Machine Learning

Curso Machine Learning - DEST/UFMG
Marcos Antonio da Cunha Santos
Marcos Oliveira Prates

27 de agosto de 2018

Expansão de Bases e Regularização

- Até agora vimos métodos lineares para classificação, e.x. regressão linear, regressão polinomial.
- Porém é extremamente improvável que a verdadeira função $f(x)$ seja linear
- Modelos de regressão representam $f(x) = E(Y|X)$ e na realidade $E(Y|X)$ será não linear e nem aditivo, porém uma representação linear aditiva pode ser uma boa aproximação e necessária para um melhor entendimento

- Nesse capítulo iremos estudar métodos que vão além da linearidade
- A idéia principal é trocar o vetor de inputs X por variáveis adicionais, que serão transformações de X e ai usar uma aproximação linear nesse novo espaço gerado pelo inputs
- Seja $h_m(X) : \Re^p \rightarrow \Re$ a m-ésima transformação de X, $m = 1, \ldots, M$

- Dessa forma temos que $f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$, ou seja, uma expansão linear de bases em X
- Uma observação interessante é que após a definição de h_m os modelos são linares sob essas novas variáveis e os métodos propostos podem então ser aplicados nesse novo espaço

- Portanto, escolhendo h_m podemos ter diferentes modelos:
  1. $h_m(X) = X_m$, $m = 1, \ldots, p$ recupera o modelo tradicional
  2. $h_m(X) = X_j^2$ ou $h_m(X) = X_j X_k$ permite uma aproximação de Taylor de maiores ordem. Note porém que a complexidade do modelo cresce exponencialmente com a ordem da aproximação de Taylor, ou seja, para uma aproximação d-variada a complexidade do modelo aumenta na order $O(p^d)$
  3. $h_m(X) = \sqrt{X_j}$ permite a transformação não linear dos inputs

- Alguns problemas práticos irão “exigir” para uma escolhar particular das funções de base h_m, como logaritmos ou funções de exponencias.
- Mais comumente, podemos escolher as fuções de base h_m para aumentar a flexibilidade $f(X)$ e portanto, do modelo
- Polinômios são um exemplo desse fato de flexibilização, porém, devido a sua natureza polinômios tendem a distorcer a “realidade” em regiões remotas

- Nesse capítulo consideramos famílias de polinômios por partes e splines que permitem uma representação local por polinômios
- Esses métodos geram um dicionário de bases D mais do que suficientes para ajustar dados. Porém, além de um dicionário vasto precisamos de maneiras para controlar a complexidade do nosso modelo.

- Existem três maneiras comuns:
  1. Método de Restrição: onde se decide a priori a limitar a classe de funções. Modelos aditivos são um exemplo

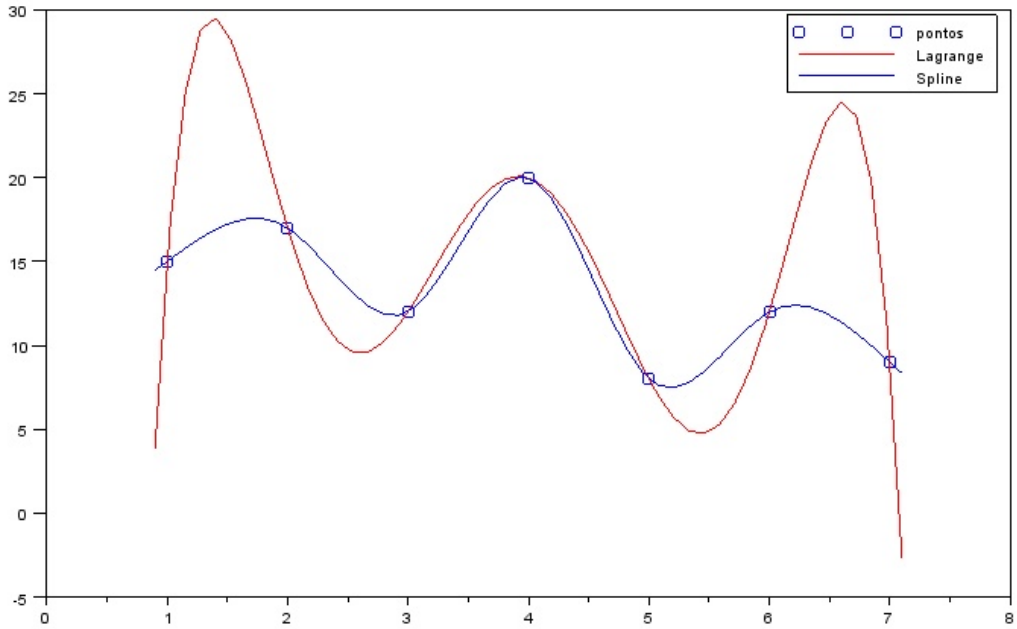
$$f(X) = \sum_{j=1}^{p} f_j(X_j) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{mj} h_{mj}(X_j)$$

     o tamanho desses modelos são controlados pelo número de funções de bases M_j em cada componente f_j do modelo
  2. Modelos de Seleção: que adaptativamente procura dentro do dicionário D e adiciona aquelas bases h_m que contribuem significativamente para melhorar o ajuste do modelo. Exemplos são os métodos apresentados anteriormente, como: stepwise, forward, backward, etc. Métodos CART e MART também são partes dessa categoria
  3. Métodos de Regularização: onde se usa todo o dicionário mas restringe os coeficientes. Ridge e Lasso são exemplos dessa categoria

Polonômios por Partes e Splines

- Para melhorar o ajuste aos dados podemos utilizar ajuste polinômiais
- Utilizando um polinômio de ordem $n-1$ temos um ajuste perfeito aos dados, ou seja, o polinômio é capaz de passar por todos os pontos dos dados
- Porém, apesar de ajuste ser pefeito é fácil ver que a curva ajustada não é robusta para predição de novos dados
- Uma alternativa é ajustar polinômios de menor graus por partes

Aproximações Polinômiais



- Para simplificar, supomos que X é unidimensional
- Um polinômio por partes é obtido separando o dominio de X em intervalos conectados e definindo f como um polinômio por intervalo
- O primeiro é o polinômio constante por partes, ou seja,

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X)$$

- Então para as 3 sub-regiões definidas o estimador de minimos quadrados para o modelo $f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$ é $\hat{\beta}_m = \bar{Y}_m$ no qual $\bar{Y}_m$ é a média dos Y's pertencentes ao m-ésimo intervalo

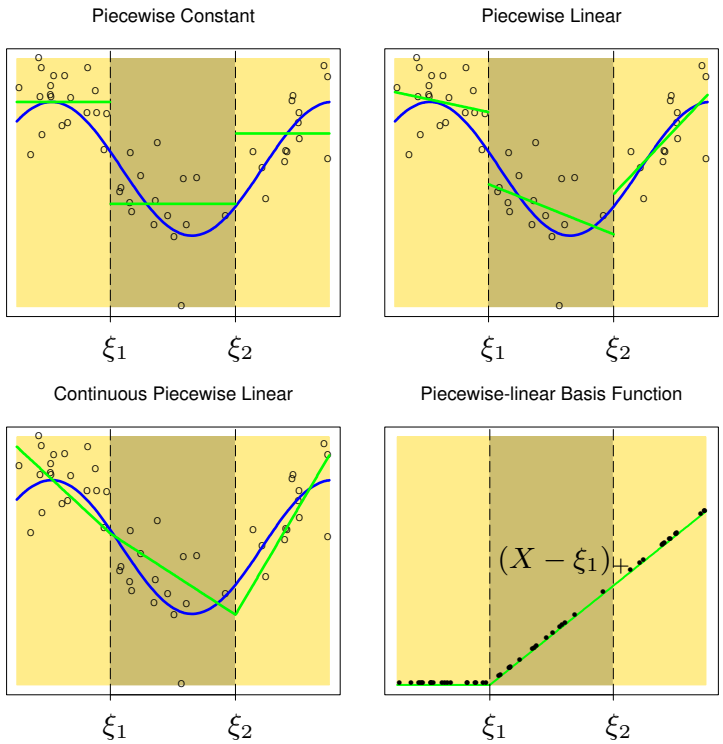


FIGURE 5.1. *The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two* knots *ξ_1 and ξ_2. The blue curve represents the true function, from which the data were generated with Gaussian noise. The remaining two panels show piecewise linear functions fit to the same data—the top right unrestricted, and the lower left restricted to be continuous at the knots. The lower right panel shows a piecewise–*

- No painel superior direito temos a aproximação linear por partes
- Para isso, um acréscimo de três bases é necessário

$$h_{m+1}(X) = h_m(X)X, \text{ para } m = 1, 2, 3$$

- Ambos exemplos apresentados não se preocupam com a continuidade da função $f(X)$

- Logo, para se garantir a continuidade precisamos colocar restrições no modelo
- Por exemplo, para garantir a continuidade linear precisamos que

$$f(\xi_1^-) = f(\xi_1^+) \text{ e } f(\xi_2^-) = f(\xi_2^+)$$

ou seja

$$\beta_1 + \beta_4\xi_1 = \beta_2 + \beta_5\xi_1 \text{ e } \beta_2 + \beta_5\xi_2 = \beta_3 + \beta_6\xi_2$$

- Dessa forma, como temos 2 restrições "perdemos" 2 graus de liberdade, nos deixando então com 4 parâmetros

- Uma maneira mais direta é utilizar uma base que diretamente incorpora tais restrições

$$h_1(X) = 1,\ h_2(X) = X,\ h_3(X) = (X - \xi_1)_+,\ h_4(X) = (X - \xi_2)_+$$

- A função $h_3(X)$ é mostrada no canto inferior da figura
- Simplesmente,

$$h_3(X) = (X - \xi_1)_+ = \begin{cases} X \text{ se } X > \xi_1 \\ 0 \text{ caso contário} \end{cases}$$

- Comumente se tem interesse em funções mais suaves, e para isso pode-se aumentar a ordem do polinômio
- Diferentes graus de suavização na curva fornecem diferentes ajustes
- Esses graus de suavização são obtidos fazendo restrições sob as derivadas do polinômio
  - Continuidade, as curvas se encontram nos knots
  - Primeira derivada, as tangentes também são as mesmas nos knots
  - Segunda derivada, as curvaturas são as memas nos knots
- Polinômios cúbicos com restrições na primeira e segunda derivadas são conhecidos como splines cúbicos

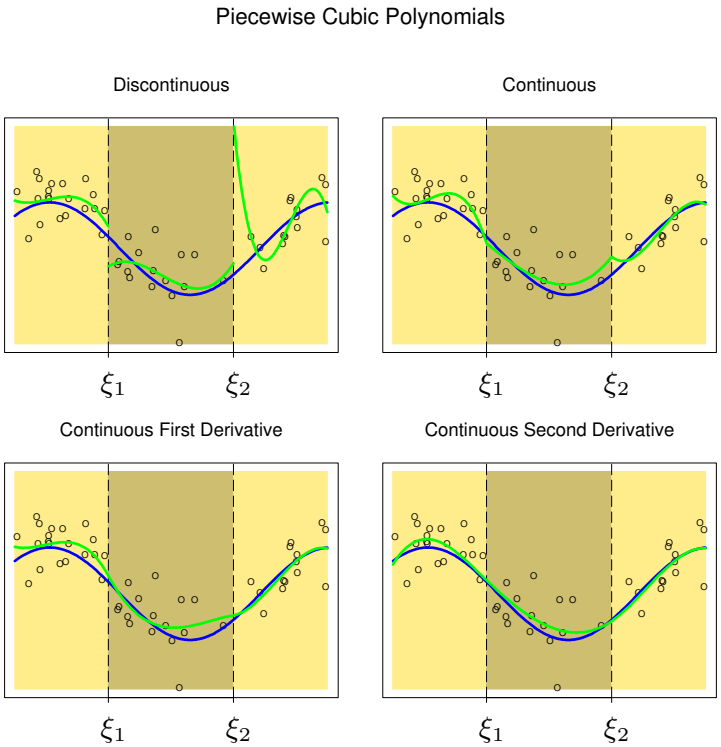


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

- É possível mostrar que a base

$$h_1(X) = 1,\ h_3(X) = X^2,\ h_5(X) = (X - \xi_1)^3_+,$$
$$h_2(X) = X,\ h_4(X) = X^3,\ h_6(X) = (X - \xi_2)^3_+.$$

- É uma base de spline cúbica com nós em ξ_1 e ξ_2
- É simples ver que existem 6 funções de base para um espaço linear de seis dimensões
- Para verificar basta contar: 3 regiões x 4 parâmetros por região - 2 knots x 3 restrições por knot (continuidade, 1a e 2a derivadas) = 6

- De forma geral um spline de ordem M com knots ξ_j, $j = 1, 2, \ldots, K$ é um polinômio de ordem M e tem derivadas continuas até ordem M-2.
- Por exemplo, uma spline cubica tem M = 4
- De forma geral temos que a base para um spline de ordem M é

$$\begin{aligned} h_j(X) &= X^{j-1},\ j = 1, \ldots, M, \\ h_{M+l}(X) &= (X - \xi_l)_+^{M-1},\ l = 1, \ldots, K \end{aligned}$$

Explicando a Spline com caso Simples

- No espaço com $n+1$ pontos $(x_0, y_0), \ldots, (X_n, y_n)$
- Encontrar a função $f(X)$ que seja um polinômio cúbico com knots $\xi_i = x_i$, $i = 0, \ldots, n$
- Passe pelos pontos (x_i, y_i)
- e seja suave nos knots, ou seja,
  - $f'(x)$ seja a mesma nos knots
  - $f''(x)$ seja a mesma nos knots
  - $f''(x_0) = f''(x_n) = 0$

- Essas condições garatem que o sistem a ser resolvido será inversível, ou seja, terá solução única
- Focamos no caso simples com os seguinte pontos $(-1,1)$, $(1,3)$, $(2,0)$
- Logo, queremos 2 polinômios cúbicos $p_1(x)$ e $p_2(x)$ tais que
  - $p_1(x)$ passa por $(-1,1)$ e $(1,3)$
  - $p_2(x)$ passa por $(1,3)$ e $(2,0)$
  - Portanto o ponto $(1,3)$ é comum a ambos polinômios

- As restrições apresentadas geram 4 sistemas lineares

$$
\begin{aligned}
1 &= a_{01} - a_{11} + a_{21} - a_{31} \\
3 &= a_{01} + a_{11} + a_{21} + a_{31} \\
3 &= a_{02} + a_{12} + a_{22} + a_{32} \\
0 &= a_{02} + 2a_{12} + 4a_{22} + 8a_{32}
\end{aligned}
$$

- Até agora temos 4 restrições e 8 incógnitas

- Mas não queremos que os polinômios simplesmente se encontrem no ponto $(1,3)$ mas também gostaríamos que o encontro fosse suave, por exemplo, que $p_1'(1) = p_2'(1)$
- Assim, colocamos mais uma restrição no sistema da seguinte forma

$$a_{11} + 2a_{21} + 3a_{31} - a_{12} - 2a_{22} - 3a_{32} = 0$$

- Logo temos agora 5 equações e 8 incógnitas
- Gostaríamos também de suavizar a curvatura, ou seja, $p_1''(1) = p_2''(1)$
- Essa restrição resulta em mais uma equação

$$2a_{21} + 6a_{31} - 2a_{22} - 6a_{32} = 0$$

- Precisamos ainda definir mais 2 equações pois temos até agora um total de 6 equações com 8 incógnitas
- O fato de termos escolhidos que $f''(-1) = f''(2) = 0$ nos fornece essas restrições, ou seja,

$$\begin{aligned} p_1''(-1) &= 2a_{21} - 6a_{31} = 0 \\ p_2''(2) &= 2a_{22} + 12a_{32} = 0 \end{aligned}$$

- As diversas possíveis escolhas para essas duas restrições finais é o que difere os diferentes tipos de splines
- O spline que restringe que a segunda derivada nos pontos extremos seja 0 é chamado de splines natural
- Isso pode ser entendido da seguinte forma, se as funções forem aproximadamente lineares nos extremos suas segundas derivadas deveriam ser 0.

Voltando as Splines

- As splines cúbicas são as splines de menor ordem nos quais a discontinuidade nos nós são suaves o suficiente para não serem vistas pelo olho humano
- Portanto, existe pouca justificativa para se ir para splines de maior ordem a não ser que por algum motivo específico precise mais derivadas suavizadas
- Esses spline com knots pré-específicados também são conhecidos como splines de regressão
- Dessa forma é necessário escolher a ordem do spline, o número de knots e suas localizações

- Uma spline natural faz a suposição que as funções são linares além das fronteiras
- Essas hipótese de linearidade nas regiões próximas as fronteiras é considerada adquada já que essa é um região com pouca informação
- A base de uma spline natural pode ser definida da seguinte forma

$$N_1(X) = 1,\ N_2(X) = X^2,\ N_{k+2}(X) = d_k(X) - d_{K+1}(X),$$

onde $d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$

Exemplo de Splines

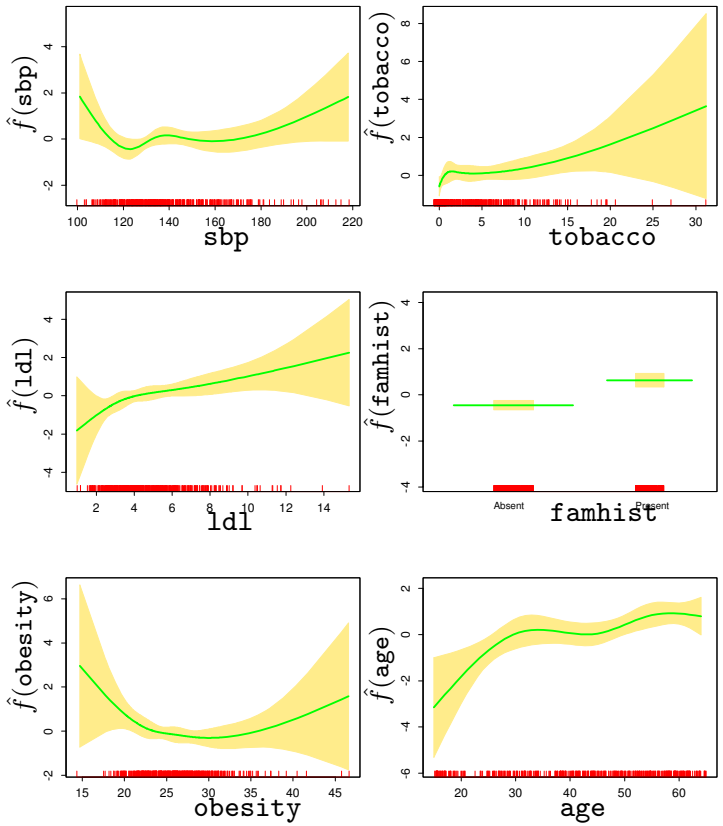


FIGURE 5.4. *Fitted natural-spline functions for each of the terms in the final model selected by the stepwise procedure. Included are pointwise standard-error bands. The* rug plot *at the base of each figure indicates the location of each of the sample values for that variable (jittered to break ties).*

Splines Suavizadas

- Agora iremos apresentar um tipo de splines que evita a escolha do knots maximizando o maior numero de knots possível
- A complexidade do ajuste é controlada por regularização
- Entre todas as possíveis $f(X)$ com segunda derivada contínua, encontre aquela que minimize a soma quadrática dos resíduos

$$RSS(f,\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 - \lambda \int (f''(t))^2 dt,$$

onde λ é o parâmetro de suavização

- O primeiro termo mede a proximidade do ajuste aos dados
- Enquanto o segundo penaliza a curvatura da função e λ define a troca entre ambos
- Dois casos especiais são
  1. $\lambda = 0$: f pode ser qualquer função que interpola os dados
  2. $\lambda = \infty$: um modelo de regressão linear já que nenhuma função com segunda derivada pode ser tolerada
- É possível mostrar que essa função possui um maximador único que é uma spline natural com knota $\xi_i = x_i$ para $i = 1, \ldots, N$

- A primeira vista parece que temos um modelo super parametrizado já que temos N knots
- Porém a penalização garante que alguns coeficientes das splines sejam encolhidos para o regressor linear
- Como a solução das splines suavizadas é uma spline natural podemo escrever

$$f(X) = \sum_{j=1}^{N} N_j(x)\theta_j,$$

onde N_j é a base de um spline natural

- Assim o problem de se minimizar

$$RSS(f,\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 - \lambda \int (f''(t))^2 dt,$$

se reduz a

$$RSS(\theta,\lambda) = (y - N\theta)^\top (y - N\theta) - \lambda\theta^\top \Omega_N \theta$$

e $N_{ij} = N_j(x_i)$ e $\{\Omega_N\}_{jk} = \int N_j''(t) N_k''(t) dt$

- Logo, esse problema é equivalente a uma solução de ridge generalizada e é dado por

$$\hat{\theta} = (N^\top N + \lambda\Omega_N)^{-1} N^\top y$$

- E a curva ajustada é $\hat{f}(x) = \sum_{j=1}^{N} N_j(x)\hat{\theta}_j$

Spline Suave

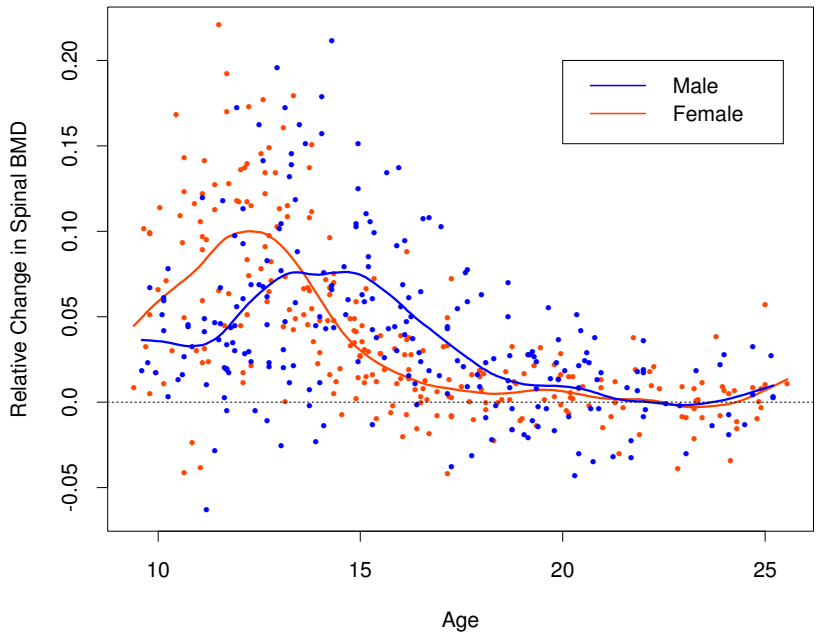


FIGURE 5.6. *The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with* $\lambda \approx 0.00022$. *This choice corresponds to about* 12 *degrees of freedom.*

Escolha da Parâmetro de Suavização

- Assim como em uma regressão linear temos que $\hat{f} = N(N^\top N + \lambda\Omega_N)^{-1}N^\top y$ ou $\hat{f} = S_\lambda y$
- Lembre-se que no caso geral de regressão temos que $\hat{f} = X(X^\top X)^{-1}X^\top y$ ou $\hat{f} = Hy$
- Ambos métodos apresentam a mesma forma mas existem semelhanças e diferenças entre eles

Análise de Discriminantes Regularizados

- Semelhanças e diferenças
  - Ambas são simétricas e semi-positiva definidas
  - H é idepotente enquanto S_λ não. Em outras palavras $S_\lambda S_\lambda \preceq S_\lambda$, ou seja, o lado direito excede o lado esquerdo por uma matriz semi-positiva definida. Dando assim, o caracter suavizador de S_λ
  - H tem posto M e S_λ tem posto N
- Pela regressão linear sabemos que $tr(H) = M$ e provê a dimensão do espaço de projeção de H
- Por analogia iremos definir os graus de liberdade efetivos de S_λ como o $df_\lambda = tr(S_\lambda)$

- Essa definição oferece uma maneira mais intuitiva de se escolher λ
- Por exemplo, na última figura escolheu-se $df_\lambda = 12$ e encontrou $\lambda = 0.00022$ resolvendo numericamente o sistema $tr(S_\lambda) = 12$
- Como S_λ é simétrica e semi-positiva definida podemos fazer uma decomposição espectral em S_λ da seguinte forma

$$S_\lambda = \sum_{k=1}^{N} \rho_k(\lambda) u_k u_k^\top$$

onde u_k são auto-vetores e $\rho_k(\lambda)$ auto-valores e

$$\rho_k(\lambda) = \frac{1}{1 + \lambda d_k}$$

no qual d_k é uma constante que depende de uma matriz de penalização

- Primeiramente vemos que u_k não depende de λ, somente $\rho_k(\lambda)$
- Claramente temos que $\rho_k(\lambda)$ é inversamente proporcional à λ
- Pode-se mostrar que $df_\lambda = tr(S_\lambda) = \sum_{k=1}^{N} \rho_k(\lambda)$
- Assim, quanto maior λ menor o df_λ e vice versa.
- Assintóticamente temos
  - Quando $\lambda \to 0 \Rightarrow df_\lambda \to N$, ou seja, $S_\lambda \to I$
  - Quando $\lambda \to \infty \Rightarrow df_\lambda \to 2$, ou seja, $S_\lambda \to H$

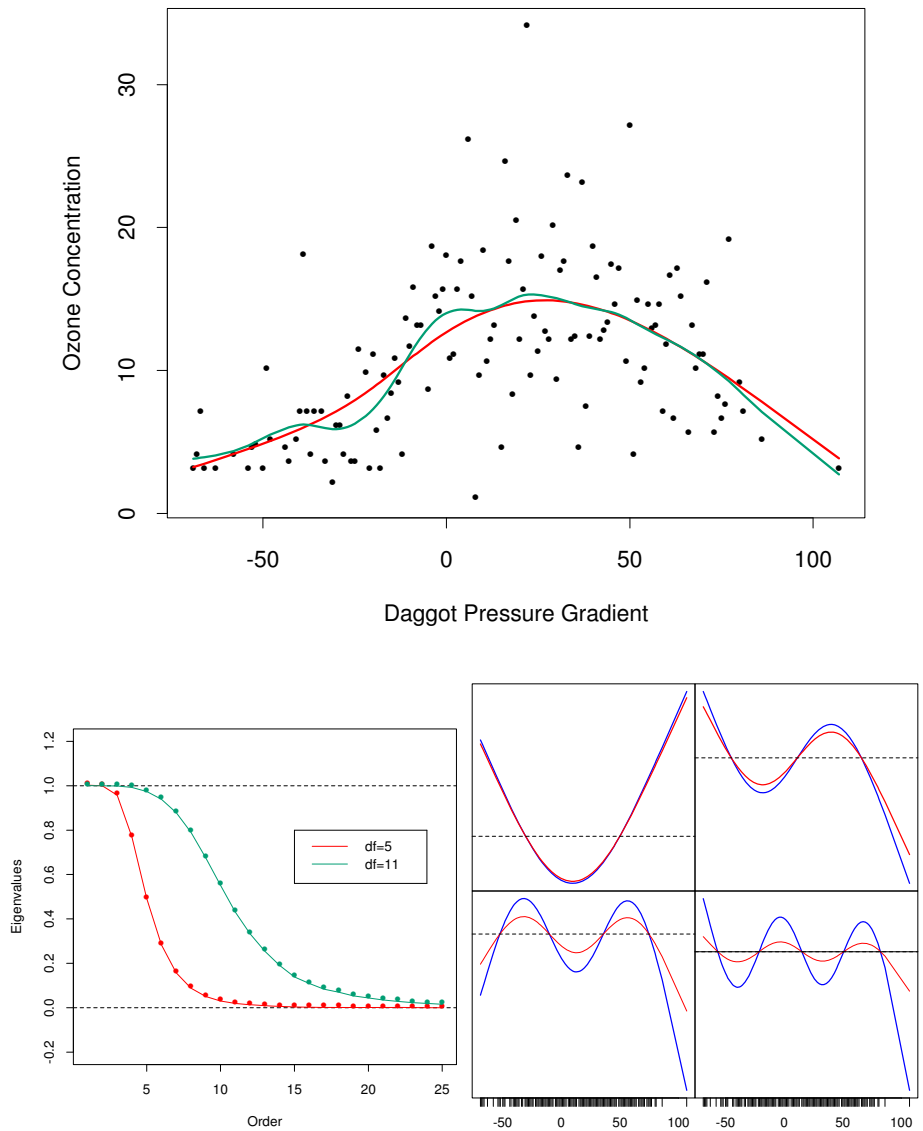


FIGURE 5.7. *(Top:) Smoothing spline fit of ozone*

Escolha dos Graus de Liberdade para Splines Suaves

- Como os graus de liberdade df_λ é uma função monotonica em λ podemos fixar tanto df_λ como λ para escolher a suavização
- Utilizando a idéia de troca entre vício e variabilidade vamos sugerir maneiras de se escolher df_λ
- Suponha o seguinte exemplo

$$\begin{aligned} Y &= f(X) + \varepsilon \\ f(X) &= \frac{\sin(12(X+0.2))}{X+0.2} \end{aligned}$$

com $X \sim U(0,1)$ ew $\varepsilon \sim N(0,1)$

- Gera uma amostra N = 100

- Utilizando as Splines Suaves sabemos que $\hat{f} = S_\lambda y$
- Logo, para o modelo proposto $Cov(\hat{f}) = S_\lambda Cov(y) S_\lambda^\top = S_\lambda S_\lambda^\top$
- E $Bias(\hat{f}) = f - E(\hat{f}) = f - S_\lambda f$
- Ajustamos a splines suaves escolhendo três graus de liberdade $df_\lambda = 5$, $df_\lambda = 9$ e $df_\lambda = 15$

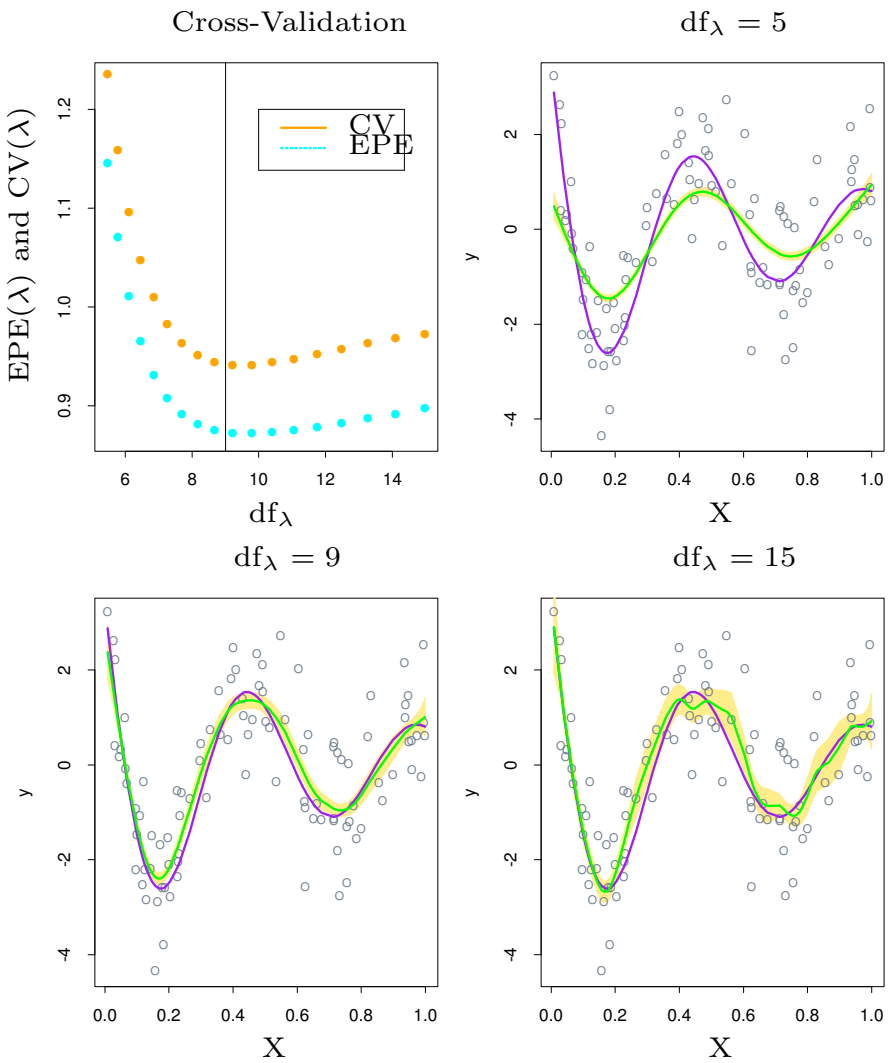


FIGURE 5.9. *The top left panel shows the* EPE(λ) *and* CV(λ) *curves for a realization from a nonlinear ad-*

- Vemos que
  - df_5 O ajuste é muito suave, ou seja, reduz demasiadamente os picos e preenche os vales. Esse fato leva a vício maiores onde temos curvatura e além disso, o desvio padrão e pequeno fazendo com que os estimadores sejam muito viciados e com alta confiabilidade
  - df_9 O ajuste é muito melhor que o anterior e apesar de existir um pequeno vício a variabilidade continiou controlada
  - df_{15} O ajuste agora é menos suave mas próximo da função verdadeira e a variabilidade da curva aumentou. Isso se deve ao fato de que com df_15 alguns pontos estãos sendo influentes para estimação da curva

- Vamos definir então o “Integrated Squared Prediction Error”

$$\begin{aligned} EPE(\hat{f}_\lambda) &= E(Y - f_\lambda(X))^2 \\ &= Var(Y) + E(Bias^2(\hat{f}_\lambda(X)) + Var(\hat{f}_\lambda(X))) \\ &= \sigma^2 + MSE(\hat{f}_\lambda) \end{aligned}$$

- Assim o *EPE* leva em conta tanto o vício como a variabilidade
- Porém o *EPE* precisa do conhecimento da verdadeira função geradora de Y

- Como não conhecemos a verdadeira função geradoras dos dados devemos estimar o EPE
- Para isso, vamos utilizar N-fold CV, ou seja, fazer a validação cruzada 1 a 1
- Assim temos

$$\begin{aligned} CV(\hat{f}_\lambda) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\lambda^{(-i)}(x_i))^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(y_i - f_\lambda(x_i))}{1 - S_\lambda(i,i)}\right)^2 \end{aligned}$$

onde $S_\lambda(i,i)$ são os elementos diagonais de S_λ

- Assim vemos que para calcular o *CV* basta ajustar o modelo uma única vez já que a igualdade depende dos parâmetros dos modelo compeleto
- Em geral *CV* é um estimador não viciado de *EPE* e para esse exemplo, apesar de estar sempre acima do *EPE* possui o mesmo comportamento selecionando o mesmo df_λ

Regressão Logística Não-Paramétrica

- Vimos que os splines suaves podem ser entendidos e calculados na forma de uma regressão linear
- Portanto, a trasferência de tecnlogia para metodologias associadas as regressões lineares se da de forma simples e direta
- Suponha que consideramos uma regressão logística com X o input univariado. Assim temos

$$\log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = f(x),$$

- Logo

$$P(Y=1|X=x) = \frac{e^{f(x)}}{1+e^{f(x)}}$$

- Escolhendo $f(x)$ de maneira suave nos oferece um estimador suave para $P(Y = 1|x) = p(x)$
- Pela aproximação linear apresentada (Cap 4) e argumentos equivalente aos apresentados anteriormente concluimos que $f(x) = \sum_{i=1}^{N} N_i(x)\theta_j$
- Utilizando a aproximação linear da regressão logística temos

$$\begin{aligned} f^{new} &= N(N^\top WN + \lambda\Omega)^{-1} N^\top W(f^{old} + W^{-1}(y - p)) \quad (1) \\ &= S_\lambda z, \end{aligned}$$

onde $z = f^{old} + W^{-1}(y - p)$

- Logo os métodos apresentados para splines suaves em regressões lineares também são válidos para regressões logísticas
- A extensão para problemas com X em maior dimensão também é direta