Regressão Não Paramétrica

Ronaldo Dias

Universidade Estadual de Campinas.

E-mail address: dias@ime.unicamp.br

1 Introdução

É sempre útil começar um estudo sobre análise de regressão fazendo uso de modelos bem simples. Para tanto, suponha que observações são coletadas de uma variável contínua Y em n valores da variável independente t. Sejam (t_j, y_j), $j = 1, \ldots, n$, tal que o seguinte modelo de regressão pode ser proposto:

$$y_j = f(t_j) + \varepsilon_j, \quad j = 1, \ldots, n$$

onde as variáveis aleatórias ε_j tem média zero, são não correlacionadas, e variância comum σ^2. Mas ainda, $f(t_j)$ são valores obtidos de alguma função f, desconhecida, calculada nos pontos $t_1, \ldots, t_n$. A função f é geralmente chamada de *função de regressão* ou *curva de regressão.*

Um modelo de regressão paramétrico assume que a forma de f é conhecida exceto por um número finito de parâmetros. Isto é, o modelo de regressão paramétrico pode ser descrito por,

$$y_j = f(t_j, \beta_1, \ldots, \beta_p) + \varepsilon_j, \quad j = 1, \ldots, n$$

onde $\beta = (\beta_1, \ldots, \beta_p)^t \in \mathbb{R}^p$. Então determinar, através dos dados, a curva f é equivalente a determinar o vetor β de parâmetros. Observe que se f tem forma

linear, i.e., $f(t,\beta) = \sum_{j=1}^{p} \beta_j x_j(t)$, estamos na situação do modelo de regressão linear paramétrico.

Certamente existem outros métodos de se ajustar curvas aos dados. A técnica de regressão não paramétrica, por exemplo, permite maior flexibilidade na possível forma de f, pois com esta técnica, assume-se que a curva de regressão pertence a uma coleção infinito dimensional de funções. Por exemplo, f pode pertencer a família de funções que são diferenciáveis e a segunda derivada é de quadrado integrável. Consequentemente, o experimentador para construir um modelo de regressão não paramétrica precisa apenas escolher o espaço de funções apropriado, ao qual se acredita que f pertença. Esta escolha, geralmente, é motivada pelo grau de suavidade que a função de regressão pode ter. Note que o caso paramétrico é mais restrito, pois nele assume-se que f pertence a uma específica família de curvas. Daí, as técnicas de regressão não paramétrica usam muito mais as informações provindas dos dados para estimar a função de regressão do que as técnicas paramétricas. De fato, regressão não paramétrica "deixa os dados falarem por si mesmos". Porém esta grande vantagem da modelagem não paramétrica tem um preço. Em geral, estimadores não paramétricos são menos eficientes que os estimadores paramétricos quando o modelo paramétrico é válido. Para muitos estimadores paramétricos o erro quadrático médio vai a zero a uma taxa de n^{-1} enquanto que nos estimadores não paramétricos esta taxa é de $n^{-\alpha}$, para algum parâmetro $\alpha \in (0,1)$, que depende da suavidade da função f. Por exemplo, se f é duas vezes diferenciável, então a taxa de decaimento a zero é de $n^{-4/5}$. Assim estimadores não paramétricos tornam-se bons candidatos a estimação da curva de regressão quando existem dúvidas sobre a forma da f. No caso em que um modelo paramétrico incorreto, ad hoc, é usado, a taxa n^{-1} não será obtida. Na verdade, o estimador não converge para a verdadeira curva de regressão.

Concluindo, os modelos de regressão paramétricos e não paramétricos representam distintas formas para a análise de regressão, isto não significa que um método

se sobrepõe ao outro. Na verdade, técnicas de regressão não paramétricas podem ser usadas para verificar a validade de um certo modelo paramétrico proposto. Reciprocamente, a forma da curva de de regressão obtida por técnicas não paramétricas podem sugerir um modelo paramétrico. Assim procedimentos de regressão não paramétricos poderão ser o estágio final de uma análise de dados ou meremente um passo confirmatório ou explantório do processo de modelagem.

2 Espaço de Funções e Estimadores por Séries

A motivação que nos leva a estudar o estimador por séries vem justamente do modelo paramétrico linear, isto é, $f(t) = \sum_{j=1}^{K} \beta_j x_j(t)$, para algum conjunto de funções $x_1, \ldots, x_K$. Portanto, estimar a função de regressão f é equivalente a estimar o vetor de parâmetros $\beta = (\beta_1, \ldots, \beta_K)^t$. Porém, como foi dito anteriormente, existe situações onde f pode pertencer a uma classe infinito dimensional, tal que $f(t) = \sum_{j=1}^{\infty} \beta_j x_j(t)$, para algum conjunto de funções conhecidas $\{x_j\}$. Consequentemente, a idéia é aproximar f por $\sum_{j=1}^{K} \beta_j x_j$, onde K é um inteiro positivo que determina o número de funções x_j necessários para estimar razoávelmente bem (em algum sentido) f. Então, para K fixo, resta estimar, por exemplo por mínimos quadrados, o vetor de parâmetros β.

2.1 Uma Breve Introdução a Teoria de Espaço de Funções

Dentre os espaços de funções mais utilizados nas aplicações, destaca-se, por sua importância e aplicabilidade, o espaço das funções definidas em $[a, b]$ de quadrado integrável denotado por $L^2[a, b]$. É muito fácil verificar que este espaço de funções é um espaço vetorial cuja norma,

$$\| f \| = (\int_a^b f^2(t)dt)^{1/2}$$

é derivada do produto interno;

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt \ , \ \forall f, g \in L^2[a, b]$$

Infelizmente, a avaliação de qualquer elemento de $L^2[a,b]$ nos pontos em $[a,b]$ não esta bem definida, pois duas funções $f, g \in L^2[a,b]$ são consideradas identicas se $\|f-g\| = 0$, ou seja, $f \equiv g$ se elas diferem apenas num conjunto de pontos em $[a,b]$ cuja a medida de Lebesgue é nula. Daí, devemos procurar a função de regressão f em um subconjunto de $L^2[a,b]$ tal que as funções deste subconjunto sejam "suaves". A escolha mais comum de tal subconjunto é o espaço de funções,

$$W_2^m[a,b] = \{f : f^{(j)} \text{ é absolutamente contínuas, } j = 0, \ldots, m-1 \text{ e } f^{(m)} \in L^2[a,b]\}.$$

Para estudarmos estimadores por séries necessitamos de algumas definições.

Definição 2.1 *Duas funções $f, g \in L^2[a,b]$ são ortogonais se o produto interno é zero, i.e, $\langle f, g\rangle = 0$.*

Definição 2.2 *Uma sequência de funções $\{x_j\}$ é dita ser ortonormal se*

$$\langle x_i, x_j\rangle = \begin{cases} 0 & \text{se } i \neq j \\ 1 & \text{se } i = j \end{cases}$$

e $\|x_j\| = 1$ para todos os i, j.

Definição 2.3 *Uma sequência de funções $\{x_j\}$ é dita ser completa se a única função g que satisfaz $\langle g, x_j\rangle = 0$, $\forall j$ é a função nula, isto é, $g \equiv 0$.*

Um exemplo de sequência completa de funções para $L^2[a,b]$ é dada por,

$$x_j = (b-a)^{-1/2} e^{2\pi i j t/(b-a)} \quad \text{para } j = 0, \pm 1, \pm 2, \ldots,$$

onde $e^{iz} = \cos z + i \sin z$.

Proposição 2.1 *Seja $\{x_j\}_{j=1}^{\infty}$ uma sequência completa e seja $f \in L^2[a,b]$. Defina $\beta_j = \langle f, x_j\rangle$, $j = 1, 2, \ldots$, então $\sum_{j=1}^{K} \beta_j x_j$ é melhor aproximação de f no sentido que $\|f - \sum_{j=1}^{K} \beta_j x_j\| \leq \|f - \sum_{j=1}^{K} \alpha_j x_j\|$ para todo $\alpha = (\alpha_1, \ldots, \alpha_K)^t \in \mathbb{R}^K$. Mais ainda, $\|f - \sum_{j=1}^{K} \beta_j x_j\|^2 \to 0$, quando $K \to \infty$.*

Esta proposição diz simplesmente que podemos aproximar f por $\sum_{j=1}^{K} \beta_j x_j$ suficientemente bem (em norma) quando $K \to \infty$ e que f e $\sum_{j=1}^{K} \beta_j x_j$ são dois elementos identicos em $L^2[a,b]$. (Note que a convergência pontual não válida). As componentes do vetor β são chamadas de coeficientes generalizados de Fourier e a expansão de f em série é chamada de série generalizada de Fourier. Devido a igualdade de Parseval, $\sum_{j=0}^{\infty} |\beta_j|^2 = \|f\|^2$, podemos dizer que se $\{\beta_j\}$ é uma sequência de quadrado integrável ($\sum_{j=0}^{\infty} \beta_j^2 < \infty$) então qualquer $f \in L^2[a,b]$ pode ser escrita como $f = \sum_{j=0}^{\infty} \beta_j x_j$, onde $\{x_j\}$ é uma sequência ortonormal completa para $L^2[a,b]$. A afirmação reciproca também é verdadeira.

2.2 Estimadores por Séries de Fourier

Vamos assumir que $\{x_j\}$ é uma sequência ortonormal completa de funções em $L^2[a,b]$ e que a $\sum_{j=0}^{K} \beta_j x_j(t)$ convirja para $f(t)$ uniformente em t quando $K \to \infty$. Isto não é uma restrição forte pois note que se $f \in C^1[a,b],\ f(a) = f(b),\ f'(a) = f'(b)$ então a série de Fourier converge uniformente em $t \in [a,b]$.

Assuma o seguinte modelo linear com infinito número de parâmetros, isto é,

$$Y_j = \sum_{m=1}^{\infty} \beta_m x_m(t_j) + \varepsilon_j,\ \ j = 1, \ldots, n.$$

Agora desde que $f \in L^2[a,b]$, sabemos que $\sum_{m=1}^{\infty} \beta_m^2 < \infty$ e daípodemos concluir que existe um inteiro positivo K tal que, qualquer seja $N > K$, $\beta_N \to 0$. Portanto, passamos a ter um número finito de parâmetros,

$$Y_j = \sum_{m=1}^{K} \beta_m x_m(t_j) + \varepsilon_j,\ \ j = 1, \ldots, n.$$

Assumiremos também que $t_j = a + (b-a)(j-1)/n - 1,\ \ j = 1, \ldots, n$ e então podemos estimar os coeficientes de Fourier analogamente a maneira que fazemos quando no caso de regressão paramétrica, isto é, minimizamos o erro quadrático médio que nesta escolha de t_j é dado por,

$$EQM(K) = \frac{1}{n} \sum_{j=1}^{n} (y_j - \sum_{m=1}^{K} \beta_m x_m(t_j))^2.$$

Defina a matriz de planejamento $X_K = (x_j(t_i),\ i = 1, \ldots, n,\ j = 1, \ldots, K)$. Se X_K é de posto completo então o minimizador do erro quadrático médio é o vetor de parâmetros $\beta_K = (\beta_{1K}, \ldots, \beta_{KK})^t = (X_K^* X_K)^{-1} X_K^* Y$, onde $Y = (Y_1, \ldots, Y_n)$ é o vetor de observações e X_K^* é a matriz conjugada de X_K. Daí, o estimador da série de Fourier generalizada de $f(t)$ é,

$$f_K(t) = \sum_{j=1}^{K} \beta_{jK} x_j(t)$$

Embora a estratégia para se obter um estimador de f seja bastante intuitiva, problemas podem ocorrer sob o ponto de vista da inferência assintótica. Não há garantia que este tipo de estimador seja pelo menos consistente.

Defina a função perda por estimar f usando como estimador f_K por

$$L_n(K) = \frac{1}{n} \sum_{j=1}^{n} (f(t_j) - f_K(t_j))^2$$

e consequentemente,

$$EL_n(K) = R_n(K) = \frac{1}{n} \sum_{j=1}^{n} (f(t_j) - Ef_K(t_j))^2 + \frac{\sigma^2 K}{n}$$

Dizemos que f_K é consistente se $R_n(K) \to 0$, quando $n \to \infty$. Observe que o primeiro termo de $R_n(K)$ é a soma de quadrados do viés. Intuitivamente, o viés deve ir a zero a medida que K cresce para infinito. Porém, quanto maior é o valor de K, maior é o número de termos da série generalizada de Fourier e assim, como já vimos anteriormente, f_K se aproxima de f. Em outras palavras, $R_n \to 0$ se $K \to \infty$, quando $n \to \infty$. Mas note que o segundo termo de R_n, $\frac{\sigma^2 K}{n}$, para decair a zero precisa necessariamente que K cresça numa taxa menor do n cresce. Isto é importante sob o ponto de vista computacional pois sugere uma maneira de ir incrementando o número de termos da série de Fourier generalizada, se o procedimento de estimação for iterativo. Observe que é imperativo obter uma estimativa do parâmetro K, pois é ele que governa a suavidade do estimador. Uns dos mais comuns métodos de estimação

de K é dado pelo critério GCV (Generalized Cross-Validation)dado por,

$$GCV_n(K) = \frac{1}{n}\sum_{j=1}^{n}\frac{(y_j - f_K(t_j))^2}{(1-\frac{K}{n})^2}$$

Sob certas condições o método GCV produz estimativas assintóticamente ótimas.

3 Estimadores por Kernel

Estimadores por Kernel são estimadores lineares de f da forma

$$f_h(t) = \sum_{j=1}^{n} K(t, t_j, h)Y_j$$

Em outras palavras, estimadores por Kernel são estimadores lineares cujos os pesos provem de uma única função indenpendente dos pontos $(t_1, \ldots, t_n)$. Por exemplo, no caso em que temos os pontos t_j igualmente espaçados

$$f_h(t) = \sum_{j=1}^{n} y_j K_h(t - t_j)$$

onde o Kernel é o Kernel de Dirichlet, $K_h(u) = \sin(2\pi u/\lambda)\sin(\pi u)$. Em geral, usa-se um kernel que satisfaz as mesmas condições para que uma função seja uma função de densidade de probabilidade. O parâmetro h é chamado de parâmetro de suavização, ou algumas vezes usando o nome em inglês "bandwidth", ou "smoothing parameter". Alguns exemplos de kernel são, unforme,

$$K(u) = \begin{cases} 1 & \text{se } |u| \leq 1 \\ 0 & \text{se } |u| > 1 \end{cases}$$

triangular,

$$K(u) = \begin{cases} 1 - |u| & \text{se } |u| \leq 1 \\ 0 & \text{se } |u| > 1 \end{cases}$$

Para entendermos o sentido do parâmetro h, tomemos por exemplo o Kernel de Dirichlet. Se $K(u)$ tem suporte em $[-1, 1]$ então $K(u/h)$ tem suporte em $[-h, h]$. Daío

parâmetro h ("bandwidth") determina, entre outras coisas, quanto das observções podem estar afastadas de t e ainda contribuir para a estimação de $f(t)$. Pequenos valores de h resultarão em uma estimativa não suave, enquanto que valores grandes de h produzirão uma estimativa muito suave para a curva de regressão f. Note que se K é uma função probabilidade de densidade em $[-1, 1]$, então h é o parâmetro de escala no sentido estatístico do termo. Em relação ao tipo de kernel a ser usado, na literatura tem aparecido com mais frequência quele desenvolvido por Watson (1964).

$$f_h(t) = \frac{\sum_{j=1}^{n} K(\frac{t-t_j}{h}) Y_j}{\sum_{j=1}^{n} K(\frac{t-t_j}{h})}$$

3.1 Escolhendo o Parâmetro de Suavização

Vimos que o parâmetro de suavização é muito importante na determinação do estimador da curva de regressão quando se usa o método de estimação por kernel. Existem várias maneiras de se obter uma estimativa para h, mas tem sido a mais utilizada aquela baseada no critério de validação cruzada, na verdade o critério de validação cruzada generalizada é o mais usado na estimação do parâmetros de suavização seja qual for o método de estimação da função de regressão. O critério de validação cruzada generalizada, GCV tem a seguinte forma quando a estimação de f é feita pelo método de kernel,

$$GCV(h) = \frac{EQM(h)}{n^{-1} tr(I - H(h))},$$

onde $H(h) = (X_h^t X_h)^{-1} X_h^t$. Especificamente, procura-se o valor de h que miminize a função objetivo dada por $GCV(h)$. Na figura abaixo vemos o efeito do parâmetro de suavização h no ajuste de curvas por kernel. O kernel usado foi o Gaussiano, X tem distribuição normal padrão e o vetor de erros é distribuido como uma normal de média zero e desvio padrão 0.1.

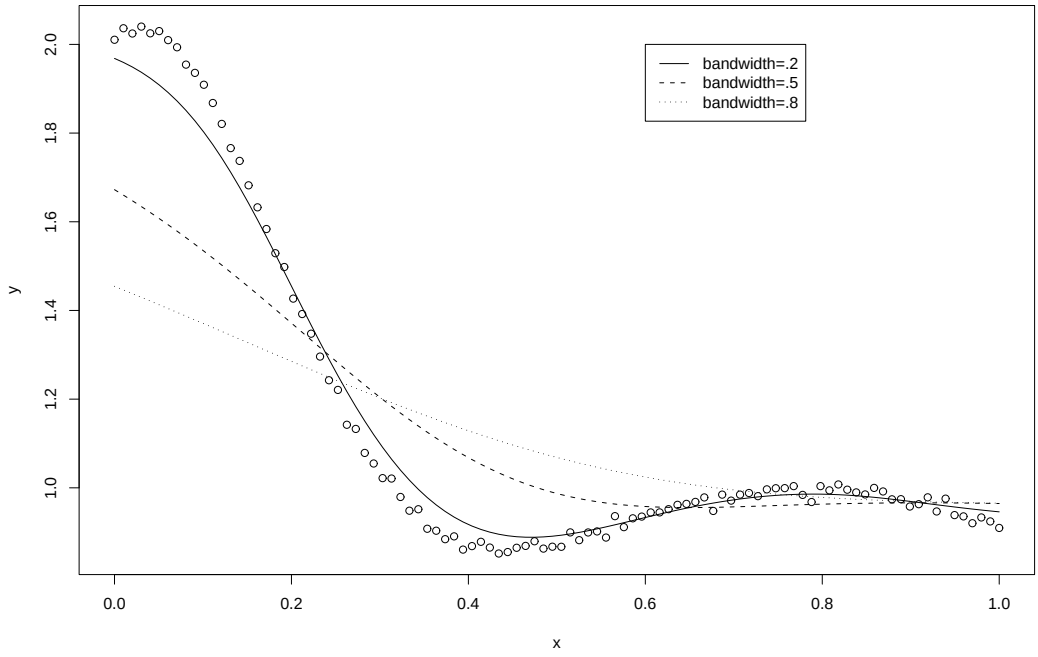

Figure 3.1: Estimativas por kernel Gaussiano

4 Obtenção de Estimadores por Splines

Aqui, surge a pergunta mais natural ao leitor que se inicia ao estudo dos métodos de regressão não paramétrica que é : Afinal, o que um spline? Deixando de lado a etmologia da palavra, passamos a definição matemática para splines.

Definição 4.1 *Um spline de ordem m com knots em $\xi_1, \ldots, \xi_k$ é qualquer função da forma*

$$s(t) = \sum_{i=0}^{m-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t - \xi_i)_+^{m-1},$$

onde $(x)_+ = \max(0, x)$ e os coeficientes $\theta_0, \ldots, \theta_{m-1}$, $\delta_1, \ldots, \delta_k$ são números reais.

Da definição acima notamos que um spline satisfaz:

1. s é um polinômio por partes de ordem m em qualquer subintervalo $[\xi_i, \xi_{i+1})$.

2. s tem $m - 2$ derivadas contínuas.

3. s tem a $(m-1)$-ésima derivada é uma função escada com saltos em $\xi_1, \ldots, \xi_k$.

Assim, um spline é um polinômio por partes cujos os diferentes segmentos são unidos nos knots $\xi_1, \ldots, \xi_k$ no sentido de garantir continuidade.

Seja $S^m(\xi_1, \ldots, \xi_k)$ o conjunto formado por todos os splines definidos em 4.1. Então $S^m(\xi_1, \ldots, \xi_k)$ é um espaço vetorial cuja dimensão é $m + k$, desde que as funções $1, t, \ldots, t^{m-1}, (t-\xi_1)_+^{m-1}, \ldots, (t-\xi_k)_+^{m-1}$ são linearmente independente.

De particular importância temos os splines naturais de ordem $m = 2p$ e $k = n$, i.e., um spline natural é um spline de ordem $2p$ com knots em $a \leq t_1, \ldots, t_n \leq b$. O nome spline natural vem do fato das chamadas condições naturais de fronteira, $s^{(j)}(a) = s^{(j)}(b) = 0$, $j = m, \ldots, 2m-1$. Note que temos mais uma condição que um spline tem que satisfazer: 4. s é um polinômio de ordem p fora do intervalo $[t_1, t_n]$.

Seja $NS^{2p}(t_1, \ldots, t_n)$ o conjunto de todos os splines naturais de ordem $2p$ com knots $a \leq t_1, \ldots, t_n \leq b$, então este conjunto de funções forma um subespaço do espaço vetorial $S^{2p}(t_1, \ldots, t_n)$ obtido por inserir $2p$ restrições lineares que provêm da propriedade 4. Assim para que um spline seja um spline natural temos que ter $\theta_p = \ldots, \theta_{2p-1} = 0$ em 4.1, desde que s deve ser um polinômio de ordem p para $t \notin [t_1, t_n]$. Pode-se verificar que a dimensão de $NS^{2p}(t_1, \ldots, t_n)$ é n.

4.1 Suavização por Splines e Regressão Polinomial

O objetivo desta seção é dar motivação para o estudo do estimador de f por splines. Para isso considere o modelo de regressão:

$$y_j = f(t_j) + \varepsilon_j, \quad j = 1, \ldots, n,$$

onde $a \leq t_1, \ldots, \leq t_n \leq b$ e o vetor ε do erros não correlacionados tem média zero e variância comum σ^2. Para motivar o conceito de suavização por splines, suponha que $f \in W_2^m[a,b] = \{f : f$ é absolutamente contínua e $\int_a^b (\ddot{f})^2$. Queremos encontrar um estimador para a curva de regressão f que ajuste bem os dados e que ao mesmo tempo tenha um certo grau de suavidade.

Uma medida natural de suavidade associada a funções em $W_2^m[a,b]$ é $\int_a^b (f^{(m)}(t))^2 dt$, enquanto uma medida natural para bondade de ajuste seria $\frac{1}{n}\sum_{j=1}^n (y_j - f(t_j))^2$. Então a idéia é combinar esses dois critérios afim de se obter uma medida geral que englobe tanto suavidade quanto bondade de ajuste. Para tanto, considere o seguinte critério

$$CPL_\alpha(f) = (1-\alpha)\sum_{j=1}^{n}(y_j - f(t_j)^2 + \alpha \int_a^b (f^{(m)}(t))^2 dt, \tag{4.1}$$

para $\alpha \in (0,1)$. Assim um estimador optimal poderia ser encontrado por minimizar 4.1 sobre $f \in W_2^m[a,b]$. Entretanto, note que se tomarmos $\lambda = \alpha/(1-\alpha)$, teremos um funcional equivalente conhecido como soma de quadrados penalizada dado pela seguinte expressão:

$$PL_\lambda(f) = \sum_{j=1}^{n}(y_j - f(t_j))^2 + \lambda \int_a^b (f^{(m)}(t))^2 dt, \quad \lambda > 0.$$

O estimador f_λ da função de regressão é aquele que minimiza 4.1 com $f \in W_2^m[a,b]$ e é chamado de estimador de suavização por splines (smoothing spline estimator).

O parâmetro λ é de grande importância neste critério de otimização, pois é ele quem controla o balanço entre suavidade e bondade de ajuste do estimador. Para valores de λ grande, a suavidade tem maior influência no estimador, enquanto que para valores pequenos de λ, a bondade de ajuste predomina no critério de soma de quadrados penalizada, daíλ ser chamado de parâmetro de suavização.

Suavização por splines teve sua origem com Whittaker (1923) e foi considerado até os anos 70 como uma ferramenta essecialmente de uso para analista númericos, quando então apareceram os trabalhos de Grace Wahba (Craven and Wahba, 1979; Wahba, 1990) mostrando de maneira pioneira que este método tem grande utilidade na análise estatística, principalmente, no caso de regressão não paramétrica.

Regressão polinomial esta bastante relacionada com o procedimento descrito acima. Para se ver isto, assuma que o estimador de f dado por f_λ seja bem definido. Desde

que $f \in W_2^m[a,b]$, segue que existem constantes, $\theta_0, \ldots, \theta_{m-1}$ tal que

$$f(t) = \sum_{i=0}^{m-1} \theta_i t^i + \text{resto}(t),$$

com

$$\text{resto}(t) = \frac{1}{(m-1)!} \int_a^b f^{(m)}(x)(t-x)_+^{m-1} dx.$$

Assim o modelo $y_j = f(t_j) + \varepsilon_j$ é equivalente ao modelo

$$y_j = \sum_{i=0}^{m-1} \theta_i t^i + \text{resto}(t) + \varepsilon_j \ , j = 1, \ldots, n.$$

Se o resto(t_j) é proximo de zero, então parece razoável estimar f usando um estimador de regressão polinomial de ordem m. Caso contrário, tem-se dúvidas sobre a aplicabilidade deste modelo polinomial.

Veremos agora como o critério 4.1 esta relacionado com o modelo de regressão polinomial. Seja o funcional $J_m(f) = \int_a^b (f^{(m)}(t))^2 dt$, então aplicando a desigualdade de Cauchy-Schawartz no resto(t) temos:

$$\text{resto}(t)^2 \leq \frac{(b-a)^{2m-1}}{[(2m-1)(m-1)!]^2} J_m(f).$$

Note que existe uma constante que depende de m mas não de f de tal modo que $\sup_{t\in[a,b]} |\text{resto}(t)| \leq C J_m(f)^{1/2}$. Em particular, $|\text{resto}(t_j)| \leq C J_m(f)^{1/2}$, $j = 1, \ldots, n$. Desta forma, o tamanho de $J_m(f)$ esta ligado ao termos resto$(t_1), \ldots,$ resto(t_n). Pois, se $J_m(f)$ é pequeno, i.e., $J_m(f) \leq \rho$, tal que $\rho > 0$, teremos grande credibilidade no modelo polinomial. Por exemplo, seja $m = 2$, daí$J_2(f) \equiv 0$ se e somente se f é linear. A condição $J_2(f) \leq \rho$ governa o quão longe f pode se afastar do modelo linear. Assim se apriori sabe-se que ρ é pequeno então isto indicará que f é aproximadamente linear. Observe que f'' representa a taxa de variação da inclinação da curva de regressão.

Podemos incluir a condição acima para determinar um procedimento de estimação. Por exemplo, encontre $f \in W_2^m[a,b]$ tal que minimize $EQM(f) = \frac{1}{n}\sum_{j=1}^n (y_j - f(t_j))^2$ com a condição que $J_m(f) \leq \rho$. Usando multiplicadores de Lagrange, tempos que encontrar $f_\lambda \in W_2^m[a,b]$ tal que $EQM(f) + \lambda(J_m(f) - \rho)$ seja mínimo. Note que este critério é em essência o mesmo que 4.1.

4.2 A Forma do Estimador

Pode-se mostrar que se $\phi_1, \ldots, \phi_n$ é uma base para o espaço $NS^{2m}(t_1, \ldots, t_n)$, com $n \leq m$, e $m \in \mathbb{N}$, então exite para $0 < \lambda < \infty$ fixo, um único estimador f_λ sob o critério 4.1 quando $f \in W_2^m[a, b]$. Mais ainda, $f_\lambda \in NS^{2m}(t_1, \ldots, t_n)$ e consequentemente este estimador pode ser escrito com uma combinação linear de $\phi_1, \ldots, \phi_n$. Isto é, $f_\lambda = \sum_{j=1}^{n} \beta_{\lambda j}\phi_j(t)$ tal que o vetor $\beta_{\lambda j} = (\beta_{\lambda 1}, \ldots, \beta_{\lambda n})^t$ satisfaz a equação $(X^tX + n\lambda\Omega)\beta_\lambda = X^ty$, onde a matriz Ω possui entradas $\Omega_{i,j} = \int_a^b \phi_i''(t)\phi_j''(t)dt$, com $i, j = 1, \ldots, n$. A escolha comum para as bases $\phi_1, \ldots, \phi_n$ é a sequência formada pelos bem conhecidos B-splines (de Boor, 1978).

É importante notar que a estimação do parâmetro de suavização pode ser feita totalmente baseada nos dados, por exemplo, pelo Critério de Validação Cruzada

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}(y_j - f_{\lambda(j)}(t_j))^2,$$

onde $f_{\lambda(j)}(t_j)$ é o estimador provindo do critério 4.1 com a j-ésima observação retirada.

Craven and Wahba (1979), demonstraram que existe um relacionamento entre os procedimentos de validação cruzada e validação cruzada generalizada (GCV). Então, um outro método a ser usado para estimar λ seria por GCV. De fato, o procedimento GCV tem sido, até o momento, o procedimento mais utilizado na estimação de parâmetros de suavização. Abaixo temos um exemplo comparando estimativas obtidas pelo método de suavização por splines com a estimativa do parâmetro de suavização via GCV e estimativas obtidas pelo método de kernel com parâmetros de suavização estimados visualmente. O modelo simulado é $y = \exp(-x^2/2)\cos(4\pi x) + \varepsilon$, onde X tem distribuição normal padrão e ε distribui normalmente com média zero e desvio padrão 0.1.

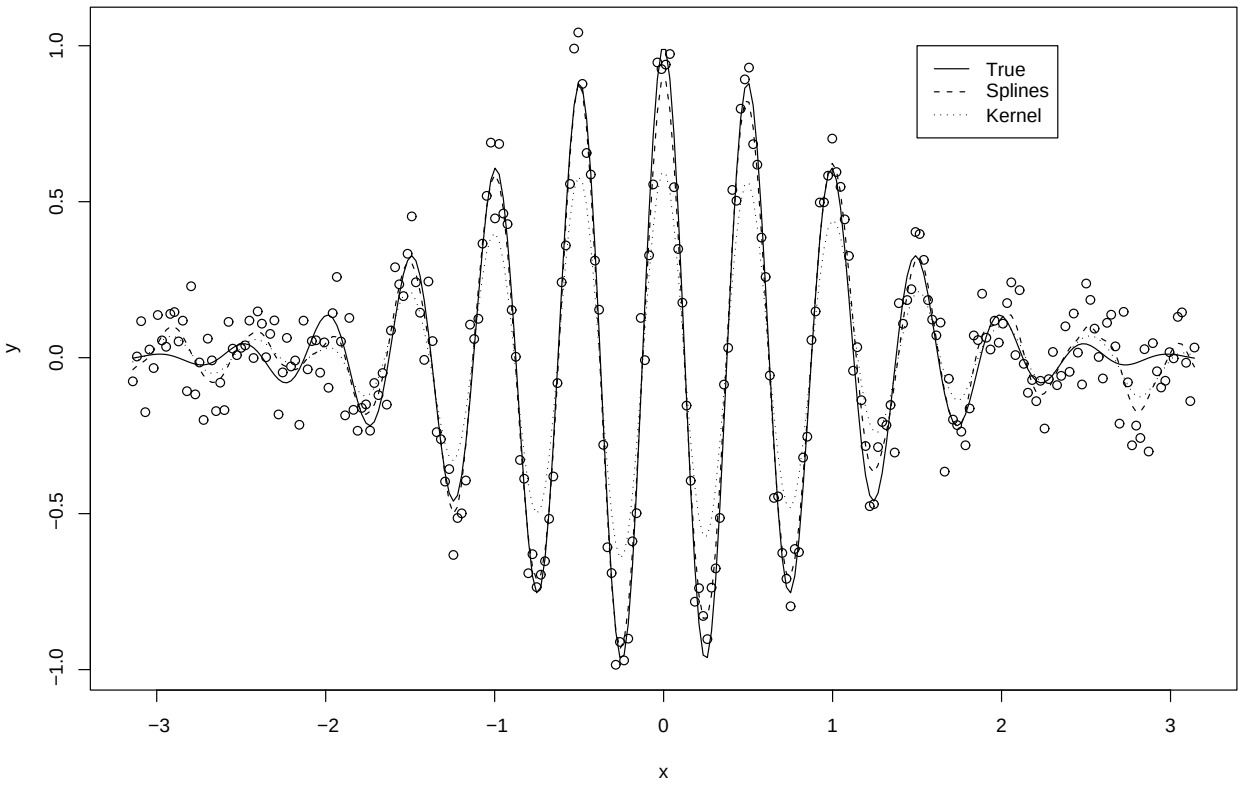

Figure 4.1: Estimativas por splines e por kernel

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

de Boor, C. (1978). *A Practical Guide to Splines*, Springer Verlag, New York.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM:PA.

Watson, G. S. (1964). Smooth regression analysis, *Sankya A* **26**: 359–372.

Whittaker, E. T. (1923). On new method of graduation, *Proc. Edinburgh Math. Soc.* **41**: 63–75.