, pero otros estadísticos de la misma época también desarrollaron trabajos similares (Akaike, ...). Para cualquier $m\in\mathscr{M}$, tenemos el riesgo siguiente

$$\mathbb{E}\left[\|s-H_ms\|^2\right]+\sigma^2(p_m+1)=\mathbb{E}\left[\|s\|^2\right]-\mathbb{E}\left[\|H_ms\|^2\right]+\sigma^2(p_m+1).$$

Entonces, queremos minimizar la cantidad

$$-\mathbb{E}\left[\|H_ms\|^2\right]+\sigma^2(p_m+1)$$

pero $\mathbb{E}\left[\|H_ms\|^2\right]$ es desconocida. La idea es reemplazarlo con un estimador sin sesgo. Se calcula fácilmente que

$$\mathbb{E}\left[\|H_mY\|^2\right]=\mathbb{E}\left[\|H_ms\|^2\right]+\sigma^2(p_m+1).$$

Por lo tanto, $\|H_mY\|^2-\sigma^2(p_m+1)$ es un estimador sin sesgo de $\mathbb{E}\left[\|H_ms\|^2\right]$ y tomamos $\hat{m}\in\mathscr{M}$ como cualquier minimizador del criterio siguiente

$$-\|H_mY\|^2+2\sigma^2(p_m+1)$$

o, de manera equivalente, del criterio de mínimos cuadrados penalizado por 2 veces el término de varianza,

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathscr{M}} \left\{ \|Y - H_m Y\|^2 + 2\sigma^2 (p_m + 1) \right\}.$$

Más generalmente, el principio es imitar el compromiso entre los términos de sesgo y de varianza por un compromiso entre el criterio de mínimos cuadrados y una penalización proporcional a la dimensión del modelo,

$$\text{pen}(m) = \lambda \sigma^2 (p_m + 1)$$

donde $\lambda \geqslant 0$ es un parámetro de ajuste. Si λ tiende a cero, encontramos el criterio de mínimos cuadrados clásico, es decir que solo tenemos en cuenta la adecuación a los datos. Para $\lambda = 2$, este es el criterio de Mallows y si λ tiende al infinito, solo la penalización cuenta y el modelo elegido es el de menor dimensión. Discutiremos este parámetro λ más detalladamente en el próximo capítulo, pero su papel es fundamental para el procedimiento y a menudo se usa un paso de *validación cruzada* para calibrarlo en la práctica.

Para acabar esta sección, notamos que la minimización de un criterio penalizado es un método muy generalizado en estadística y no es limitado a la selección de modelos. Se puede penalizar con otras cantidades dependiendo del compromiso buscado. Por ejemplo, para estimar la esperanza de Y en un modelo lineal generado por las $p+1$ columnas de una matriz X, la *regresión Ridge* utiliza la *regularización de Tíjonov* de los coeficientes de regresión $\alpha \in \mathbb{R}^{p+1}$ como una penalización para favorecer a las más bajas,

$$\gamma_{Ridge}(\alpha) = \|Y - X\alpha\|^2 + \lambda \|\alpha\|^2$$

con $\lambda \geqslant 0$. La idea de este criterio es de regularizar el estimador obtenido como un filtro paso bajo que elimina las variaciones rápidas para un operador de Fourier, por ejemplo.

1.4. Ejemplo: regresión polinomial

Para ilustrar el procedimiento descrito anteriormente, proponemos considerar el caso particular de la regresión polinomial con soporte fijo en $[0,1]$. Aquí, los datos son pares $(x_1, Y_1), \ldots, (x_n, Y_n) \in [0,1] \times \mathbb{R}$ con $x_i = i/n$, $i \in \{1, \ldots, n\}$. Se supone que las variables aleatorias $Y_1, \ldots, Y_n$ son independientes y de misma varianza $\sigma^2 > 0$. Dado un entero $p \geqslant 0$, buscamos el "mejor" polinomio de grado p para estimar $s = \mathbb{E}[Y]$ a partir de los x_i. Entonces, minimizamos el criterio de mínimos cuadrados siguiente con respecto a $\alpha = (\alpha_0, \ldots, \alpha_p)^\top \in \mathbb{R}^{p+1}$,

$$\gamma_n^{(p)}(\alpha) = \sum_{i=1}^{n} \left(Y_i - \alpha_0 - \alpha_1 x_i - \cdots - \alpha_p x_i^p \right)^2.$$

Definiendo la matriz $X^{(p)}$ de tamaño $n \times (p+1)$ por

$$X^{(p)} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^p \\ 1 & x_2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^p \end{pmatrix},$$

vemos que el minimizador de $\gamma_n^{(p)}$ lleva al estimador $\hat{s}_p = H_p Y$ donde H_p es la proyección ortogonal del vector $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ en el espacio generado por las columnas de $X^{(p)}$.

De acuerdo con los resultados obtenidos para los modelos lineales en la sección anterior, el riesgo de $\hat{s}_p$ es dado por

$$\mathbb{E}[\ell(s,\hat{s}_p)] = \|s - H_p s\|^2 + \sigma^2(p+1).$$

Este marco es interesante porque los modelos son anidados: si $p' \geqslant p \geqslant 0$, el modelo lineal de los polinomios de grado p es incluido en el modelo lineal de los polinomios de grado p'. Entonces, cuando p aumenta, el término de sesgo $\|s - H_p s\|^2$ disminuye y el de varianza $\sigma^2(p+1)$ aumenta. La pregunta natural es la de elegir un "buen" grado p para tener un riesgo bajo sin conocer s. Este es un ejemplo simple de búsqueda de un compromiso entre el sesgo y la varianza.

Si solo tomamos un grado que minimice $p \mapsto \gamma_n^{(p)}(\hat{s}_p)$, vamos a tener el grado $n-1$ máximo. Aunque este modelo tiene una buena capacidad de aproximación (*i.e.* un término de sesgo bajo), su riesgo es grande debido al término de varianza. El estimador asociado con un modelo de este tipo es demasiado cerca a los datos y hablamos de un fenómeno de *sobreajuste* (o *overfitting* en inglés). La Figura 1.1 ilustra el comportamiento del término de sesgo y el del riesgo con respecto al grado p. Para evitar el sobreajuste, el enfoque de la selección de modelos descrito anteriormente consiste en elegir un grado $\hat{p}$ como un minimizador del criterio de mínimos cuadrados penalizado,

$$\hat{p} \in \operatorname*{argmin}_{0 \leqslant p \leqslant n-1} \left\{ \gamma_n^{(p)}(\hat{s}_p) + \lambda \sigma^2(p+1) \right\}$$

para un cierto $\lambda > 0$.

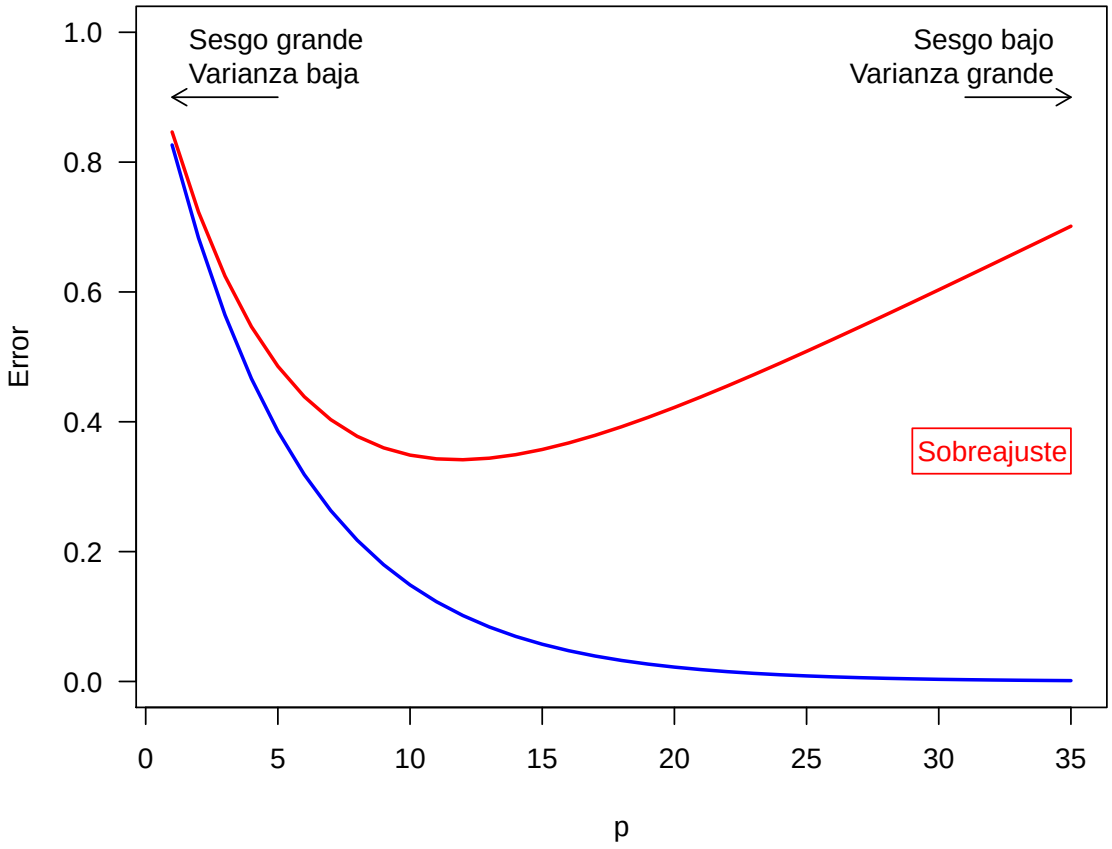

Figura 1.1: Término de sesgo (azul) y riesgo (roja) del estimador en el modelo lineal de los polinomios de grado p con respecto a p.

2 — Un teorema de selección de modelos

2.1. Aperitivo: desigualdades

En este capítulo, consideramos de nuevo el marco estadístico de la regresión para el cual los datos observados son un vector $Y \in \mathbb{R}^n$ que se puede escribir en la forma siguiente

$$Y = s + \varepsilon$$

donde $s \in \mathbb{R}^n$ es el vector a estimar y $\varepsilon \in \mathbb{R}^n$ es un vector aleatorio centrado que asumiremos gaussiano más tarde. Dada una colección contable de modelos lineales $\{S_m\}_{m \in \mathscr{M}}$, *i.e.* subespacios vectoriales de $\mathbb{R}^n$, tenemos a nuestra disposición la colección de proyectores asociados $\{H_m\}_{m \in \mathscr{M}}$ definida como en el primer capítulo. Por cada $m \in \mathscr{M}$, la dimensión del modelo S_m es $D_m = \mathrm{tr}(H_m)$ y notamos el estimador $\hat{s}_m = H_m Y$. Elegimos un modelo $S_{\hat{m}}$ tal que

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathscr{M}} \{\gamma_n(\hat{s}_m) + \mathrm{pen}(m)\}$$

donde $\gamma_n(\hat{s}_m) = \|Y - \hat{s}_m\|^2$ es el criterio de mínimos cuadrados y $\mathrm{pen} : \mathscr{M} \to \mathbb{R}_+$ es una función de penalización determinista. Hemos visto que tenemos la límite superior siguiente para el riesgo del estimador $\tilde{s} = \hat{s}_{\hat{m}}$,

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leqslant \inf_{m \in \mathscr{M}} \{\mathbb{E}\left[\gamma_n(\hat{s}_m) - \gamma(s)\right] + \mathrm{pen}(m)\} + R$$

donde el término restante R es dado por

$$R = \sum_{m \in \mathscr{M}} \mathbb{E}\left[(\gamma(\hat{s}_m) - \gamma_n(\hat{s}_m) - \mathrm{pen}(m))_+\right]$$

con $\gamma(t) = \mathbb{E}[\gamma_n(t)]$. Para controlar este término restante, vemos que necesitamos entender cómo se comporta el criterio empírico $\gamma_n(\hat{s}_m)$ alrededor de $\gamma(\hat{s}_m)$. Herramientas matemáticas que pueden ayudarnos en esta dirección son las desigualdades de concentración y las de desviación.

Si Z es una variable aleatoria real, una *desigualdad de concentración* es una desigualdad como, para cualquier $t > 0$,

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geqslant t\right) \leqslant \rho(t)$$

donde ρ es una función decreciente a cero a definir. Una desigualdad similar sin el valor absoluto se llama *desigualdad de desviación*,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geqslant t) \leqslant \rho(t).$$

Un método genérico para establecer tales desigualdades es usar la *transformada de Laplace* para tener *cotas de Chernoff*. Dado $\lambda > 0$, la desigualdad de Márkov nos da

$$\begin{aligned}\mathbb{P}(Z - \mathbb{E}[Z] \geqslant t) &= \mathbb{P}\left(e^{\lambda(Z-\mathbb{E}[Z])} \geqslant e^{\lambda t}\right) \\ &\leqslant e^{-\lambda t}\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \\ &= \exp\left(\log\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] - \lambda t\right).\end{aligned}$$

Entonces, obtenemos la *desigualdad de Chernoff*,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geqslant t) \leqslant \exp\left(-\sup_{\lambda>0}\left\{\lambda t - \log\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right]\right\}\right).$$

En el caso gaussiano estándar $Z \sim \mathcal{N}(0,1)$, para cualquier $\lambda > 0$, sabemos que

$$\mathbb{E}\left[e^{\lambda Z}\right] = \exp\left(\frac{\lambda^2}{2}\right).$$

Entonces, para cualquier $t > 0$, la desigualdad de Chernoff nos da

$$\mathbb{P}(Z \geqslant t) \leqslant \exp\left(-\sup_{\lambda>0}\left\{\lambda t - \frac{\lambda^2}{2}\right\}\right) = \exp\left(-\frac{t^2}{2}\right).$$

Por simetría, podemos deducir la desigualdad de concentración

$$\mathbb{P}(|Z| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2}\right).$$

Un otro caso que será útil en lo que sigue es $Z \sim \chi^2(D)$, *i.e.* Z tiene la misma distribución que la suma de D variables gaussianas estándares independientes al cuadrado. Se calcula fácilmente $\mathbb{E}[Z] = D$ y, para cualquier $\lambda \in (0, 1/2)$,

$$\mathbb{E}\left[e^{\lambda(Z-D)}\right] = (1-2\lambda)^{-D/2}e^{-\lambda D}.$$

Entonces, para cualquier $t > 0$, la desigualdad de Chernoff nos da la desigualdad de desviación siguiente

$$\mathbb{P}(Z - D \geqslant t) \leqslant \exp\left(-\sup_{0<\lambda<1/2}\left\{\lambda t + \frac{D}{2}\log(1-2\lambda) + \lambda D\right\}\right) = \left(\frac{D}{t+D}\right)^{-D/2} e^{-t/2}$$

donde se alcanza el supremo para $\lambda = \frac{t}{2(t+D)}$. En esta forma, esta desigualdad no siempre es fácil a manipular y se puede reescribirla con $t = 2\sqrt{Dx} + 2x$ donde $x > 0$ para tener

$$\mathbb{P}\left(Z - D \geqslant 2\sqrt{Dx} + 2x\right) \leqslant e^{-x} \times \exp\left(-\sqrt{Dx} - \frac{D}{2}\log\left(\frac{D}{D + 2x + 2\sqrt{Dx}}\right)\right).$$

Para todo $u = 2\sqrt{x/D} > 0$, sabemos que $u \geqslant \log(1+u+u^2/2)$, así

$$-\sqrt{Dx} - \frac{D}{2}\log\left(\frac{D}{D+2x+2\sqrt{Dx}}\right) = -\frac{D}{2}\left(u - \log\left(1+u+\frac{u^2}{2}\right)\right) \leqslant 0.$$

Entonces, se tiene la variante de la desigualdad de desviación anterior, para cualquier $x > 0$,

$$\mathbb{P}\left(Z - D \geqslant 2\sqrt{Dx} + 2x\right) \leqslant e^{-x}.$$

2.2. Entrada: un teorema de Birgé y Massart (2001)

Ahora podemos establecer un resultado teórico de selección de modelos en el marco de la regresión gaussiana. Observamos los datos

$$Y = s + \sigma\varepsilon$$

donde $s \in \mathbb{R}^n$ es el vector a estimar, $\sigma^2 > 0$ es el factor de varianza conocido y $\varepsilon \in \mathbb{R}^n$ es un vector gaussiano estándar. Para estimar s, consideramos los estimadores $\{\hat{s}_m\}_{m\in\mathscr{M}}$ dados por las proyecciones $\hat{s}_m = H_m Y$.

Teorema 2.1. *Dada una colección de reales positivos $\{x_m\}_{m\in\mathscr{M}}$ con*

$$\Sigma = \sum_{m\in\mathscr{M}} e^{-x_m} < \infty,$$

suponemos que existe $\kappa > 1$ tal que, para cualquier $m \in \mathscr{M}$,

$$pen(m) \geqslant \kappa\sigma^2\left(D_m + \frac{2\kappa^2 x_m}{(\kappa-1)^2}\right).$$

Si $\hat{m} \in \mathscr{M}$ es elegido por

$$\hat{m} \in \operatorname*{argmin}_{m\in\mathscr{M}} \left\{\|Y - \hat{s}_m\|^2 + pen(m)\right\}$$

entonces, el estimador $\tilde{s} = \hat{s}_{\hat{m}}$ verifica

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leqslant C \inf_{m\in\mathscr{M}} \left\{\|s - H_m s\|^2 - \sigma^2 D_m + pen(m)\right\} + C'\sigma^2\Sigma \tag{2.1}$$

donde $C > 1$ y $C' > 26$ solo dependen de κ.

Antes de probar este teorema, debemos hacer varias observaciones importantes. En el resultado original, Birgé y Massart hacen una hipótesis más débil sobre la penalización, a saber

$$\text{pen}(m) \geqslant \kappa\sigma^2\left(\sqrt{D_m} + \sqrt{x_m}\right)^2.$$

Esto no cambia mucho en el resultado, pero la prueba se vuelve un poco más complicada.

El papel de los pesos x_m es importante y permiten introducir un conocimiento a priori sobre los modelos en el procedimiento (opinión de especialistas, ...). En efecto, si un peso x_m es grande, el modelo S_m asociado será más penalizado y, por lo tanto, menos fácil a elegir. Pero este no es el único uso de estos pesos en el método. La condición de finitud en Σ es

relacionada con el número de modelos considerados, *i.e.* el tamaño de la colección de modelos. Si tenemos pocos modelos en competencia en el sentido de que existe $K \geqslant 0$ tal que, para cualquier entero $D \geqslant 0$,

$$\operatorname{card}\{m \in \mathscr{M} \text{ tal que } D_m = D\} \leqslant K$$

entonces, podemos tomar $x_m = LD_m$ para un cierto $L > 0$ (*i.e.* $\operatorname{pen}(m) = \kappa'\sigma^2 D_m$ con $\kappa' = \kappa(1 + 2\kappa^2 L/(\kappa-1)^2) > 1$) y tenemos

$$\Sigma = \sum_{m\in\mathscr{M}} e^{-LD_m} \leqslant K \sum_{D\geqslant 0} e^{-LD} = \frac{K}{1-e^{-L}} < \infty.$$

Así, la desigualdad (2.1) lleva a

$$\begin{aligned}\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] &\leqslant C \inf_{m\in\mathscr{M}} \left\{ \|s - H_m s\|_n^2 + \frac{(\kappa'-1)\sigma^2 D_m}{n} \right\} + \frac{C'\sigma^2 K}{n(1-e^{-L})} \\ &\leqslant C\kappa' \inf_{m\in\mathscr{M}} \mathbb{E}\left[\|s-\hat{s}_m\|_n^2\right] + \frac{C'\sigma^2 K}{n(1-e^{-L})}\end{aligned}$$

donde $\|\cdot\|_n^2 = \|\cdot\|^2/n$ es la norma normalizada en $\mathbb{R}^n$. Vemos que el riesgo del estimador $\tilde{s}$ es comparable al riesgo mínimo entre los estimadores $\hat{s}_m$ aparte de un término aditivo que tiende a cero con n. Tal resultado se llama *desigualdad de oráculo*.

2.3. Plato: prueba del teorema

En esta parte, notamos $\langle\cdot,\cdot\rangle$ el producto escalar en $\mathbb{R}^n$ asociado con la norma $\|\cdot\|$. Para cualquier $m \in \mathscr{M}$, tenemos

$$\|Y - \hat{s}_m\|^2 = \|s - \hat{s}_m\|^2 + 2\sigma\langle s - \hat{s}_m, \varepsilon\rangle + \sigma^2\|\varepsilon\|^2$$

y, por el teorema de Pitágoras,

$$\|s - \hat{s}_m\|^2 = \|s - H_m s\|^2 + \sigma^2\|H_m\varepsilon\|^2.$$

Entonces, ambas igualdades y la definición de $\hat{m}$,

$$\|Y - \tilde{s}\|^2 + \operatorname{pen}(\hat{m}) \leqslant \|Y - \hat{s}_m\|^2 + \operatorname{pen}(m),$$

llevan a

$$\begin{aligned}\|s-\tilde{s}\|^2 &\leqslant \|s - H_m s\|^2 + \sigma^2\|H_m\varepsilon\|^2 + 2\sigma\langle s - \hat{s}_m, \varepsilon\rangle + \operatorname{pen}(m) - 2\sigma\langle s - \tilde{s}, \varepsilon\rangle - \operatorname{pen}(\hat{m}) \\ &\leqslant \|s - H_m s\|^2 + Z_m - \sigma^2 D_m + \operatorname{pen}(m) - 2\sigma\langle s - H_{\hat{m}} s, \varepsilon\rangle + 2\sigma^2\|H_{\hat{m}}\varepsilon\|^2 - \operatorname{pen}(\hat{m})\end{aligned}$$

donde $Z_m = 2\sigma\langle s - H_m s, \varepsilon\rangle - \sigma^2\|H_m\varepsilon\|^2 + \sigma^2 D_m$ es una variable tal que $\mathbb{E}[Z_m] = 0$. Dado $\alpha \in (0,1)$, se obtiene

$$\begin{aligned}2\sigma|\langle s - H_{\hat{m}} s, \varepsilon\rangle| &= 2\sigma\|s - H_{\hat{m}} s\| \times |\langle u_{\hat{m}}, \varepsilon\rangle| \\ &\leqslant \alpha\|s - H_{\hat{m}} s\|^2 + \alpha^{-1}\sigma^2\langle u_{\hat{m}}, \varepsilon\rangle^2 \\ &= \alpha\|s-\tilde{s}\|^2 - \alpha\sigma^2\|H_{\hat{m}}\varepsilon\|^2 + \alpha^{-1}\sigma^2\langle u_{\hat{m}}, \varepsilon\rangle^2\end{aligned}$$

donde, para cualquier $m' \in \mathscr{M}$, $u_{m'} \in S_{m'}^{\perp}$ es tal que $\|u_{m'}\|^2 = 1$. Así, obtenemos

$$(1-\alpha)\|s-\tilde{s}\|^2 \leqslant \|s-H_m s\|^2 + Z_m - \sigma^2 D_m + \mathrm{pen}(m) \\ + \left(\alpha^{-1}\sigma^2\langle u_{\hat{m}}, \varepsilon\rangle^2 + (2-\alpha)\sigma^2\|H_{\hat{m}}\varepsilon\|^2 - \mathrm{pen}(\hat{m})\right)_+ .$$

Tomando la esperanza en ambos lados de esta desigualdad, tenemos

$$(1-\alpha)\mathbb{E}[\|s-\tilde{s}\|^2] \leqslant \inf_{m\in\mathscr{M}} \left\{\|s-H_m s\|^2 - \sigma^2 D_m + \mathrm{pen}(m)\right\} + R \tag{2.2}$$

donde el término restante R es dado por

$$R = \sum_{m\in\mathscr{M}} \mathbb{E}\left[\left(\alpha^{-1}\sigma^2\langle u_m, \varepsilon\rangle^2 + (2-\alpha)\sigma^2\|H_m\varepsilon\|^2 - \mathrm{pen}(m)\right)_+\right].$$

Para obtener el resultado anunciado, debemos controlar este término restante R. Por eso, lo dividimos en dos partes

$$R \leqslant \alpha^{-1}R_1 + (2-\alpha)R_2$$

dadas por

$$R_1 = \sum_{m\in\mathscr{M}} \mathbb{E}\left[\left(\sigma^2\langle u_m, \varepsilon\rangle^2 - p_1(m)\right)_+\right] \quad \text{y} \quad R_2 = \sum_{m\in\mathscr{M}} \mathbb{E}\left[\left(\sigma^2\|H_m\varepsilon\|^2 - p_2(m)\right)_+\right]$$

donde $p_1, p_2 : \mathscr{M} \to \mathbb{R}_+$ serán definidas más tarde y verifican

$$\mathrm{pen}(m) \geqslant \alpha^{-1}p_1(m) + (2-\alpha)p_2(m).$$

Empezamos con R_1, la variable $\langle u_m, \varepsilon\rangle$ es gaussiana estándar y la desigualdad de concentración presentada en la primera sección lleva a

$$\begin{aligned}
\mathbb{E}\left[\left(\sigma^2\langle u_m, \varepsilon\rangle^2 - p_1(m)\right)_+\right] &= \int_0^\infty \mathbb{P}\left(\sigma^2\langle u_m, \varepsilon\rangle^2 - p_1(m) \geqslant t\right) dt \\
&= \int_0^\infty \mathbb{P}\left(|\langle u_m, \varepsilon\rangle| \geqslant \sqrt{\frac{t+p_1(m)}{\sigma^2}}\right) dt \\
&\leqslant \int_0^\infty 2\exp\left(-\frac{t+p_1(m)}{2\sigma^2}\right) dt \\
&= 4\sigma^2\exp\left(-\frac{p_1(m)}{2\sigma^2}\right).
\end{aligned}$$

Entonces, tomando $p_1(m) = 2\sigma^2 x_m$, se obtiene

$$R_1 \leqslant 4\sigma^2\Sigma.$$

Procedemos de la misma manera para la segunda parte R_2. Observamos que $\|H_m\varepsilon\|^2 \sim \chi^2(D_m)$ por el teorema de Cochran. La desigualdad de desviación que obtuvimos en la primera sección dio, para todo $x > 0$,

$$\mathbb{P}\left(\|H_m\varepsilon\|^2 \geqslant D_m + 2\sqrt{D_m x} + 2x\right) \leqslant e^{-x}.$$

Para cualquier $\beta > 0$, sabemos que $2\sqrt{D_m x} \leqslant \beta D_m + \beta^{-1}x$ y podemos aflojar la desigualdad de desviación en

$$\mathbb{P}\left(\|H_m\varepsilon\|^2 \geqslant (1+\beta)D_m + (2+\beta^{-1})x\right) \leqslant e^{-x}.$$

Entonces, tomando $p_2(m) = \sigma^2((1+\beta)D_m + (2+\beta^{-1})x_m)$, se obtiene

$$\begin{aligned}
&\mathbb{E}\left[\left(\sigma^2\|H_m\varepsilon\|^2 - p_2(m)\right)_+\right] \\
&= \int_0^\infty \mathbb{P}\left(\sigma^2\|H_m\varepsilon\|^2 - p_2(m) \geqslant t\right) dt \\
&= \int_0^\infty \mathbb{P}\left(\|H_m\varepsilon\|^2 - (1+\beta)D_m - (2+\beta^{-1})x_m \geqslant \frac{t}{\sigma^2}\right) dt \\
&= (2+\beta^{-1})\sigma^2 \int_0^\infty \mathbb{P}\left(\|H_m\varepsilon\|^2 \geqslant (1+\beta)D_m + (2+\beta^{-1})(x_m+u)\right) du \\
&\leqslant (2+\beta^{-1})\sigma^2 e^{-x_m} \int_0^\infty e^{-u} du \\
&\leqslant (2+\beta^{-1})\sigma^2 e^{-x_m}.
\end{aligned}$$

Pues, tenemos

$$R_2 \leqslant (2+\beta^{-1})\sigma^2\Sigma$$

y deducimos que el término restante es tal que

$$R \leqslant \left(4\alpha^{-1} + (2-\alpha)(2+\beta^{-1})\right)\sigma^2\Sigma.$$

Para acabar esta prueba, debemos ahora dar los valores de α y β. La desigualdad (2.2) se convierte en

$$\mathbb{E}[\|s-\tilde{s}\|^2] \leqslant \frac{1}{1-\alpha} \inf_{m\in\mathscr{M}} \left\{\|s - H_m s\|^2 - \sigma^2 D_m + \text{pen}(m)\right\} + \frac{4\alpha^{-1} + (2-\alpha)(2+\beta^{-1})}{1-\alpha}\sigma^2\Sigma$$

y tenemos la límite inferior siguiente para la función de penalización,

$$\begin{aligned}
\text{pen}(m) &\geqslant \alpha^{-1}p_1(m) + (2-\alpha)p_2(m) \\
&= 2\alpha^{-1}\sigma^2 x_m + (2-\alpha)\sigma^2((1+\beta)D_m + (2+\beta^{-1})x_m) \\
&= \sigma^2\left[(2-\alpha)(1+\beta)D_m + \left(2\alpha^{-1} + (2-\alpha)(2+\beta^{-1})\right)x_m\right] \\
&= (2-\alpha)(1+\beta)\sigma^2\left[D_m + \frac{2\alpha^{-1} + (2-\alpha)(2+\beta^{-1})}{(2-\alpha)(1+\beta)}x_m\right].
\end{aligned}$$

Tomando $\alpha = \kappa^{-1}$ y $\beta = \frac{(\kappa-1)^2}{2\kappa-1}$, se obtiene $(2-\alpha)(1+\beta) = \kappa$ y

$$\begin{aligned}
\frac{2\alpha^{-1} + (2-\alpha)(2+\beta^{-1})}{(2-\alpha)(1+\beta)} &= 2 + \frac{2\kappa-1}{\kappa^2}\left(2 + \frac{2\kappa-1}{(\kappa-1)^2}\right) \\
&= \frac{2\kappa^4 - 4\kappa^2 + 4\kappa - 1}{\kappa^2(\kappa-1)^2} \\
&= \frac{2}{(\kappa-1)^2} \times \frac{\kappa^4 - 2(\kappa-1/2)^2}{\kappa^2} \\
&\leqslant \frac{2\kappa^2}{(\kappa-1)^2}.
\end{aligned}$$

Así, si la función de penalización es tal que

$$\text{pen}(m) \geqslant \kappa\sigma^2\left(D_m + \frac{2\kappa^2 x_m}{(\kappa-1)^2}\right),$$

hemos mostrado que

$$\mathbb{E}[\|s-\tilde{s}\|^2] \leqslant C \inf_{m\in\mathscr{M}} \left\{ \|s-H_m s\|^2 - \sigma^2 D_m + \mathrm{pen}(m) \right\} + C'\sigma^2\Sigma$$

donde

$$C = \frac{1}{1-\alpha} = \frac{\kappa}{\kappa-1} > 1$$

y

$$C' = \frac{4\alpha^{-1} + (2-\alpha)(2+\beta^{-1})}{1-\alpha} = \frac{4\kappa^4 - 4\kappa^3 - 2\kappa^2 + 4\kappa - 1}{(\kappa-1)^3} > 26.$$

2.4. Postre: adaptabilidad

En esta última sección, queremos ilustrar una de las ventajas de las desigualdades de oráculo, a saber la facilidad de obtener procedimientos adaptativos. Por eso, volvemos a considerar el marco estadístico de la regresión funcional con soporte fijo en $[0,1]$. Los datos son las pares $(x_1,Y_1),\dots,(x_n,Y_n) \in [0,1]\times\mathbb{R}$ con $x_i = i/n$, $i \in \{1,\dots,n\}$. Se supone también que el vector $Y = (Y_1,\dots,Y_n)^\top \in \mathbb{R}^n$ es gaussiano con componentes independientes y de misma varianza $\sigma^2 > 0$. Se puede ver el vector a estimar $s = \mathbb{E}[Y]$ como la discretización de una función $s^* : [0,1] \to \mathbb{R}$ en los puntos x_i,

$$s_i = E[Y_i] = s^*(x_i),\ i \in \{1,\dots,n\}.$$

Así, nuestro objetivo aquí es la estimación no paramétrica de esta función s^*. El riesgo es medido por la norma habitual en el espacio $L^2([0,1])$, a saber, para cualquier estimador $\hat{s}^*$,

$$\mathbb{E}\left[\|s^* - \hat{s}^*\|_{L^2}^2\right] = \mathbb{E}\left[\int_0^1 (s^*(t) - \hat{s}^*(t))^2\, dt\right].$$

Proponemos utilizar histogramas regulares en $[0,1]$. Para cualquier entero $D \geqslant 1$, consideramos las D funciones ortonormales $\varphi^*_{D,1},\dots,\varphi^*_{D,D} : [0,1] \to \mathbb{R}$ dadas por,

$$\varphi^*_{D,j}(x) = \begin{cases} \sqrt{D} & \text{si } \frac{j-1}{D} < x \leqslant \frac{j}{D} \\ 0 & \text{en caso contrario} \end{cases},\ x \in [0,1],\ j \in \{1,\dots,D\}.$$

Notando π_D la proyección ortogonal en el subespacio $S^*_D \subset L^2([0,1])$ generado por las funciones $\varphi^*_{D,j}$, tenemos el histograma

$$\pi_D s^* = \sum_{j=1}^{D} s^*_{D,j}\varphi^*_{D,j}$$

donde, para cualquier $j \in \{1,\dots,D\}$,

$$s^*_{D,j} = \sqrt{D}\int_{\frac{j-1}{D}}^{\frac{j}{D}} s^*(t)dt.$$

Al definir los D vectores $\varphi_{D,1},\dots,\varphi_{D,D} \in \mathbb{R}^n$ como las discretizaciones respectivas de las funciones $\varphi^*_{D,1},\dots,\varphi^*_{D,D}$ en los puntos x_i, podemos considerar el modelo lineal $S_D \subset \mathbb{R}^n$

generado por $\varphi_{D,1},\dots,\varphi_{D,D}$. Si $D \leqslant n$, el estimador $\hat{s}_D = H_D Y$ de s, donde H_D es la proyección ortogonal en S_D, se escribe como la combinación lineal

$$\hat{s}_D = \sum_{j=1}^{D} \hat{s}_{D,j}\varphi_{D,j}$$

donde $\hat{s}_{D,1},\dots,\hat{s}_{D,D} \in \mathbb{R}$. Entonces, definimos el estimador funcional $\hat{s}^*_D \in S^*_D$ de s^* por

$$\hat{s}^*_D = \sum_{j=1}^{D} \hat{s}_{D,j}\varphi^*_{D,j}.$$

El riesgo de este estimador $\hat{s}_D \in S_D$ es dado por

$$\mathbb{E}\left[\|s-\hat{s}_D\|_n^2\right] = \|s-H_D s\|_n^2 + \frac{\sigma^2 D}{n}.$$

Considerando la colección de modelos $\{S_D\}_{1\leqslant D\leqslant n}$ en $\mathbb{R}^n$, podemos aplicar el teorema 2.1 para elegir $\hat{D} \in \{1,\dots,n\}$ y definir un estimador $\tilde{s} = \hat{s}_{\hat{D}}$ y su versión funcional $\tilde{s}^* = \hat{s}^*_{\hat{D}} \in S^*_{\hat{D}}$. Como no hay más de $K = 1$ modelo por dimensión, deducimos una desigualdad de oráculo como en la sección 2.2,

$$\mathbb{E}\left[\|s-\tilde{s}\|_n^2\right] \leqslant C_1 \inf_{1\leqslant D\leqslant n}\left\{\|s-H_D s\|_n^2 + \frac{\sigma^2 D}{n}\right\} + \frac{C_2\sigma^2}{n}$$

donde $C_1, C_2 > 0$ dependen de los varios parámetros del procedimiento. Para n suficientemente grande, podemos deducir (con *sumas de Riemann*, por ejemplo) una desigualdad similar para los estimadores funcionales,

$$\mathbb{E}\left[\|s^*-\tilde{s}^*\|_{L^2}^2\right] \leqslant C_1' \inf_{1\leqslant D\leqslant n}\left\{\|s^*-\pi_D s^*\|_{L^2}^2 + \frac{\sigma^2 D}{n}\right\} + \frac{C_2'\sigma^2}{n} \tag{2.3}$$

con $C_1', C_2' > 0$ que pueden además depender de s^*.

Dados $\alpha \in (0,1]$ y $L > 0$, presentamos la bola de Hölder de regularidad α y de radio L,

$$\mathscr{H}_\alpha(L) = \left\{f : [0,1] \to \mathbb{R} \text{ tal que } \forall x,y \in [0,1], |f(x)-f(y)| \leqslant L|x-y|^\alpha\right\}.$$

Si $s^* \in \mathscr{H}_\alpha(L)$, no es difícil de mostrar que existe $C_{\alpha,L} > 0$ tal que

$$\|s^*-\pi_D s^*\|_{L^2}^2 \leqslant C_{\alpha,L} D^{-2\alpha}.$$

En efecto, tenemos

$$\begin{aligned}
\|s^* - \pi_D s^*\|_{L^2}^2 &= \int_0^1 (s^*(x) - \pi_D s^*(x))^2 \, dx \\
&= \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} \left(s^*(x) - \sqrt{D} \times s^*_{D,j} \right)^2 dx \\
&= \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} \left(s^*(x) - D \int_{\frac{j-1}{D}}^{\frac{j}{D}} s^*(y) dy \right)^2 dx \\
&= \frac{D}{2} \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} \int_{\frac{j-1}{D}}^{\frac{j}{D}} (s^*(x) - s^*(y))^2 \, dxdy \\
&\leqslant \frac{DL^2}{2} \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} \int_{\frac{j-1}{D}}^{\frac{j}{D}} |x-y|^{2\alpha} dxdy \\
&= \frac{DL^2}{2D^{2+2\alpha}} \sum_{j=1}^{D} \int_0^1 \int_0^1 |x-y|^{2\alpha} dxdy \\
&= C_{\alpha,L} D^{-2\alpha}.
\end{aligned}$$

La desigualdad de oráculo 2.3 lleva a

$$\begin{aligned}
\mathbb{E}\left[\|s^* - \tilde{s}^*\|_{L^2}^2 \right] &\leqslant C_1'(C_{\alpha,L} + \sigma^2) \inf_{1 \leqslant D \leqslant n} \left\{ D^{-2\alpha} + \frac{D}{n} \right\} + \frac{C_2' \sigma^2}{n} \\
&\leqslant C_1'' n^{-2\alpha/(2\alpha+1)} + \frac{C_2''}{n} \\
&\leqslant C^* n^{-2\alpha/(2\alpha+1)}
\end{aligned}$$

donde $C^* > 0$. En el ínfimo, hemos considerado el valor particular $D = \lceil n^{1/(2\alpha+1)} \rceil$.

En conclusión, sin hacer ninguna hipótesis sobre la función s^*, hemos construido un procedimiento de estimación no paramétrica de s^*. Nuestro estimador $\tilde{s}^*$ es tal que su riesgo converge a cero con la velocidad $n^{-2\alpha/(2\alpha+1)}$ sobre $\mathscr{H}_\alpha(L)$,

$$\sup_{s^* \in \mathscr{H}_\alpha(L)} \mathbb{E}\left[\|s^* - \tilde{s}^*\|_{L^2}^2 \right] \leqslant C^* n^{-2\alpha/(2\alpha+1)}.$$

Tal velocidad es llamada *minimax* porque es posible demostrar (ver *Introduction to Nonparametric Estimation*, Tsybakov) que existe $c^* > 0$ tal que

$$c^* n^{-2\alpha/(2\alpha+1)} \leqslant \inf_T \sup_{s^* \in \mathscr{H}_\alpha(L)} \mathbb{E}\left[\|s^* - T\|_{L^2}^2 \right].$$

donde el ínfimo es tomado con respecto a todos los estimadores de s^*. Por lo tanto, decimos que el estimador $\tilde{s}^*$ *se adapta* a la regularidad α de la función s^* porque la velocidad óptima se alcanza sin asumir el conocimiento de α.

Práctica: Algunas experiencias

2.5. Introducción

Ahora ofrecemos algunas experiencias para ilustrar los diferentes aspectos de la selección de modelos presentados en este curso. Estas experiencias se realizarán con el software libre *R*. La versión de *R* utilizada es 3.3.3 "Another Canoe" pero cualquier versión bastante reciente es adecuada. Los datos son disponibles en la página web del autor y se pueden cargarlas de la manera siguiente en *R*.

```
# Ozono
ozono_raw <- read.table("ozono", header=TRUE)
ozono <- as.matrix(ozono_raw[,c(1,3:14)])
```

Los datos `ozono` dan la concentración máxima de ozono `maxO3` para un día determinado en la ciudad de Rennes (Francia) y también las variables:

- `T6`, `T9`, `T12`, `T15` y `T18`: las temperaturas esperadas a las 6, 9, 12, 15 y 18 respectivamente,
- `Ne6`, `Ne9`, `Ne12`, `Ne15` y `Ne18`: la nebulosidad esperada a las 6, 9, 12, 15 y 18 respectivamente,
- `Vx`: velocidad del viento en el eje este-oeste,
- `maxO3v`: concentración máxima de ozono observada el dia anterior.

El objetivo es predecir esta concentración de ozono a partir de estas variables. Como primer ejemplo, consideramos la regresión polinomial de `maxO3` con respecto a `T18` para diferentes grados.

```
# Datos
T18 <- ozono[,'T18']
maxO3 <- ozono[,'maxO3']
```

```
plot(T18, maxO3)
T18_pred <- seq(min(T18), max(T18), length.out=256)

# Regresión lineal
mod1 <- lm(maxO3 ~ T18)
abline(mod1, col='blue', lty=3, lwd=2)
print(mod1$coefficients)

# Regresión polinomial de grado 6
mod6 <- lm(maxO3 ~ poly(T18, 6))
maxO3_pred <- predict(mod6, data.frame(T18=T18_pred))
points(T18_pred, maxO3_pred, type='l', col='red', lty=3, lwd=2)
print(mod6$coefficients)

# Regresión polinomial de grado 22
mod22 <- lm(maxO3 ~ poly(T18, 22))
maxO3_pred <- predict(mod22, data.frame(T18=T18_pred))
points(T18_pred, maxO3_pred, type='l', col='green', lty=3, lwd=2)
print(mod22$coefficients)

# Residuos cuadráticos
res <- sum(mod1$residuals^2)
for(p in 2:22) {
  model <- lm(maxO3 ~ poly(T18, p))
  res <- c(res, sum(model$residuals^2))
}
plot(res, type='b')
```

- ¿Qué comentarios se puede hacer sobre estos primeros modelos? En particular, ¿el modelo polinomial de mayor grado proporciona un buen estimador?
- ¿Qué miden los residuos cuadráticos y qué significa un valor bajo?
- El modelo de mayor grado tiene los residuos más bajos pero el estimador asociado no es ideal. Describe el fenómeno observado en relación con el riesgo del estimador.

Queremos elegir un grado $\hat{p} \in \{0,\dots,22\}$ por selección de modelos con un criterio de mínimos cuadrados y una función de penalización de la forma

$$\text{pen}(p) = \lambda\sigma^2(p+1)$$

donde $\lambda > 1$. Como la varianza σ^2 no es conocida, debemos estimarla en primer lugar.

```
# Proyección en un gran espacio de histogramas
n <- length(maxO3)
```

```
p <- n %/% 2
X <- matrix(0, nrow=n, ncol=p)
for(j in seq_len(p)) X[c(2*j-1, 2*j), j] <- 1
X[n, p] <- 1
H <- X %*% solve(t(X) %*% X) %*% t(X)

# Estimador de la varianza
sigma2 <- sum((maxO3 - H %*% maxO3)^2) / p
```

- Al descomponer el riesgo de `H %*% maxO3` en un término de sesgo y un término de varianza, explique cómo construimos nuestro estimador de la varianza arriba.

Ahora podemos calcular el criterio de los mínimos cuadrados penalizado para seleccionar un modelo polinomial de grado $\hat{p} \in \{1, \ldots, 22\}$.

```
crit <- NULL
pen <- NULL

# Criterio y penalización
for(p in 1:22) {
  model <- lm(maxO3 ~ poly(T18, p))
  crit <- c(crit, sum(model$residuals^2))
  pen <- c(pen, sigma2*(p+1))
}

# Resultados para varios valores de lambda
Lambda <- c(1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)
Lambda_col <- rainbow(length(Lambda))
p_hat <- rep(0, length(Lambda))
names(p_hat) <- Lambda

plot(crit, type='l', xlab='Grado', ylab='Criterio')
for(i in seq_along(Lambda)) {
  lambda <- Lambda[i]
  crit_pen <- crit + lambda*pen
  p_hat[i] <- which.min(crit_pen)
  points(crit_pen, type='l', lty=2, col=Lambda_col[i])
}
print(p_hat)
```

- ¿Cómo se comporta el criterio penalizado cuando λ aumenta? ¿Cuál es la consecuencia para el valor de $\hat{p}$ seleccionado?
- Calcule y visualice la función de regresión obtenida para los diferentes valores de λ.
- ¿Qué valor de λ parece ser "buena" aquí?

2.6. Modelos multivariantes

Para ir más lejos que en la sección anterior, podemos considerar modelos multivariantes de polinomios con respecto a las 12 variables en el conjunto de datos `ozono`. Por ejemplo, calculamos abajo la regresión polinomial de grado 2.

```
# 12 variables y grado 2
model <- lm(maxO3 ~ poly(ozono[,-1], 2, raw=TRUE))

# Resultados
print(sum(model$residuals^2))
print(length(model$coefficients))
print(model$rank)
```

- ¿Qué decir sobre los residuos cuadráticos?
- ¿Por qué hay una diferencia entre el número de coeficientes y el rango del modelo?
- Repita el mismo cálculo para el grado 3. ¿Qué nota? ¿Tenemos algo que ganar al aumentar el grado?

Más en general, podemos considerar todos los modelos polinomiales basados en los $2^{12} - 1 = 4095$ subconjuntos no vacíos de variables. Para limitar el tiempo de cálculo, solo consideramos las regresiones para las cuales *R* da un resultado en menos de 0,1 segundo.

```
# Subconjuntos de variables
sub_var <- expand.grid(rep(list(0:1), 12))[-1,]
colnames(sub_var) <- colnames(ozono[,-1])

# Proceso de los modelos
res <- list()
for(i in seq_len(nrow(sub_var))) {
  # Preparación
  cat('Model ', i, ':', sep='')
  res[[i]] <- list(crit=NULL, pen=NULL)
  X <- ozono[, which(sub_var[i,] == 1) + 1]

  # Aumentar el grado al máximo
  p <- 1
  calculando <- TRUE
  while(calculando) {
    # Soporte
    XX <- tryCatch(poly(X, p, raw=TRUE), error=function(e) {})
    if(is.null(XX)) {
      calculando <- FALSE
    } else {
```

```
      # Modelo de grado p
      start <- Sys.time()
      model <- tryCatch(lm(maxO3 ~ XX), error=function(e) {})
      end <- Sys.time()
      if(is.null(model)) {
        calculando <- FALSE
      } else {
        # Guardar los resultados
        cat(' ', p, sep='')
        res[[i]]$crit <- c(res[[i]]$crit, sum(model$residuals^2))
        res[[i]]$pen <- c(res[[i]]$pen, sigma2 * model$rank)

        # Limitar el tiempo de cálculo
        if((model$rank == nrow(ozono)) || (end - start > 0.1)) {
          calculando <- FALSE
        } else {
          p <- p + 1
        }
      }
    }
  }

  cat('\n')
}

print(sum(sapply(res, function(l) length(l$crit))))
```

- Observe cómo usamos `tryCatch` para evitar los errores. Este es un mecanismo muy útil de *R*.
- Observe también cómo limitamos el tiempo de cálculo de las regresiones lineales con `Sys.time`.
- ¿Para qué es la condición `model$rank == nrow(ozono)`?
- ¿Cuántos modelos tenemos al final?

Con una gran cantidad de modelos como es el caso aquí, necesariamente hay más candidatos para el sobreajuste. La dificultad que surge se relaciona principalmente con los modelos pequeños.

```
crit_p <- do.call('rbind',
  lapply(res, function(item) {
    matrix(c(seq_along(item$crit), item$crit), ncol=2)
  })
)
```

```
crit_p20 <- crit_p[crit_p[,1] <= 20,]
boxplot(crit_p20[,2] ~ crit_p20[,1])
```

- ¿Qué se puede decir de estos diagramas de caja?
- En relación con la selección de modelos, ¿por qué consideramos que los modelos pequeños serán problemáticos?

Para cualquier subconjunto V de variables y cualquier grado $p \geqslant 1$, podemos elegir un modelo como en la primera sección minimizando un criterio de mínimos cuadrados penalizados con respecto al par $m = (V, p)$,

$$\|Y - H_m Y\|^2 + \lambda \sigma^2 D_m$$

donde H_m es la proyección ortogonal en el modelo, $D_m = \mathrm{tr}(H_m)$ y $\lambda > 1$.

```
# Selección de modelos
m_hat <- matrix(0, nrow=2, ncol=length(Lambda))
dimnames(m_hat) <- list(c('Variables', 'Grado'), Lambda)

for(i in seq_along(Lambda)) {
  lambda <- Lambda[i]
  crit_pen <- sapply(res, function(l) min(l$crit + lambda*l$pen))
  m_hat[1,i] <- which.min(crit_pen)
  model <- res[[m_hat[1,i]]]
  m_hat[2,i] <- which.min(model$crit + lambda*model$pen)
}

print(m_hat)
```

- En función de λ, ¿cuáles son las variables y el grado del modelo seleccionado?
- ¿Qué valor de λ parece ser "buena" aquí? Compare este valor con el de la primera sección.

2.7. Regresión funcional y validación cruzada

En esta última sección, ya no manipularemos datos reales sino datos simulados para poder realizar varios cálculos. Nos ubicamos dentro del marco de la regresión funcional gaussiana con soporte fijo en $[0, 1]$.

```
# Simulación de datos
n <- 1024
x <- seq_len(n) / n
```

```
s <- sin(pi*x)
sigma2 <- 1
epsilon <- sqrt(sigma2) * rnorm(n)
Y <- s + epsilon

# Visualización
plot(x, Y, pch=4, cex=0.5)
points(x, s, type='l', col='red', lty=2, lwd=2)
```

Proponemos estimar s como en el curso usando histogramas regulares. Es decir que, para cualquier entero $D \geqslant 1$, definimos el modelo S_D de los histogramas regulares a D bloques con valores dados por la media de las observaciones entre $(j-1)/D$ y j/D, $j \in \{1,\dots,D\}$.

```
# Cálculo de la proyección
histograma <- function(D, x, v) {
  res <- rep(0, length(x))
  for(j in seq_len(D)) {
    idx <- ((j-1)/D < x) & (x <= j/D)
    res[idx] <- mean(v[idx])
  }
  return(res)
}

# Ejemplo
s50 <- histograma(50, x, Y)
Hs50 <- histograma(50, x, s)
plot(x, Y, pch=4, cex=0.5, col='grey')
points(x, s50, type='l')
points(x, s, type='l', col='orange', lty=2, lwd=2)
points(x, Hs50, type='l', col='red')
```

- ¿Cuál es la diferencia entre el histograma negro y el histograma rojo? ¿A cuál tenemos acceso como estadístico?

Notando H_D la proyección sobre el modelo S_D, sabemos que el riesgo del estimador $\hat{s}_D = H_D Y$ se divide en dos términos,

$$\mathbb{E}\left[\|s-\hat{s}_D\|_n^2\right] = \|s-H_D s\|_n^2 + \frac{\sigma^2 D}{n}.$$

```
Dmax <- 100
sesgo <- rep(0, Dmax)
varianza <- rep(0, Dmax)
```

```
for(D in seq_len(Dmax)) {
  Hs <- histograma(D, x, s)
  sesgo[D] <- mean((s - Hs)^2)
  varianza[D] <- sigma2 * D / n
}

# Visualización del riesgo
plot(sesgo, type='l', lty=2, xlab='Dimensión', ylab='')
points(sesgo + varianza, type='l')
```

- ¿Cómo se comporta el término de sesgo cuando la dimensión aumenta?
- ¿Cómo se comporta el riesgo cuando la dimensión aumenta?
- ¿Cuál es la dimensión óptima D^* aquí para estimar el vector s? Visualiza el estimador $\hat{s}_{D^*}$ y la proyección $H_{D^*}s$ en S_{D^*}.

Claro, cuando tratamos de estimar s, este vector es desconocido y no podemos calcular este riesgo para los diferentes valores de D. Para encontrar un compromiso entre los términos de sesgo y de varianza, minimizamos el siguiente criterio penalizado con respecto a $D \in \{1,\dots,D_{max}\}$,

$$\|Y - H_D Y\|_n^2 + \lambda \frac{\sigma^2 D}{n}$$

donde $\lambda > 1$.

```
# Criterio y penalización
Dmax <- 256
crit <- NULL
pen <- NULL
for(D in seq_len(Dmax)) {
  HY <- histograma(D, x, Y)
  crit <- c(crit, mean((Y - HY)^2))
  pen <- c(pen, sigma2 * D / n)
}

# Selección de modelos
Lambda <- seq(1, 5, length.out=256)
D_hat <- sapply(Lambda, function(lambda) {
    which.min(crit + lambda * pen)
  }
)
plot(Lambda, D_hat, type='l', ylab='Dimensión seleccionada')
```

- ¿Cómo se comporta la dimensión seleccionada cuando λ aumenta?

- Para $\lambda = 2$, la función de penalización es la de Mallows. ¿Cuál es la dimensión $\hat{D}_{Mallows}$ seleccionada? Visualiza el estimador y la proyección asociados.

- ¿Qué valor de λ parece ser "buena" aquí?

En la práctica, debemos hacer una elección del parámetro λ. Buscar a tientas como acabamos de hacer es un método muy cuestionable y necesitamos un enfoque más automático. La validación cruzada generalmente se usa para elegir un "buen" valor de λ en la práctica. Este enfoque consiste básicamente en separar los datos en dos partes: *datos de entrenamiento* para aplicar nuestro procedimiento con diferentes valores de λ y *datos de prueba* para elegir un valor de λ.

```
# Parámetros
Dmax <- 256
rho <- 0.7

# Separación de los datos
n_entrenamiento <- floor(rho * n)
id_entrenamiento <- sample.int(n, n_entrenamiento)
x_entrenamiento <- x[id_entrenamiento]
Y_entrenamiento <- Y[id_entrenamiento]

id_prueba <- seq_len(n)[-id_entrenamiento]
n_prueba <- length(id_prueba)
x_prueba <- x[id_prueba]
Y_prueba <- Y[id_prueba]

# Visualización de las errores
crit <- NULL
crit_prueba <- NULL
for(D in seq_len(Dmax)) {
  HY <- histograma(D, x_entrenamiento, Y_entrenamiento)
  crit <- c(crit, mean((Y_entrenamiento - HY)^2))
  HY_prueba <- sapply(x_prueba, function(x) {
    id <- which.min(abs(x_entrenamiento - x))
    return(HY[id])
  })
  crit_prueba <- c(crit_prueba, mean((Y_prueba - HY_prueba)^2))
}

plot(crit, type='l', ylim=range(c(crit, crit_prueba)),
     xlab='Dimensión', ylab='Criterio')
points(crit_prueba, type='l', col='red')
legend('bottomleft', legend=c('entrenamiento', 'prueba'),
       lty=1, col=c('black', 'red'))
```

- ¿Cuál es el papel del parámetro `rho`?
- ¿Por qué separamos los datos? ¿Cuál es la ventaja de esta separación?
- Ejecute este código algunas veces. ¿Qué observa?

Por lo tanto, podemos seleccionar un valor de λ automáticamente minimizando el error en los datos de prueba.

```
Lambda <- seq(1, 5, length.out=256)

# Procedimiento con los datos de entrenamiento
pen <- sigma2 * seq_len(Dmax) / n_entrenamiento
Dhat <- sapply(Lambda, function(l) { which.min(crit + l * pen) })

# Elección de lambda con los datos de prueba
err_vc <- crit_prueba[Dhat]
lambda_vc <- Lambda[which.min(err_vc)]
print(lambda_vc)
```

- Explica cómo se elige `lambda_vc`.
- Ejecute todo el código de validación cruzada algunas veces. ¿Qué problema nota sobre el valor de `lambda_vc`?

Una manera de estabilizar un poco el valor de λ elegido por validación cruzada consiste en dividir los datos en k partes, cada una de las cuales desempeñará el papel de los datos de prueba sucesivamente. Así, obtenemos k valores $\lambda_1, \ldots, \lambda_k$ de los cuales tomamos la media. Este método se llama *k-fold*.

```
# Preparación
k <- 8
fold <- matrix(sample(n), nrow=k)
lambda <- rep(0, k)
n_entrenamiento <- (k - 1) * n / k
pen <- sigma2 * seq_len(Dmax) / n_entrenamiento

# Hacer k iteraciones
for(i in seq_len(k)) {
  crit <- NULL
  crit_prueba <- NULL
  x_entrenamiento <- x[-fold[i,]]
  Y_entrenamiento <- Y[-fold[i,]]
  x_prueba <- x[fold[i,]]
  Y_prueba <- Y[fold[i,]]
```

```
  for(D in seq_len(Dmax)) {
    HY <- histograma(D, x_entrenamiento, Y_entrenamiento)
    crit <- c(crit, mean((Y_entrenamiento - HY)^2))
    HY_prueba <- sapply(x_prueba, function(x) {
      id <- which.min(abs(x_entrenamiento - x))
      return(HY[id])
    })
    crit_prueba <- c(crit_prueba, mean((Y_prueba - HY_prueba)^2))
  }

  Dhat <- sapply(Lambda, function(l) which.min(crit + l * pen))
  err_vc <- crit_prueba[Dhat]
  lambda[i] <- Lambda[which.min(err_vc)]
}

# Tomar la media
lambda_fold <- mean(lambda)
print(lambda_fold)
```

- Ejecute este código algunas veces para notar que `lambda_fold` es un poco más estable que `lambda_vc`.
- Varíe el valor de k. ¿Qué observa cuando k aumenta?
- Use el valor de λ obtenido con k-fold en el procedimiento de selección de modelos. Visualice el estimador obtenido.
- Repita el trabajo de esta sección con un vector s dado por una función menos regular que la función sinusoidal.