2
Revisão teórico-metodológica

2.1
Modelos Poisson-gama

Os modelos Poisson-gama foram introduzidos como uma proposta de modelos para lidar com observações de séries temporais de processos de contagem. O problema consiste essencialmente em formular um modelo que forneça a distribuição de y_t dado o passado da série, ou seja, a seqüência $y_1, \ldots, y_{t-1}$ denotada por Y_{t-1}. A solução do problema reside no uso das distribuições conjugadas como as usadas no contexto bayesiano, contudo, a abordagem utilizada é a clássica [13]. Embora o artigo original [20] compreenda um maior número de distribuições não-gaussianas, neste trabalho, apenas a distribuição de Poisson é abordada.

Considere-se a seqüência $y_1, \ldots, y_n$ como n realizações de um processo estocástico de Poisson [29]. Para cada instante t a distribuição de y_t condicionada no nível μ_t é dada por

$$p\left(y_t|\mu_t\right) = \frac{\mu_t^{y_t} e^{-\mu_t}}{y_t!}. \tag{2-1}$$

Suponha-se que a distribuição de μ_{t-1} condicionada em todas as observações da série até o instante $t-1$ seja gama com parâmetros a_{t-1} e b_{t-1} estimados a partir da seqüência Y_{t-1}. Sob normalidade dos erros nas equações de um modelo de nível local gaussiano [21, 9], a média de $\mu_t|Y_{t-1}$ é igual a de $\mu_{t-1}|Y_{t-1}$ e a variância é maior. Este mesmo comportamento pode ser replicado na distribuição gama aplicando aos parâmetros um fator menor que 1, denotado por ω e denominado fator de desconto. Ou seja, assume-se que a distribuição a priori $p\left(\mu_t|Y_{t-1}\right)$ é uma gama com parâmetros $a_{t|t-1}$ e $b_{t|t-1}$ da forma

$$a_{t|t-1} = \omega a_{t-1} \tag{2-2}$$

$$b_{t|t-1} = \omega b_{t-1} \tag{2-3}$$

com $0 < \omega \leq 1$. As equações 2-2 e 2-3 são as equações de previsão do filtro Poisson-gama.

Com a observação y_t disponível, devido à conjugação das distribuições de probabilidades, a distribuição a posteriori $p(\mu_t|Y_t)$ também é gama com parâmetros dados por

$$a_t = \omega a_{t-1} + y_t \tag{2-4}$$

$$b_t = \omega b_{t-1} + 1 \tag{2-5}$$

As equações 2-4 e 2-5 são as equações de atualização do filtro Poisson-gama. As equações de previsão e de atualização deste filtro podem, na prática, ser combinadas. Neste caso, apenas $a_{t|t-1}$ e $b_{t|t-1}$ são estimados [4].

A distribuição de μ_t é difusa se $a = 0$ e $b = 0$. Entretanto, a iniciação das recursões do filtro no instante $t = 0$ com $a_0 = 0$ e $b_0 = 0$ permite a obtenção de uma distribuição própria para μ_t no instante $t = \tau$, em que τ é o índice da primeira observação com valor diferente de zero [20].

Condicionada em Y_τ, a distribuição conjunta de $y_{\tau+1}, \ldots, y_n$ é

$$p(y_{\tau+1}, \ldots, y_n; \omega) = \prod_{t=\tau+1}^{n} p(y_t|Y_{t-1}) \tag{2-6}$$

e a função de densidade de probabilidade preditiva é dada por

$$p(y_t|Y_{t-1}) = \int_0^\infty p(y_t|\mu_t)\, p(\mu_t|Y_{t-1})\, d\mu_t. \tag{2-7}$$

Para observações de um processo de Poisson e uma priori gama, a equação 2-7 leva a distribuição binomial negativa com parâmetros $a_{t|t-1}$ e $b_{t|t-1}$. A função de log-verossimilhança do hiperparâmetro ω a ser estimado é dada por

$$\begin{aligned} \log L(\omega) = \sum_{t=\tau+1}^{n} \{ & \log \Gamma\left(a_{t|t-1} + y_t\right) - \log y_t! - \log \Gamma\left(a_{t|t-1}\right) + \\ & a_{t|t-1} \log b_{t|t-1} - \left(a_{t|t-1} + y_t\right) \log\left(1 + b_{t|t-1}\right)\}. \end{aligned} \tag{2-8}$$

Das propriedades da binomial negativa se obtêm a média e a variância da distribuição preditiva dadas por

$$E(y_t|Y_{t-1}) = \frac{a_{t|t-1}}{b_{t|t-1}} \tag{2-9}$$

$$Var(y_t|Y_{t-1}) = \frac{a_{t|t-1}\left(1 + b_{t|t-1}\right)}{b_{t|t-1}^2} \tag{2-10}$$

Usando substituições sucessivas, verifica-se que a função de previsão L passos à frente do modelo Poisson-gama sem variáveis explicativas equivale a um amortecimento exponencial ponderado (EWMA) das observações passadas com constante de suavização igual a $1 - \omega$. Nos modelos com variáveis explicativas estas formas não são equivalentes [20].

Para introduzir variáveis explicativas no modelo Poisson-gama, considere-se que o efeito do nível do componente estrutural μ_t da série temporal é separado do efeito das covariáveis no vetor x_t. Este nível pode ser combinado de forma multiplicativa com uma função de ligação exponencial das covariáveis, denotada por $\exp\left(\eta_t^+\right)$. Logo, a distribuição de y_t condicionada em μ_t é Poisson com média

$$\mu_t^+ = \mu_t \exp\left(\eta_t^+\right) \tag{2-11}$$

$$\eta_t^+ = \sum_{j=1}^{p} \beta_j x_{jt} \tag{2-12}$$

em que η_t^+ é o preditor linear.

Seja gama a distribuição de μ_{t-1} condicionada em Y_{t-1}. A distribuição de μ_t condicionada em Y_{t-1} também é gama, com parâmetros $a_{t|t-1}$ e $b_{t|t-1}$. As médias de $\mu_{t-1}|Y_{t-1}$ e $\mu_t|Y_{t-1}$ são iguais, porém a variância de $\mu_t|Y_{t-1}$ é maior que a de $\mu_{t-1}|Y_{t-1}$ [20]. Então, as equações 2-2 e 2-3 de previsão do filtro Poisson-gama para o modelo com variáveis explicativas são dadas por

$$a_{t|t-1} = \omega a_{t-1} \tag{2-13}$$

$$b_{t|t-1} = \omega b_{t-1} \exp\left(-\eta_t^+\right) \tag{2-14}$$

e as equações 2-4 e 2-5 de atualização do filtro para o modelo com variáveis explicativas são

$$a_t = \omega a_{t-1} + y_t \tag{2-15}$$

$$b_t = \omega b_{t-1} + \exp\left(\eta_t^+\right) \tag{2-16}$$

com $t = \tau + 1, \ldots, n$.

Os hiperparâmetros ω e β_j são estimados pelo método da máxima verossimilhança cuja função é dada pela equação 2-8. A média e variância da distribuição preditiva do modelo com variáveis explicativas permanecem as mesmas que nas equações 2-9 e 2-10, exceto pelos parâmetros que agora são calculados como nas equações 2-13 e 2-14.

Muitas das técnicas de diagnóstico usualmente empregadas em mo-

delos lineares generalizados (MLG) [28, 12] são válidos para os modelos Poisson-gama. Contudo, observa-se que para o diagnóstico que depende da distribuição deve tomar por referência a distribuição preditiva que é binomial negativa. Como exemplo, pode-se definir a função desvio para os modelos Poisson-gama da seguinte forma

$$D\left(y;\hat{\mu}\right)=2\sum_{t=\tau+1}^{n}\left\{a_{t|t-1}\log\left(\frac{a_{t|t-1}}{y_t b_{t|t-1}}\right)-\left(a_{t|t-1}+y_t\right)\log\frac{\left(y_t+a_{t|t-1}\right)}{\left(1+b_{t|t-1}\right)y_t}\right\}. \tag{2-17}$$

O número de graus de liberdades do modelo ajustado é dado por $n-p-\tau$ [20]. Fazendo uso da equação 2-17, pode ser definido, por exemplo, o resíduo de desvio dado por $r_{d_t}=\text{sign}(y_t-\mu_t)\sqrt{d_t}$ em que d_t é o valor da parcela da função desvio referente ao instante t [28]. Os resíduos de desvio são considerados superiores e mais apropriados para diagnóstico e validação de modelos que usam a abordagem dos modelos lineares generalizados que os resíduos de Pearson [30].

Outra possibilidade é o resíduo de desvio padronizado definido por $r_{dp_t}=r_{d_t}/\sqrt{1-h_{tt}}$, em que r_{d_t} é o resíduo de desvio e a quantidade h_{tt} é a contribuição da t-ésima observação para o valor previsto, ou seja, é o t-ésimo elemento da diagonal da matriz chapéu estimada. A matriz chapéu é equivalente à matriz de projeção dos modelos de regressão linear e não é definida explicitamente nos modelos Poisson-gama. Campos e colaboradores (2003) [4] propõem uma quantidade equivalente para h_{tt} e conduzem um estudo de simulação para investigar a eficácia da padronização dos resíduos usando esta quantidade.

2.2 Regressão não-paramétrica

2.2.1 Splines cúbicas

Nos modelos lineares generalizados [28], a média de uma variável resposta Y é modelada como uma função linear $\sum_{j=1}^{p}\beta_j X_j$ de um conjunto de covariáveis $X_1,\ldots,X_p$. Estes modelos assumem uma forma linear ou paramétrica para o efeito das covariáveis. Os MLG podem ser estendidos, substituindo o preditor linear $\eta=\sum_{j=1}^{p}\beta_j X_j$ por um preditor aditivo $\eta=\sum_{j=1}^{p}g_j\left(X_j\right)$, em que $\{g_j\left(X_j\right)\}$, com $j=1,\ldots,p$, são funções quaisquer das covariáveis $X_1,\ldots,X_p$. Por não possuir restrição na forma

funcional de nenhuma das covariáveis este modelo é dito não paramétrico. Na notação do preditor de ambos os modelos, o intercepto foi omitido por simplificação. Os modelos semi-paramétricos são aqueles nos quais uma ou mais funções $g_j(X_j)$ do preditor aditivo são lineares, ou seja, são da forma $\beta_j X_j$ [22, 23, 3, 16].

Os pressupostos tradicionais dos modelos de regressão são relaxados e o problema agora passa a ser escolher as funções $\{g_j(X_j)\}$ de tal forma que alguma norma seja minimizada. A norma comumente utilizada na análise de regressão é a L_2. Então, é necessário escolher g tal que a soma dos quadrados dos resíduos seja mínima. Apesar de as funções trigonométricas e as funções polinomiais serem mais flexíveis que uma reta, essas ainda definem uma estrutura rígida para a associação entre as covariáveis e a variável resposta. Além disto, uma observação individual pode exercer efeitos imprevisíveis em outras regiões da curva. A escolha natural para funções g são funções suaves estimadas a partir dos próprios dados, tal que a soma de quadrados penalizada seja minimizada [18, 12].

Para estimar g considere-se minimizar o funcional

$$S(g) = \sum_{i=1}^{n} \{Y_i - g(k_i)\}^2 + \lambda \int_a^b \{g''\}^2 dx \tag{2-18}$$

em que k_i, com $i = 1, \ldots, n$, são pontos ordenados num intervalo $[a, b]$ qualquer, g tem primeira e segunda derivadas contínuas g' e g'', o quadrado de g'' é uma função integrável e $0 < \lambda < \infty$ é o parâmetro de suavização da curva g. A solução $\hat{g}_\lambda$ do problema de otimização acima é uma *spline* cúbica natural [11].

Suponha-se que a seqüência de pontos $k_1, \ldots, k_n$ pertença ao intervalo $[a, b]$ tal que $a < k_1 < k_2 < \cdots < k_n < b$. Uma função g definida sobre o intervalo $[a, b]$ é uma *spline* cúbica se satisfaz as seguintes condições: (1) sobre cada intervalo $(a, k_1), (k_1, k_2), (k_2, k_3), \ldots, (k_n, b)$, g é uma função polinomial cúbica e (2) cada dois polinômios em partes vizinhos se conectam no ponto k_i de tal modo que a própria g e sua primeira e segunda derivadas sejam contínuas em todos os pontos k_i e, portanto, sobre todo o intervalo $[a, b]$. Pode ser definido então o espaço $\mathcal{S}[a, b]$ de todas as funções suaves g em $[a, b]$. Os pontos k_i são chamados *nós*[1]. A fim de simplificar a notação, defina-se $k_0 = a$ e $k_{n+1} = b$ os limites do intervalo sobre o qual a função g é definida.

Uma representação natural de um polinômio em partes é da forma de

[1] do termo em inglês *knots*.

quatro coeficientes polinomiais

$$g(x) = d_i (x - k_i)^3 + c_i (x - k_i)^2 + b_i (x - k_i) + a_i \qquad (2\text{-}19)$$

para $k_i \leq x \leq k_{i+1}$ e constantes a_i, b_i, c_i, d_i com $i = 0, \ldots, n$. Uma *spline* cúbica no intervalo $[k_0, k_{n+1}]$ é dita *spline* cúbica natural se as segunda e terceira derivadas nos pontos k_0 e k_{n+1} são iguais a zero. A implicação destas condições é que $d_0 = 0$, $c_0 = 0$, $d_n = 0$ e $c_n = 0$, logo g é linear nos intervalos $[k_0, k_1]$ e $[k_n, k_{n+1}]$ [18].

Uma representação mais eficiente do ponto de vista computacional e matemático que aquela na equação 2-19 é a representação do valor da segunda derivada. Nesta representação, uma *spline* cúbica natural g é completamente especificada pelo seu valor e o valor da segunda derivada em cada nó k_i. Supondo que g é uma *spline* cúbica natural com nós $k_1, \ldots, k_n$, defina-se $g_i = g(k_i)$ e $\gamma_i = g''(k_i)$ para $i = 1, \ldots, n$. Uma *spline* cúbica natural g tem segunda derivada nos pontos k_1 e k_n igual a zero, logo $\gamma_1 = 0$ e $\gamma_n = 0$. Considere-se os vetores $g = (g_1, \ldots, g_n)'$ e $\gamma = (\gamma_2, \ldots, \gamma_{n-1})'$. Os valores de g e de suas derivadas em qualquer ponto x podem ser calculados explicitamente em termos dos vetores g e γ. Deste modo g pode ser descrita em um gráfico com qualquer grau de precisão.

A condição necessária e suficiente para que os vetores g e γ representem uma autêntica *spline* cúbica natural para uma dada seqüência de nós depende de duas matrizes R e Q. A matriz Q tem dimensão $n \times (n-2)$ com elementos q_{ij}, com $i = 1, \ldots, n$ e $j = 2, \ldots, n-1$. Os elementos de Q têm a seguinte forma

$$\begin{aligned} q_{j-1,j} &= h_{j-1}^{-1} \\ q_{jj} &= -h_{j-1}^{-1} - h_j^{-1} \\ q_{j+1,j} &= h_j^{-1} \\ q_{ij} &= 0 \text{ se } \mid i - j \mid \geq 2 \end{aligned} \qquad (2\text{-}20)$$

com $h_i = k_{i+1} - k_i$.

A matriz R é simétrica e tem dimensão $(n-2) \times (n-2)$ com seus elementos dados por

$$\begin{aligned} r_{ii} &= (1/3)(h_{i-1} + h_i) \\ r_{i,i+1} &= (1/6) h_i \\ r_{i+1,i} &= (1/6) h_i \\ r_{ij} &= 0 \text{ se } \mid i - j \mid \geq 2 \end{aligned} \qquad (2\text{-}21)$$

com $i = 2, \ldots, n-1$ e $j = 2, \ldots, n-1$. A matriz R é estritamente positiva definida.

Com as matrizes R e Q definidas, pode-se enunciar um dos teoremas que formam a base da interpolação e da suavização por *splines*. As provas desses teoremas podem ser consultadas em Green e Silverman (1985) [18].

Teorema 2.1 *Os vetores g e γ especificam uma* spline *cúbica natural se e somente se*

$$Q'g = R\gamma \tag{2-22}$$

Se a condição acima é satisfeita, então o termo de penalização em 2-18 satisfaz

$$\int_a^b g''(x)^2 \, dx = \gamma' R\gamma = g'QR^{-1}Q'g. \tag{2-23}$$

A *spline* de interpolação tem como motivação mecânica um antigo dispositivo usado para desenhar cascos de navios e trilhos de linhas férreas. Considere-se que para cada nó k_i existe um ponto (k_i, z_i). Considere-se também uma peça de madeira ou metal flexível forçada a passar pelos pivôs fixos nos pontos dados (k_i, z_i) – nos nós k_i – e livre para tomar qualquer forma nos outros pontos. Com os pivôs presos nos nós, a lâmina toma a forma de mínima energia sujeita às restrições nos nós [18].

A fim de simplificar o entendimento da suavização por *spline*, considere-se a interpolação por *spline*. Seja $\mathcal{S}[a, b]$ o espaço de todas as funções g suaves no sentido de que possuem primeira e segunda derivadas contínuas. A curva mais suave em $\mathcal{S}[a, b]$ para interpolar os pontos dados é a que tem menor termo de penalização $\int g''^2$ entre todas as curvas que interpolam os dados.

Entre todas as curvas g em $\mathcal{S}[a, b]$ que interpolam os pontos (k_i, z_i), aquela que minimiza $\int g''^2$ é uma *spline* cúbica natural com nós em k_i. Se $n \geq 2$, então existe uma única *spline* cúbica natural que interpola os dados. Assim, o problema de minimizar o termo de penalização $\int g''^2$ é equivalente a encontrar uma única *spline* cúbica natural com nós k_i e valores $g(k_i) = z_i$ para todo i. Logo, uma *spline* cúbica natural é a solução de um sistema de equações lineares. O segundo teorema trata da unicidade da *spline* cúbica natural de interpolação.

Teorema 2.2 *Suponha-se $n \geq 2$ e $k_1 < \cdots < k_n$. Dados os valores $z_1, \ldots, z_n$, existe uma e apenas uma* spline *cúbica natural g com nós nos pontos k_i que satisfaz*

$$g(k_i) = z_i \tag{2-24}$$

para $i = 1, \ldots, n$.

A *spline* cúbica natural de interpolação é ótima em uma classe ainda maior de funções suaves. Seja $\mathcal{S}_2[a,b]$ o espaço das funções contínuas e com primeira derivada contínua g' sobre o intervalo $[a,b]$. Isto implica a existência de uma função g'' integrável tal que $\int_a^x g''(k)\,dk = g'(x) - g'(a)$ para todo $x \in [a,b]$. Este resultado é garantido pelo terceiro teorema.

Teorema 2.3 *Suponha-se $n \geq 2$ e que g é uma* spline *cúbica natural de interpolação com valores $z_1, \ldots, z_n$ nos pontos $k_1, \ldots, k_n$ satisfazendo $a < k_1 < \cdots < k_n < b$. Seja $\tilde{g}$ uma função em $\mathcal{S}_2[a,b]$ tal que $\tilde{g}(k_i) = z_i$ para $i = 1, \ldots, n$. Então $\int \tilde{g}''^2 = \int g''^2$. A igualdade só é satisfeita se $\tilde{g}$ e g são idênticas.*

Nas aplicações estatísticas, o que se deseja é estimar uma curva cujos valores observados são realizações de uma variável aleatória, ou seja, sujeitos a erros aleatórios. Neste caso, o objetivo é obter uma curva g que suaviza os dados observados. Tal como no problema de interpolação, considere-se $k_1, \ldots, k_n$ pontos pertencentes ao intervalo $[a,b]$ tal que $a < k_1 < \cdots < k_n < b$. Sejam $y_1, \ldots, y_n$ observações de uma variável aleatória. A fim de garantir as condições do teorema 2.1, considere-se $n \geq 3$ [18]. Dada uma função g em $\mathcal{S}_2[a,b]$, seja $S(g)$ a soma de quadrados penalizada como definida na equação 2-18. A curva $\hat{g}$ estimada será aquela que minimiza $S(g)$ entre todas as funções do espaço $\mathcal{S}_2[a,b]$.

Aplicando as propriedades das *splines* de interpolação, pode ser mostrado que a curva estimada $\hat{g}$ é uma *spline* cúbica natural com nós nos pontos k_i. Reescrevendo $S(g)$ em função dos vetores g e γ e das matrizes R e Q é possível concluir que a função ótima $\hat{g}$ existe e é única.

Seja $Y = (Y_1, \ldots, Y_n)'$. A soma de quadrados penalizada $S(g)$ pode ser reescrita na forma matricial

$$\begin{aligned} S(g) &= (Y-g)'(Y-g) + \lambda g' Q R^{-1} Q' g \\ &= g'\left(I + \lambda Q R^{-1} Q'\right) g - 2Y'g + Y. \end{aligned} \tag{2-25}$$

Fazendo $K = QR^{-1}Q'$, λK é não-negativa definida e portanto, $(I + \lambda K)$ é estritamente positiva definida. Logo, a função na equação 2-25 tem um único mínimo obtido pela expressão

$$g = (I + \lambda K)^{-1} Y. \tag{2-26}$$

O teorema 2.2 garante que o vetor g define unicamente uma *spline* g. Então, $S(g)$ tem um único mínimo dado pela equação 2-26 sobre o espaço de todas as *splines* cúbicas naturais com nós nos pontos k_i.

Teorema 2.4 *Suponha-se* $n \geq 3$ *e que* $k_1, \ldots, k_n$ *sejam pontos tais que* $a < k_1 < \cdots < k_n < b$. *Dados os pontos* $Y_1, \ldots, Y_n$ *e o parâmetro de suavização* λ *estritamente positivo e seja* $\hat{g}$ *a* spline *cúbica natural com nós em* $k_1, \ldots, k_n$ *tal que* $g = (I + \lambda K)^{-1} Y$. *Então, para qualquer* $g \in \mathcal{S}_2[a, b]$,

$$S(\hat{g}) \leq S(g). \tag{2-27}$$

A igualdade só é satisfeita se g *e* $\hat{g}$ *são idênticas.*

Alguns algoritmos para encontrar $\hat{g}$, estimativa da curva g, estão descritos em detalhes em Green e Silverman (1985) [18].

A partir da equação 2-26 e considerando a natureza quadrática da equação 2-18, pode ser mostrado que $\hat{g}$ é linear nas observações [33, 18], no sentido que existe uma matriz $H(\lambda)$, tal que

$$\hat{g} = H(\lambda)\, y \tag{2-28}$$

e

$$H(\lambda) = (I + \lambda K)^{-1}. \tag{2-29}$$

Considere-se na regressão linear uma matriz H tal que $\hat{y} = Hy$ e $H = X(X'X)^{-1}X'$. A matriz de suavização $H(\lambda)$ tem um papel equivalente à matriz chapéu H da regressão linear, pois mapeia os valores observados y_i nos valores previstos $\hat{g}(k_i)$. Entretanto, $H(\lambda)$ não pode ser interpretada como uma matriz de projeção [10].

Por analogia, podem ser estentidas à matriz $H(\lambda)$ as propriedades básicas da matriz chapéu da regressão linear. Denote-se $H(\lambda)$ a matriz chapéu da *spline* de regressão e os elementos $h_{ii}(\lambda)$ da diagonal principal os valores de influência. Os elementos de $H(\lambda)$ têm a mesma interpretação que aqueles de H na regressão linear [10].

A fim de derivar as propriedades básicas de $H(\lambda)$, considere-se uma matriz T de dimensão $n \times m$ com os elementos t_{ij} iguais a t_i^j, respectivamente, com $i = 1, \ldots, n$ e $j = 0, \ldots, m-1$ e defina-se

$$H^* = T(T'T)^{-1}T'. \tag{2-30}$$

A matriz H^* é conhecida como a matriz chapéu da regressão polinomial. O teorema a seguir mostra as propriedades da matriz de suavização.

Teorema 2.5 *A matriz* $H(\lambda) = \{h_{ij}(\lambda)\}$ *satisfaz as seguintes propriedades:*

$$0 \leq h_{ii}(\lambda) \leq 1 \tag{2-31}$$

$$-1 \leq h_{ij}(\lambda) \leq 1 \tag{2-32}$$

para $i \neq j$

$$h_{ii}(\lambda) = 1 \text{ se e somente se } h_{ij}(\lambda) = 0$$

para todo $i \neq j$ *e* $\sum_{j=1}^{n} h_{ij}(\lambda) = 1$. *Ela e fortemente correlacionada com* $H^* = \{h_{ij}^*\}$ *no sentido que*

$$h_{ii}(\lambda) \downarrow h_{ii}^* \text{ como } \lambda \to \infty \text{ se } h_{ii}^* \neq 1$$

Ainda, $h_{ij}(\lambda) \to h_{ij}^*$ *como* $\lambda \to \infty$ *e para* λ *suficientemente grande com* $h_{ij}^* \neq 0$, *tanto* $h_{ij}(\lambda)$ *quanto* h_{ij}^* *têm o mesmo sinal. Se* $\lambda \to 0$ *e* $h_{ii}^* \neq 1$, *então* $h_{ii}(\lambda) \uparrow 1$.

A prova deste teorema pode ser obtida em Eubank (1984) [10].

Seja $e = (I - H)\, y$ o vetor de resíduos de um modelo na regressão linear usual. Ainda por analogia, pode ser definido um vetor e_λ tal que $e_\lambda = (I - H(\lambda))\, y$ e $Var(e_\lambda) = \sigma^2 (I - H(\lambda))$. E, como resultado do teorema 2.5, os elementos da matriz $H(\lambda)$ podem ser utilizados como ferramenta de diagnóstico tal como a matriz chapéu dos modelos de regressão linear [10, 1].

As *splines* cúbicas também podem ser ponderadas e, neste caso, é atribuído um peso w_i para cada observação y_i. Esta abordagem é especialmente importante quando alguns pontos do conjunto de dados tem grande influência sobre os valores previstos $\hat{g}(k_i)$. Estimar a função g, agora, consiste em minimizar o funcional

$$S_W(g) = \sum_{i=1}^{n} w_i \{Y_i - g(k_i)\}^2 + \lambda \int_a^b \{g''\}^2 dx \tag{2-33}$$

em que w_i com $i = 1, \ldots, n$. Se $n \geq 3$ e λ e os pesos w_i são estritamente positivos, então a função na equação 2-33 tem um único mínimo dado por

$$g = (W + \lambda K)^{-1} WY \tag{2-34}$$

em que W é uma matriz diagonal de dimensão $n \times n$ cujos elementos são os pesos w_i com $i = 1, \ldots, n$ [18].

As *splines* cúbicas podem ser generalizadas para polinômios de ordens mais elevadas se introduzindo condições nas derivadas de ordens superiores. A idéia de suavização por *splines* pode ser estendida para problemas de dimensão superior [39]. Entre as opções de estimação podem ser consideradas, por exemplo, a redução da dimensionalidade usando funções aditivas ajustadas de forma iterativa [22, 23] ou *thin plate splines*, na qual toda a hiper-superfície é ajustada de uma só vez [38, 37, 36].

2.2.2 Seleção do parâmetro de suavização

O parâmetro de suavização é denotado por λ e controla a contribuição do termo $\int g''^2$ para $S(g)$. Um dos problemas na estimação de g reside na escolha do valor de λ com melhor relação viés $\times$ variância. Existem duas abordagens filosóficas para a escolha do parâmetro de suavização. Em alguns contextos, o parâmetro λ pode ser selecionado de forma empírica e subjetiva. Em outros casos, o parâmetro de suavização pode ser selecionado por um método automático. Então, os próprios dados determinam o valor de λ. O valor selecionado de forma automática pode ser também usado como valor inicial para um ajuste fino manual do parâmetro de suavização.

No processo de seleção do parâmetro de suavização, é necessário minimizar uma medida global de erro como, por exemplo, a média do erro quadrático médio. O método mais comum para a seleção automática do parâmetro de suavização é a validação cruzada. Este método é motivado em termos de erro de previsão. Supondo um erro com média zero, a curva g tem a propriedade de que, dada uma observação y_k, $g(y_k)$ é a melhor previsão de y_k em termos de erro quadrático médio. Então, é razoável escolher o estimador $\hat{g}(k)$ tal que este dê o menor valor de $\{y_k - \hat{g}(y_k)\}^2$ para uma nova observação y_k no ponto k. Na prática, como não há novas observações disponíveis, a validação cruzada reproduz o efeito de uma nova observação y_k removendo a observação y_i referente ao ponto k_i do conjunto de dados [18, 17, 23].

Seja y_i a observação referente ao ponto k_i. Considere-se que y_i é uma nova observação omitindo-a do conjunto de dados utilizado para a estimação da curva g. Denote-se por $\hat{g}^{(-i)}(k;\lambda)$ a curva estimada usando o parâmetro de suavização λ e sem a observação y_i. Então, $\hat{g}^{(-i)}(k;\lambda)$ é a curva que minimiza

$$\sum_{j\neq i}\{Y_j - g(k_j)\}^2 + \lambda\int_a^b\{g''\}^2dx. \tag{2-35}$$

O ajuste da curva estimada $\hat{g}^{(-i)}$ pode ser avaliado se verificando quão bem $\hat{g}^{(-i)}(k_i;\lambda)$ prevê y_i. Seja $y_i - \hat{g}^{(-i)}(k_i;\lambda)$ o resíduo referente à observação y_i prevista pela curva $\hat{g}^{(-i)}$ estimada com $n-1$ observações e com parâmetro de suavização λ, que será denotado por resíduo deletado. Uma medida de ajuste orientada a previsão é o erro preditivo quadrático médio dado por

$$EPQ(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left\{y_i^* - \hat{g}(k_i;\lambda)\right\}^2 \tag{2-36}$$

em que y_i^* é a nova observação referente ao ponto k_i e $\hat{g}$ é a curva estimada com n observações e parâmetro λ. A validação cruzada é uma estimativa do erro preditivo quadrado médio [23, 12].

Dado que a escolha de qual observação y_i é retirada do ajuste de $\hat{g}^{(-i)}$, uma avaliação total da adequação do parâmetro de suavização λ pode ser obtida por meio da função escore da validação cruzada

$$VC(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{g}^{(-i)}(k_i;\lambda)\right\}^2. \tag{2-37}$$

O objetivo da validação cruzada é encontrar o valor de λ que minimiza $VC(\lambda)$. Não há garantias de que a função na equação 2-37 tenha um único mínimo. Uma busca numa grade de valores de λ pode ser o melhor método para a minimização [18].

Para calcular $VC(\lambda)$ não é necessário resolver n problemas de suavização separados para achar n curvas $\hat{g}^{(-i)}$. Usando o fato de que a curva g depende linearmente dos dados y, como mostra a equação 2-28, pode ser desenvolvida uma forma computacionalmente econômica para calcular o escore $VC(\lambda)$.

Teorema 2.6 *A função escore da validação cruzada satisfaz a seguinte equação*

$$VC(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{g}(k_i)}{1 - h_{ii}(\lambda)}\right\}^2 \tag{2-38}$$

em que $\hat{g}$ é a spline *calculada a partir de todo o conjunto de dados $\{(k_i, y_i)\}$, onde $i = 1, \ldots, n$, com parâmetro de suavização λ.*

O teorema 2.6, cuja demostração pode ser consultada em Green e Silverman (1985) [18], mostra que uma vez conhecidos os elementos $h_{ii}(\lambda)$ da diagonal principal da matriz de suavização, o escore da validação cruzada pode ser calculado a partir dos resíduos em torno da *spline* estimada com todas as n observações. Usando uma abordagem semelhante àquela dos resíduos deletados no contexto de regressão linear [6], obtém-se o resíduo deletado

$$y_i - \hat{g}^{(-i)}(k_i) = \frac{y_i - \hat{g}(k_i)}{1 - h_{ii}(\lambda)}. \tag{2-39}$$

Uma extensão da validação cruzada é validação cruzada generalizada [7, 36, 18, 22, 23]. A idéia básica da validação cruzada generalizada é substituir o fator $1 - h_{ii}(\lambda)$ na equação 2-39 por $1 - (1/n)\,\mathrm{tr}\{H(\lambda)\}$. Por analogia a 2-38, a função escore da validação cruzada é então obtida da forma

$$VCG(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - \hat{g}(k_i)\}^2}{\{1 - (1/n)\,\mathrm{tr}\{H(\lambda)\}\}^2} \tag{2-40}$$

em que $\hat{g}$ é a *spline* calculada a partir de todo o conjunto de dados $\{(k_i, y_i)\}$, onde $i = 1, \ldots, n$, com parâmetro de suavização λ. A função $VCG(\lambda)$ deve ser minimizada sobre os valores de λ.

A validação cruzada e a validação cruzada generalizada podem ser facilmente estendidas para seleção do parâmetro de suavização λ em *splines* cúbicas ponderadas [7, 36, 18].

2.2.3
Graus de liberdade do suavizador

A quantidade de suavização de um estimador pode ser expressa em termos do número de parâmetros estimados ou graus de liberdade do suavizador. Esta quantidade tem sua motivação na regressão clássica e é referida como graus de liberdade equivalentes. Suponha-se que a curva g esteja sendo estimada por uma regressão paramétrica. Assumindo que os parâmetros sejam identificáveis com base nas obsevações, a matriz H é uma projeção sobre um espaço de dimensão k. Então, o número de parâmetros ajustados é k, assim como o traço de H é igual a k. Logo, o número de graus de liberdade do modelo é igual ao traço de H. O número de graus de liberdade dos resíduos é $n - k$ que é dado por $\mathrm{tr}(I - H)$ [18, 23, 3].

Por analogia, os graus de liberdade equivalentes dos resíduos na regressão por *spline* são definidos por

$$GLER = \mathrm{tr}\{I - H(\lambda)\} \tag{2-41}$$

em que $H(\lambda)$ é a matriz de suavização associada com o parâmetro de suavização λ. Os números de graus de liberdade equivalentes dos resíduos variam de 0 quando $\lambda = 0$, a curva g interpola todos os pontos e a matriz $H(\lambda)$ é a identidade, até $n-2$ quando $\lambda = \infty$ e a curva g é a reta de regressão linear. O número de graus de liberdade equivalentes está associado com a

relação viés × variância do estimador da curva g. Da definição de validação cruzada generalizada, esta pode ser escrita em função do número de graus de liberdade equivalentes dos resíduos da forma

$$VCG\left(\lambda\right) = n \cdot \frac{SQR}{\left(GLER\right)^2} \tag{2-42}$$

em que SQR é a soma de quadrados dos resíduos.

A definição de graus de liberdade equivalentes é discutida mais profundamente em Buja, Hastie e Tibshirani (1989) [3] e Hastie e Tibshirani (1990) [23].

2.2.4 Algoritmo de estimação com múltiplas covariáveis

Modelos nos quais se tenta estabelecer a dependência de uma variável resposta Y com apenas uma covariável X não caracterizam uma ferramenta apropriada para a análise estatística de problemas complexos. Por analogia aos modelos de regressão linear clássica, na regressão não-paramétrica um modelo no qual Y depende de uma função de apenas uma covariável X pode ser escrito da forma

$$Y = g\left(X\right) + \varepsilon \tag{2-43}$$

em que ε é um vetor de erros independentemente distribuídos. A estimação do modelo 2-43 foi discutida ao longo desta seção. Entretanto, este modelo não tem muita utilidade na prática.

Admita-se, agora, que X é um vetor aleatório de dimensão p da forma $X = (X_1, \ldots, X_p)$. Suponha-se um modelo no qual a dependência da variável Y é expressa como uma combinação de funções dos componentes do vetor X. Então, um modelo com múltiplas covariáveis pode ser formulado de acordo com a seguinte equação

$$Y = g_1\left(X_1\right) + \cdots + g_p\left(X_p\right) + \varepsilon \tag{2-44}$$

onde g_j, com $j = 1, \ldots, p$, são curvas suaves das covariáveis X_j, respectivamente, e ε é um vetor de erros independentemente distribuídos. Na notação dos modelos lineares generalizados [28, 12, 23], o modelo 2-44 pode ser reescrito da seguinte forma

$$E\left(Y|X\right) \;\; = \;\; f\left(\eta\right)$$

$$\eta = \sum_{j=1}^{p} g_j(X_j) \tag{2-45}$$

tal que $f(\cdot)$ é a inversa da função de ligação apropriada para a família de distribuição de Y, η é o preditor aditivo da função de regressão e, por simplificação e sem perda de generalidade, o intercepto é igual a 0. O problema agora consiste em estimar as funções g_j dados os valores observados de X_j. Note-se que para algum j, g_j pode ser linear, isto é, da forma $g_j = \beta_j X_j$ e, neste caso, o modelo é dito semi-paramétrico.

Como cada covariável no modelo aditivo é representada separadamente, a característica de interpretabilidade é herdada do modelo linear, isto é, a variabilidade de superfície estimada depende apenas da covariável X_s quando todas as outras covariáveis $X_{j\neq s}$ são fixadas. Devido a esta simplificação, os modelos aditivos são aproximações da verdadeira superfície de regressão por uma soma de funções individuais dos preditores. Entretanto, os modelos aditivos não lidam de forma trivial com interações entre os preditores [23, 14].

Os suavizadores multidimensionais de alta dimensão não funcionam adequadamente pois herdam a esparsividade das amostras de dimensão alta, a chamada "maldição da dimensionalidade" [14]. Uma discussão detalhada da abordagem dos suavizadores multidimensionais de baixa ordem, por exemplo, *thin plate splines* pode ser consultada em Wood (2003) [38], Wood(2000) [37], Wahba (2000) [36] e Green e Silverman (1994) [18]. Os modelos aditivos caracterizam uma abordagem para lidar com problemas de alta dimensão, decompondo-os em problemas de baixa dimensão, normalmente $d = 1$ [14, 22].

Considere-se estimar as funções $g_1, \ldots, g_p$ do modelo aditivo 2-45. O modelo pode ser estimado por meio do algoritmo *backfitting*, também conhecido como *projection pursuit* [22, 14, 12]. O algoritmo consiste em estimar uma função g_s dadas as estimativas das funções $g_{j\neq s}$, com $j = 1, \ldots, p$, num procedimento iterativo até que um critério de convergência seja satisfeito. Um exemplo de critério de convergência pode ser dado pela diferença entre a soma de quadrados dos resíduos entre duas iterações consecutivas comparado com um valor fixo tão pequeno quanto se deseje.

No algoritmo 2.1, m é o contador de iterações, R_s é o resíduo parcial do modelo aditivo com todas as curvas $g_{j\neq s}$, $h_s(\cdot)$ é uma função suavizadora arbitrária aplicada à covariável X_s e SQR é a soma de quadrados dos resíduos.

A regressão por *projection pursuit* é uma forma direta de atacar o

Algoritmo 2.1 O algoritmo *backfitting*

1. Inicia-se $g_1^{(0)} = \cdots = g_p^{(0)} = 0$ e $m = 0$

2. Itera-se: $m = m + 1$
   Para $j = 1$ até p faz-se:

$$R_s = Y - \sum_{j=1}^{s-1} g_j^{(m)}(X_j) - \sum_{j=s+1}^{p} g_j^{(m-1)}(X_j)$$

3. Estima-se

$$g_s^{(m)} = h_s(R_s|X_s)$$

4. Até que

$$SQR = \left\| Y - \sum_{j=1}^{p} g_j^{(m)}(X_j) \right\| \leq \epsilon$$

problema da dimensionalidade. Considere-se o modelo

$$Y = \sum_{k=1}^{K} h_k(\alpha_k' X) + \varepsilon \tag{2-46}$$

no qual $\alpha_k' X$ denota uma projeção unidimensional do vetor X, h_k é uma função univariada arbitrária da projeção e os erros são independentes de X com média zero e variância σ^2. O algoritmo constrói a superfície de regressão escolhendo as projeções definidas pelo vetor α_k. As direções α_k e o número de termos K, em 2-46, são escolhidos de forma a oferecer o melhor ajuste aos dados. O algoritmo *backfitting* é um algoritmo Gauss-Seidel para solução de sistemas de equações [5]. Se os suavizadores $h(\cdot)$ são operadores de projeção, a convergência do algorimo é garantida. Alguns suavizadores como as *splines* embora não sejam operadores de projeção, possuem as propriedades requeridas para a convergência [23]. O modelo 2-46 procura explicar a variabilidade da variável resposta não por uma seqüência suavizada, mas por uma soma de suavizações de várias seqüências da variável resposta induzida por várias combinações lineares do preditor [14].

Nos modelos de regressão linear múltipla, a interpretação dos coeficientes pode ser seriamente comprometida se existe colinearidade entre as covariáveis. Um fenômeno análogo pode ocorrer nos modelos não-paramétricos

chamado *concurvidade*[2]. Seus efeitos na interpretação das curvas individuais nos modelos aditivos ainda não são bem conhecidos [22].

[2]do termo em inglês *concurvity*.