



UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE MATEMÁTICA  
DEPARTAMENTO DE ESTATÍSTICA



# **Um estudo comparativo de estimadores de regressões não-paramétricas aditivas: Performance em Amostras Finitas**

Autor: Renan Xavier Cortes  
Orientador: Fernando Augusto Boeira Sabino da Silva

Porto Alegre, 4 de Junho de 2010.

Universidade Federal do Rio Grande do Sul  
Instituto de Matemática  
Departamento de Estatística

# Um estudo comparativo de estimadores de regressões não-paramétricas aditivas: Performance em Amostras Finitas

Autor: Renan Xavier Cortes

Monografia apresentada para obtenção  
do grau de Bacharel em Estatística.

Banca Examinadora:  
Professor Fernando Augusto Boeira Sabino da Silva  
Professor Dr. Flávio Augusto Ziegelmann

Porto Alegre, 4 de Junho de 2010.

*Dedico este trabalho ao Popó por ser atencioso e estar sempre presente nos momentos mais frios e difíceis da minha vida.*

*“A genialidade é feita de 10% de talento e 90% de esforço.”*  
*Thomas Alva Edison*

# Agradecimentos

Agradeço primeiramente a minha família que me deu todo o suporte necessário para conclusão de mais esta etapa na minha vida.

Agradeço ao meu pai por ter sido sempre atencioso, dedicado e compreensivo e por ter me dado a excelente herança genética inclinada para a área das ciências exatas.

Agradeço a minha mãe por ser um exemplo materno e por ser sempre atenciosa, dedicada, carinhosa e carismática e que, quando nos momentos mais difíceis, me levava para esfriar a cabeça e fazer compras no supermercado.

Agradeço ao meu irmão por ser uma pessoa extremamente inteligente, engraçada, companheira e carismática que sempre, quando podia, perdia para mim nas partidas de Winning Eleven para levantar o meu astral.

Agradeço a todos os professores do departamento de Estatística da UFRGS, não apenas pelos valiosos ensinamentos, mas também pelos conselhos de experiências pessoais transmitidos.

Em especial, agradeço aos professores Fernando Augusto Boeira Sabino da Silva, Suzi Alves Camey e Jandyra Maria Guimarães Fachel por ter me auxiliado bastante no meu trabalho de conclusão de curso, ter me fornecido uma excelente experiência na bolsa de iniciação científica, e na bolsa de extensão, respectivamente.

Por fim, agradeço a todas as pessoas com quem eu conheci ou me relacionei de maneira direta ou indireta ao longo da minha jornada no curso de Bacharelado em Estatística. Todas elas foram importantes e ocupam um espaço no meu coração.

## Resumo

Nesta monografia, conduzimos um exercício de simulação de Monte Carlo a fim de investigar algumas características em amostras finitas de três estimadores baseados em *kernels* disponíveis: o estimador Backfitting Clássico (CBF), o estimador Smooth Backfitting (SBF) e o estimador de 2 estágios (2E) proposto por Kim et. al. (1999). Diferentes escolhas do parâmetro de suavização foram estudadas, diferentes métodos de suavização, bem como diferentes níveis de correlação. A comparação é baseada nas estimativas da média dos erros quadráticos médios de cada estimador. Os resultados do nosso estudo sugerem que o estimador Backfitting Clássico é superior aos outros dentro das especificações dos cenários propostos.

Palavras-Chave: modelos aditivos, regressão não-paramétrica aditiva, backfitting, smooth backfitting, integração marginal

# Sumário

1. Introdução.....	8
2. Revisão de Regressões Paramétricas.....	10
2.1. Regressão Linear Simples.....	11
2.2. Regressão Linear Múltipla.....	12
2.3. Regressão Não-Linear.....	14
2.4. Regressões Paramétricas: Considerações Finais.....	15
3. Regressão Não-Paramétrica.....	18
3.1. Estimação de Densidades.....	19
3.1.1. Histograma.....	19
3.1.2. Método de <i>kernel</i> .....	21
3.1.3. Parâmetro de Suavização.....	22
3.2. Estimação de Curvas de Regressão.....	24
3.2.1. Regressão Polinomial Local.....	26
3.2.2. Nadaraya-Watson e Regressão Linear Local.....	27
3.3. Problemas da Regressão Não-Paramétrica.....	31
3.3.1. <i>Curse of Dimensionality</i> .....	32
4. Modelos Aditivos.....	35
4.1. Definição.....	35
4.2. Estimação.....	37
4.3. Backfitting Clássico.....	38
4.4. Smooth Backfitting.....	40
4.5. Estimador de 2 estágios.....	41
5. Escolha do parâmetro de suavização.....	43
5.1. Escolha de $h_n$ para Backfitting Clássico e Smooth Backfitting.....	44
5.2. Escolha de $h_n$ para o Estimador de 2 estágios.....	45
5.3. Janelas Estimadas.....	46
5.4. Janelas Verdadeiras.....	47
6. Processo gerador dos dados.....	48
7. Resultados e Conclusões.....	50
Referências Bibliográficas.....	58
Anexo I.....	61
Anexo II.....	63

## 1. Introdução

A estimação de regressões não-paramétricas tem sido o objetivo de vários artigos recentes. Regressão não-paramétrica tornou-se um campo de rápido desenvolvimento desde que se notou que regressão paramétrica poderia ter uma grande demanda de tempo para a obtenção de modelos adequados para muitos conjuntos de dados. Sem dúvida, um dos problemas mais estudados em Estatística e Econometria está relacionado à estimação de funções paramétricas de regressão. A necessidade, bem como a adequação da modelagem de funções de regressões paramétricas, são amplamente discutidas em um grande número de livros, ver, por exemplo, Draper & Smith (1998), Johnson & Dinardo (1996) e Greene (1999). Porém, as boas propriedades dos estimadores de máxima verossimilhança e dos estimadores de mínimos quadrados dependem da forma funcional paramétrica assumida estar corretamente especificada. Caso contrário, estes estimadores não serão eficientes, assintoticamente normais e nem mesmo consistentes. Uma alternativa menos restrita para estimação de uma função de regressão é o modelo de regressão não-paramétrica  $Y = E(Y | X = x) + \varepsilon = m(x) + \varepsilon$ , onde  $m(x)$  pertence a uma classe de funções bem mais abrangente do que a de modelos paramétricos.

No entanto, diversos problemas são levantados quando se tratando de modelos não-paramétricos com múltiplos regressores. Um deles é de que, infelizmente, a convergência entre os estimadores é mais lenta à medida que aumentamos o número de regressores. No entanto, uma maneira de evitar tais problemas é estabelecendo uma restrição aditiva na estrutura do modelo não-paramétrico através de um modelo aditivo.

Modelos aditivos são modelos semi-paramétricos da forma:

$$Y = \alpha + \sum_{d=1}^p m_d(x_d) + \varepsilon ,$$

onde cada função individual  $m_d(x_d)$  será estimada não parametricamente.

O objetivo da monografia é comparar diferentes maneiras disponíveis atualmente para estimar modelos aditivos: o estimador *Backfitting* Clássico (CBF), o estimador *Smooth Backfitting* (SBF) e o estimador de 2 estágios (2E). A estrutura dá-se da seguinte forma: o capítulo 2 dá uma breve descrição e revisão de modelos paramétricos lineares e não-lineares, contrastando suas vantagens e desvantagens na análise de dados. O capítulo 3 versa sobre a estimação não-paramétrica, abordando estimação de densidades de probabilidade de variáveis contínuas, estimação de curvas de regressão, o método de kernel, a importância do parâmetro de suavização na estimação não-

paramétrica, o método de estimação polinomial local, o método de Nadaraya-Watson, assim como os problemas agregados. O capítulo 4 introduz a classe de modelos estudados no presente trabalho: os modelos aditivos. Definimos sua estrutura, estimação bem como os estimadores sob estudo. O capítulo 5 apresenta os métodos de escolha dos parâmetros de suavização utilizados para cada estimador. O capítulo 6 descreve o processo gerador dos dados simulados no estudo de Monte Carlo realizado, bem como as especificações dos cenários construídos. Por fim, o capítulo 7 traz os resultados obtidos após a realização das simulações, tabelas de resumo do desempenho de cada estimador, tabelas de demanda computacional, gráficos comparativos e conclusões a respeito dos achados do trabalho.

Nas últimas páginas deste trabalho, encontram-se as referências bibliográficas utilizadas, gráficos e todos os códigos em R utilizados no trabalho.

## 2. Revisão de Regressões Paramétricas

Modelos estatísticos constituem ferramentas extremamente úteis para resumir e interpretar dados. Em particular, eles podem facilitar a avaliação da forma e da intensidade de associações de interesse em estudos das mais diversas áreas como, por exemplo, em economia, agronomia, epidemiologia, administração, etc. (Conceição, Saldiva e Singer, 2001).

De um modo geral, os principais objetivos da modelagem estatística é descrever o relacionamento entre uma ou várias variáveis exploratórias  $X$ 's (independentes) e uma (ou, em alguns casos, várias) variável resposta de interesse  $Y$  (dependente). Comumente, em regressão, estamos interessados em modelar a *expectativa condicional de  $Y$*  tendo os valores das variáveis preditoras. Seja para avaliar o impacto dos regressores sobre a variável de interesse, seja para prever o valor de  $Y$  com base nos valores dos  $X$ 's, ou ainda para qualquer outra finalidade da modelagem, a abordagem paramétrica exige *a priori* o estabelecimento da relação funcional entre a variável resposta e as variáveis preditoras. Em outras palavras, no modelo abaixo

$$Y_i = f(X_1, X_2, \dots, X_p) + \varepsilon_i \quad i = 1, 2, \dots, n, \quad (2.1)$$

necessitamos estabelecer previamente qual é a estrutura funcional de  $f(X_1, X_2, \dots, X_p)$  que estamos querendo estimar para conduzir as estimativas dos parâmetros e realizar testes, previsões, etc.

Supondo que nós temos disponíveis  $n$  observações da variável  $Y$  em  $n$  vetores  $x^i = (x_{i1}, x_{i2}, \dots, x_{ip})$ , nosso objetivo é modelar a dependência do primeiro momento de  $Y$  dado  $X_1, X_2, \dots, X_p$ :

$$E(Y_i | X_1=x_1, X_2=x_2, \dots, X_p=x_p) = f(X_1, X_2, \dots, X_p). \quad (2.2)$$

Existem várias razões para os porquês de estarmos interessados nisto. A primeira delas é *descrição*: queremos um modelo que descreva a dependência da resposta nos preditores para que possamos aprender mais sobre o processo gerador de  $Y$ . A segunda é *inferência*: gostaríamos de inferir valores de parâmetros populacionais a partir de valores obtidos da amostra. Por último, a *predição*: há interesse na predição de  $Y$  para um conjunto de valores  $X_1, \dots, X_p$ .

Um dos casos mais simples e muito utilizado na prática para avaliar esta relação funcional é descrito a seguir.

## 2.1 Regressão Linear Simples

De fato, a Análise de Regressão Linear Paramétrica é uma das técnicas estatísticas mais populares e úteis na análise de dados (Gujarati, 2000). A análise de regressão linear simples tenta estabelecer a relação funcional entre a variável dependente e a variável independente através de uma reta, como na figura 2.1:

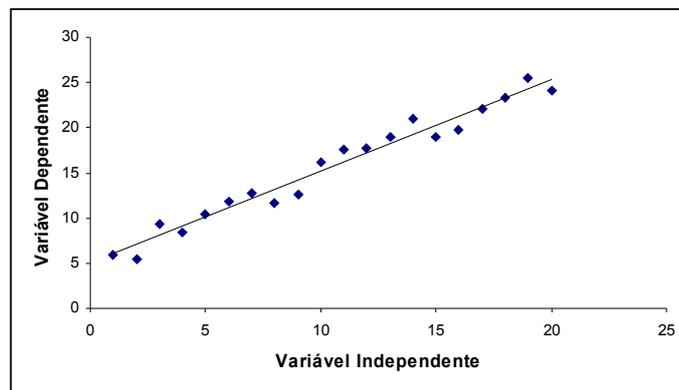


Figura 2.1: diagrama de dispersão de 20 pontos simulados, com ajuste de regressão linear

No caso simples em que a expectativa condicional de  $Y$  é baseada em relação funcional linear nos parâmetros e na variável explicativa  $X$ , o modelo de regressão é o que segue:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (2.1.1)$$

onde  $\beta_0$  e  $\beta_1$  são constantes e o termo  $\varepsilon_i$  é o erro não-determinístico associado ao modelo proposto. O parâmetro  $\beta_0$  é denominado intercepto e o grau de inclinação da reta é dado por  $\beta_1$ . Comumente, nos modelos paramétricos, necessitamos estabelecer uma distribuição de probabilidade para os erros associados ao modelo que, conseqüentemente, será a distribuição de probabilidade condicional da variável dependente  $Y$ , pois a idéia por trás da estimação dos parâmetros pode ser feita sob a condição de que as variáveis independentes são *constantes*, isto é, *já são dadas*. No caso mais simples, o modelo descrito acima assume normalidade dos erros, esperança nula, variância constante e valores independentes e identicamente distribuídos. Isto é:

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

Os parâmetros  $\beta_0$  e  $\beta_1$  do modelo (2.1.1) são chamados coeficientes da regressão e podem ser estimados utilizando uma amostra de dados. Uma possibilidade é via o método de mínimos quadrados. Ele consiste em minimizar a soma dos quadrados das diferenças entre os valores observados e a reta ajustada, levando a estimativas dos parâmetros que são combinações lineares dos valores observados da variável resposta.

Note que se o pesquisador utilizar este modelo de regressão, bastante utilizado na prática, pode estar cometendo um equívoco, pois estaria estabelecendo a relação linear entre as variáveis mesmo que esta relação seja inexistente. Um caso mais geral, em que adicionamos múltiplas variáveis na modelagem da expectativa condicional, é descrito a seguir.

## 2.2 Regressão Linear Múltipla

Quando temos  $p$  variáveis para descrever a relação funcional entre a expectativa condicional da variável  $Y$  de interesse e as demais variáveis, podemos optar pela regressão linear múltipla. Este modelo é semelhante ao de regressão simples, porém obtido com mais variáveis regressoras e é dado pela equação descrita abaixo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \varepsilon_i, \quad (2.2.1)$$

onde  $\beta_k$ ,  $k = 0, 1, \dots, p$ , são constantes desconhecidas. Novamente, é comum assumirmos que os erros associados têm, por suposição, média zero, variância desconhecida constante  $\sigma^2$ , e também podemos utilizar as suposições de independência, distribuição idêntica e, para fins de testes, seguir uma distribuição Normal.

Este modelo faz uma forte suposição entre a relação de  $Y$  com os  $X$ 's, sugerindo uma dependência linear em cada preditor. Se essa suposição se sustenta, mesmo que aproximadamente, então este modelo é extremamente útil e conveniente, pois:

- Permite uma simples descrição dos dados;
- Resume a contribuição de cada preditor com um único coeficiente;
- Promove um método simples para predição de novas observações.

Sendo  $p$  o número de variáveis regressoras, o modelo de regressão linear múltipla define uma superfície em um espaço  $(p+1)$ -dimensional que, dependendo do número de variáveis independentes inseridas, inviabiliza a visualização gráfica desta superfície.

Um exemplo de superfície, gerada com duas variáveis preditoras, em um espaço tridimensional segue abaixo:

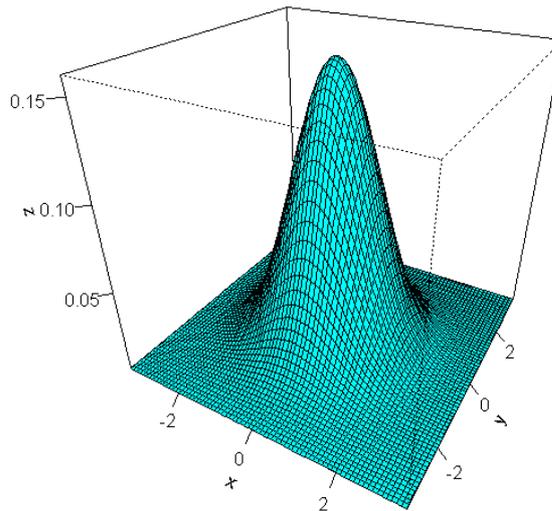


Figura 2.2: exemplo de superfície tridimensional

No modelo (2.2.1) os parâmetros estimados representam a alteração esperada na variável resposta  $Y$  por unidade de  $X_j$ , quando todas as demais variáveis regressoras  $X_i$  ( $i \neq j$ ) são mantidas constantes. Por essa razão são chamados de coeficientes parciais da regressão.

O modelo descrito em (2.2.1) com  $p$  variáveis regressoras é chamado de modelo de primeira-ordem com  $p$  variáveis regressoras. Isto é, quando o efeito de  $X_i$  na média de  $Y$  não depende do nível de  $X_j$  ( $i \neq j$ ), e vice-versa, dizemos que as variáveis possuem *efeitos aditivos* ou *não interagidas* (Neter et. al., 1996).

Neste caso, as estimativas dos parâmetros são comumente dadas também pelo método dos mínimos quadrados e, em notação matricial, temos:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (2.2.2)$$

onde  $A^T$  e  $A^{-1}$  são as matrizes transposta e inversa de  $A$ , respectivamente, e:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{e} \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdot & \cdot & X_{1p} \\ 1 & X_{21} & X_{22} & \cdot & \cdot & X_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{n1} & X_{n2} & \cdot & \cdot & X_{np} \end{pmatrix}$$

No caso em que  $p < n$ , o Teorema de Gauss-Markov (que também vale para o caso da regressão simples) postula que os estimadores  $\hat{\beta}_k$  obtidos por mínimos quadrados

são os melhores estimadores lineares não-viciados de  $\beta_k$ . Se os erros forem considerados normais, então, os estimadores  $\hat{\beta}_k$  são, também, estimadores de máxima verossimilhança e são não-viciados de variância mínima entre todos não-viciados.

### 2.3 Regressão Não-Linear

Em um caso mais geral, podemos supor um modelo da forma

$$Y_i = f(X_i, \beta) + \varepsilon_i. \quad (2.3.1)$$

Se o modelo escolhido for da forma

$$f(X_i, \beta) = X_i^T \beta, \quad (2.3.2)$$

teríamos os modelos descritos anteriormente pois a estrutura é *linear nos parâmetros*.

Modelos de regressão não-linear são, basicamente, da mesma forma do que (2.3.1) para regressão linear, porém o vetor de parâmetros pode ter uma relação não-linear nos preditores. Eles podem ser escritos na forma:

$$Y_i = f(X_i, \gamma) + \varepsilon_i \quad (2.3.3)$$

Uma observação  $Y_i$  é a soma de uma resposta média  $f(X_i, \gamma)$  dada por uma função de resposta não-linear  $f(X, \gamma)$  e um termo de erro  $\varepsilon_i$ . As suposições usuais do termo do erro são mantidas.

O vetor de parâmetros da função de resposta  $f(X, \gamma)$  é agora denotado por  $\gamma$  ao invés de  $\beta$  como diferencial de a resposta não ser mais linear nos parâmetros. Um exemplo de estrutura não linear segue:

$$Y_i = \gamma_0 \exp(\gamma_1 X_{1i}) + \frac{\gamma_2}{\sqrt{X_{2i} \gamma_3}} + \varepsilon_i, \quad (2.3.4)$$

onde

$\gamma_0, \gamma_1, \gamma_2$  e  $\gamma_3$  são parâmetros a serem estimados,

$X_{1i}$  e  $X_{2i}$  são usualmente regressoras constantes conhecidas,

$\varepsilon_i$  são erros que usualmente podemos assumir independência e distribuição  $N(0, \sigma^2)$ .

A função resposta para este modelo é

$$f(X, \gamma) = \gamma_0 \exp(\gamma_1 X_1) + \frac{\gamma_2}{\sqrt{X_2 \gamma_3}}. \quad (2.3.5)$$

A estimação dos parâmetros do modelo de regressão não-linear é usualmente feita através do método de mínimos quadrados ou através de máxima verossimilhança. No entanto, ao contrário dos modelos lineares, não é possível encontrar expressões com as soluções analíticas para mínimos quadrados, como em (2.2.2), e estimadores de máxima verossimilhança para modelos não-lineares. A estimação é feita através de procedimentos iterativos numéricos que podem exigir intensa demanda computacional.

Para mais detalhes sobre modelos de regressões não-lineares ver Neter et. al. (1996, Cáp. 13).

## 2.4 Regressões Paramétricas: Considerações Finais

Os modelos paramétricos lineares e não-lineares vistos anteriormente englobam uma grande parte da modelagem estatística aplicada nas mais diversas áreas da ciência. Tanto para modelos de regressão, análise de variância e planejamentos de experimentos, estes modelos são amplamente utilizados em administração, economia, engenharia e ciências sociais, da saúde e biológicas. Aplicações bem sucedidas destes modelos requerem um conhecimento tanto teórico, quanto prático por trás do problema real em questão (Neter, 1996).

No entanto, os modelos paramétricos possuem uma estrutura rígida e exigem um estabelecimento *a priori* da relação funcional entre a variável resposta e suas preditoras. Mesmo no caso linear múltiplo onde poderíamos incluir termos não lineares nas variáveis preditoras, incluindo termos polinomiais (ou até mesmo termos de interação), a fim de detectar uma relação não-linear entre Y e os demais X's, como por exemplo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2 + \varepsilon, \quad \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

ainda estaríamos determinando um número finito de parâmetros e adicionando as suposições usuais para os erros associados a ele. O ajuste é facilmente computado, através de mínimos quadrados, interpretável e estimado. Mas, se as suposições do modelo são violadas as estimativas podem tornar-se inconsistentes e viesadas.

### 3. Regressão Não-Paramétrica

Dado um vetor aleatório  $(\mathbf{Y}, \mathbf{X})$ ,  $\mathbf{Y} \in \mathfrak{R}$  e  $\mathbf{X} \in \mathfrak{R}^d$ , a expectativa condicional  $E(Y | X=x) = m(x)$ , onde  $x^T = (x_1, \dots, x_d)$  pode ser estimada não-parametricamente sob algumas condições de regularidade. Em outras palavras, no caso univariado, desejamos ajustar uma função que represente a relação entre  $y$  e  $x$ , isto é,  $m(x)$  tal que

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, n. \quad (3.1)$$

No caso mais simples, podemos assumir que  $E(\varepsilon_i)=0$ ,  $\text{Var}(\varepsilon_i)=\sigma^2$ ,  $i=1,\dots,n$  e que os erros são independentes e que a função  $m(\cdot)$  é contínua e diferenciável um número  $q$  de vezes.

Note que na abordagem paramétrica discutida anteriormente a maneira mais comum de estimar uma função de regressão é optar por um modelo paramétrico linear ou não-linear, permitindo previsões para valores de  $Y$  e a construção de intervalos de confiança e testes para os parâmetros.

No entanto, na abordagem paramétrica é possível especificar-se uma família de formas funcionais em (3.1) para  $m$  de maneira errada. Este problema, possivelmente desastroso para a abordagem paramétrica, inexistente no enfoque não-paramétrico. Além disso, a adoção de abordagens flexíveis para a estimação de  $m$  pode levar à descoberta de características consideradas insuspeitas quando da adoção de um modelo paramétrico. Por esta razão, é de interesse explorar o que se pode aprender sobre a função  $m$  sem restringi-la a modelos estabelecidos *a priori*.

O objetivo da regressão não-paramétrica é aproximar  $m$ , que pode pertencer a uma classe bastante vasta, sendo a única restrição imposta de que seja uma função continuamente diferenciável até determinada ordem. Cabe comentar, no entanto, que existe um preço a pagar pela flexibilidade da modelagem não-paramétrica. O tamanho da amostra para conseguir-se a mesma eficiência será maior no caso não-paramétrico do que no paramétrico, quando o modelo paramétrico especificado for correto, devido às taxas de convergência dos estimadores não-paramétricos serem mais lentas do que a dos estimadores paramétricos. Este problema de convergência será discutido posteriormente.

Dentro do contexto não-paramétrico, os dois problemas mais abordados na literatura são a estimação de densidades e a estimação de funções de regressão. Esta última sendo a de interesse para a estimação de modelos aditivos que serão vistos

posteriormente. A idéia principal é “permitir que os dados falem por si só”, ou seja, os dados que vão ditar o formato que a densidade ou curva de regressão vão se comportar no processo de estimação. Este formato se dará através de um *suavizador*, que representa uma ferramenta para resumir a tendência da medida da resposta que seja *menos variável do que o próprio Y coletado* da amostra (Hastie & Tibhirani, 1990).

Existe uma vasta literatura para diferentes abordagens de suavizadores, como suavizadores usando núcleo-estimadores (*kernels*), splines (Green e Silverman, 1994), loess (*lowess*), médias-móveis, ondaletas (*wavelets*) (Efromovich, 1999), etc.

Será explicado, a partir de agora, a metodologia de estimação de densidades de probabilidade com enfoque na suavização usando *kernels* que faz parte do método de estimação de regressões de curvas desta monografia.

### 3.1 Estimação de Densidades

Este capítulo é dedicado para a compreensão de alguns métodos de estimação de densidades de probabilidades que são extremamente úteis na prática e servem como base para a compreensão da estimação de curvas de regressão vistas logo a seguir.

Considerando uma variável aleatória contínua e sua função densidade de probabilidade (fdp), o seu formato descreve a distribuição desta variável aleatória e podemos calcular não somente a média e a variância, mas também a probabilidade de ocorrência de um determinado valor em um intervalo.

O primeiro método abordado é o método mais popular dentre os estimadores de fdp.

#### 3.1.1 Histograma

O histograma é o método de estimação de densidade mais utilizado na área da estatística descritiva e compreende um método direto e rápido de acessar um formato “grosseiro” da distribuição de probabilidade univariada contínua.

O histograma é obtido a partir da divisão do intervalo a que os dados pertencem em classes que, em geral, possuem o mesmo tamanho  $h$ . Cada observação amostral é alocada em sua devida classe e é calculada a proporção da amostra contida em cada uma delas dividida pela largura da mesma, resultando nas estimativas da fdp, que são nada mais do que as alturas destas barras do histograma.

A estimativa em um ponto  $x$  é dada por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n I(x - \tilde{x}_i, h/2), \quad (3.1.1.1)$$

Sendo:

$n$  - tamanho da amostra

$h$  - largura da classe

$\tilde{x}_i$  - é o ponto central da classe da observação  $x_i$

$I(x - \tilde{x}_i, h/2)$  - é uma função indicadora do intervalo  $[-h/2, h/2]$ .

Tendo estabelecido formalmente a construção deste estimador de densidade, cabe ressaltar alguns comentários. Primeiramente, poderíamos variar o valor de “âncora” da estimativa variando o valor onde se inicia cada classe. Em segundo, o tamanho da largura da classe não precisa ser necessariamente igual para todo o domínio amostral podendo variar para tamanhos maiores onde temos menos observações. Em terceiro lugar, o formato do histograma depende crucialmente do número de classes que será utilizado na sua estimação. Neste último comentário, introduz-se o conceito de **parâmetro de suavização  $h$**  que representa o papel mais importante na estimação da densidade de probabilidade de variáveis contínuas e na modelagem não-paramétrica.

Os gráficos abaixo ilustram visualmente de que maneira o parâmetro  $h$  influencia diretamente no formato do gráfico do histograma quando variamos seu valor para o mesmo conjunto de dados:

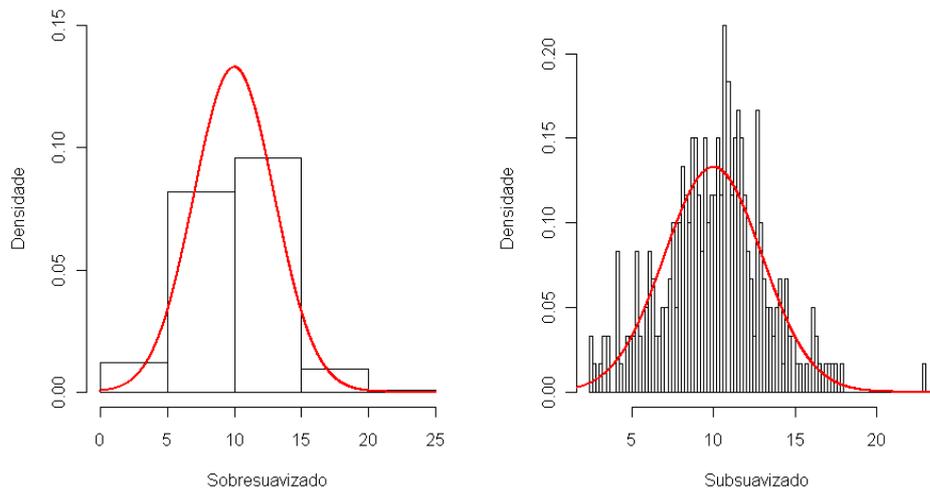


Figura 3.2: histogramas com diferentes parâmetros de suavização

Os gráficos acima mostram uma amostra simulada de 300 valores de uma distribuição normal com média 10 e desvio padrão 3. No primeiro gráfico temos um histograma com parâmetro  $h$  maior do que no segundo gráfico causando assim um comportamento sobresuavizado quando comparado com o segundo gráfico, que possui

um parâmetro  $h$  menor e, conseqüentemente, possui um maior número de classes e formato subsuavizado quando comparado com o primeiro gráfico. Para ilustração, foi plotada, em vermelho, uma linha que representa a verdadeira fdp dos dados simulados.

O termo “grosseiro” utilizado no primeiro parágrafo deste subcapítulo se deve ao fato de que o formato deste estimador possui saltos descontínuos na estimação e falta de suavidade, apesar da distribuição subjacente ser suave.

Para maiores detalhes sobre as propriedades dos estimadores, veja Silverman (1986), Scott (1992) e Simonoff(1996).

### 3.1.2 Método de *kernel*

Diferente do histograma, o método de *kernel* não possui a necessidade de definir classes. É um procedimento para estimação de uma função que consiste em estabelecer uma média localmente ponderada.

O estimador de *kernel*, também chamado de estimador baseado em núcleos, é baseado em uma função de probabilidade  $K$  que, por Silverman (1996), deve ser contínua, unimodal, simétrica em torno de zero, com variabilidade controlada pelo parâmetro  $h$  e limitada. Cline (1988) mostra que os estimadores que não possuem tais propriedades são inadmissíveis no sentido em que eles não têm o menor erro quadrático médio integrado.

O estimador de Rosenblatt-Parzen é um estimador de fdp que resolve o problema do histograma substituindo a função indicadora por uma função *kernel* ponderadora e é definido da seguinte forma:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (3.1.2)$$

onde a estimativa  $\hat{f}_h(x)$  também será uma fdp.

Um suavizador *kernel* usa uma explícita e definida maneira de ponderar as observações para produzir a estimativa de cada valor. A idéia é fazer a *estimação global com ponderação local*. Usualmente este suavizador se utiliza de ponderação que decai de maneira suave à medida que os valores vão se afastando do ponto em que se está realizando a estimação. Logo, a descontinuidade do histograma é grosseira e desnecessária, pois podemos, ao invés de ponderar igualmente as observações através da função indicadora, acessar ponderações menores para valores mais afastados do ponto de interesse.

O parâmetro de suavização  $h$  define uma espécie de controlador entre vício e variância da estimativa. Se  $h$  for muito baixo, pode-se estar deixando de suavizar a função de maneira eficiente (subsuaivizar), enquanto que se pode estar suavizando a função em excesso (sobresuaivizar) caso  $h$  for muito grande.

A escolha da função núcleo  $K$ , apesar de desempenhar papel fundamental na estimação não-paramétrica, é uma tarefa de menor importância que a escolha de  $h$ . Estudos demonstram que a função núcleo dita ótima, no sentido que minimiza as funções critérios comumente utilizadas, leva a pequenas melhorias em relação à maioria das funções núcleo escolhida. Consequentemente, a simplicidade, a facilidade de manipulação analítica, o custo computacional e a velocidade de convergência dos estimadores frequentemente determinam a escolha da função núcleo. A função núcleo ótima, no sentido em que minimizamos o critério do Erro Quadrático Médio Integrado ou MISE (Pagan, A. and A. Ullah, 1999), foi primeiramente sugerida por Epanechnikov (1969) e é, por isso, frequentemente, chamada de função núcleo de Epanechnikov e, para o caso univariado, é dada por:

$$K(\psi) = \begin{cases} \frac{3}{4}(1 - \psi^2), & \text{se } |\psi| \leq 1, \psi \in \mathfrak{R} \\ 0, & \text{c.c.} \end{cases}$$

A função núcleo Epanechnikov tem formato quadrático e pode ser vista em Härdle (1990, p.44) e Silverman (1986, p.42). Outra função núcleo comumente utilizada e que não apresenta descontinuidades é a função núcleo gaussiana, isto é,

$$K(\Psi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\Psi^2\right\}, \Psi \in \mathfrak{R}$$

Simonoff (1996, p.44) e Silverman (1986, p.43) mostram uma tabela onde se comparam a eficiência das funções núcleo comumente utilizadas em relação à função ótima de Epanechnikov. Pode-se notar que até a função uniforme, que pondera as observações de maneira homogênea e, por isso, é dita mais “ingênua”, tem uma eficiência de aproximadamente 0,93 da função ótima. A função núcleo gaussiana, que será utilizada nesta monografia, possui uma eficiência de aproximadamente 0,95 nesta comparação.

### 3.1.3 Parâmetro de Suavização

Como já mencionado, a escolha do parâmetro de suavização  $h$  é crucial para um bom ajuste não-paramétrico.

Uma maneira rudimentar de escolher  $h$  é por tentativa e erro baseado no uso de análises gráficas. Este método consiste em olhar vários gráficos de  $\hat{f}(x)$  contra  $x$ , quando  $\hat{f}(x)$  é calculado através de diferentes valores de  $h$ . Esta maneira pode ser efetiva, embora seja vantajoso escolher  $h$  de uma maneira mais objetiva.

Para fazer uma escolha de  $h$  é necessário ter algum critério. O objetivo é simplesmente avaliar se  $\hat{f}(x)$  é um bom estimador para  $f(x)$ . Os critérios mais populares envolvem a minimização de uma função perda ou risco quadrático, dentre as quais se destacam: Erro Quadrático Integrado (ISE, do inglês *Integrated Squared Error*), Erro Quadrático Médio Integrado (MISE, do inglês *Integrated Mean Squared Error*), Erro Quadrático Médio (ASE, do inglês *Averaged Squared Error*), Média dos Erros Quadráticos Médios (MASE, do inglês *Mean Averaged Squared Error*), Erro Quadrático Médio Assintótico (AMSE, do inglês *Asymptotic Mean Squared Error*) e a aproximação assintótica da Média dos Erros Quadráticos Médios (AMASE, do inglês *Asymptotic Mean Averaged Squared Error*). Para maiores detalhes veja Simonoff (1996), Pagan & Ullah (1999) e Opsomer e Ruppert (1998).

Como se pode notar pela discussão acima, as preferências do usuário no que tange a vício e variância serão refletidas na escolha de  $h$ , de algum modo, quando ele escolhe a função critério a ser minimizada. A variância da  $\hat{f}(x)$  diminui quando  $h$  torna-se maior e vice-versa. O vício da  $\hat{f}(x)$  aumenta quando  $h$  aumenta e vice-versa. Esta relação é conhecida como *trade-off* entre vício e variância desta escolha.

Dentre os métodos de suavização existentes, o que facilmente ilustra a compreensão desta relação é, por exemplo, o suavizador média-móvel. Ele pode ser escrito como:

$$\hat{f}_k(x_i) = \sum_{j \in V_k^s(x_i)} \frac{y_j}{2k+1}$$

onde,

$$E[\hat{f}_k(x_i)] = \sum_{j \in V_k^s(x_i)} \frac{f(x_j)}{2k+1}$$

e

$$Var[\hat{f}_k(x_i)] = \frac{\sigma^2}{2k+1}$$

Para interpretação de notação, assumimos que  $x_i$  é próximo ao “meio” dos dados para que  $V_k^s(x_i)$ , a vizinhança de  $x_i$ , contenha todos  $2k+1$  pontos (sendo  $k$  à esquerda e  $k$  à direita). Desta forma, o valor de  $k$  funciona como o parâmetro de suavização  $h$  e

constata-se que à medida que  $k$  aumenta a variância diminui, porém a esperança inflaciona o vício porque  $\sum_{j \in V_k^s(x_i)} f(x_j)/(2k+1)$  envolveria mais termos com valores de  $f(\cdot)$  diferentes de  $f(x_i)$ .

Isto é facilmente visualizado graficamente no contexto de estimação de densidades. Foi gerada uma distribuição de uma variável aleatória com densidade normal com  $\sigma=2$  e três estimações de densidades seguem a seguir:

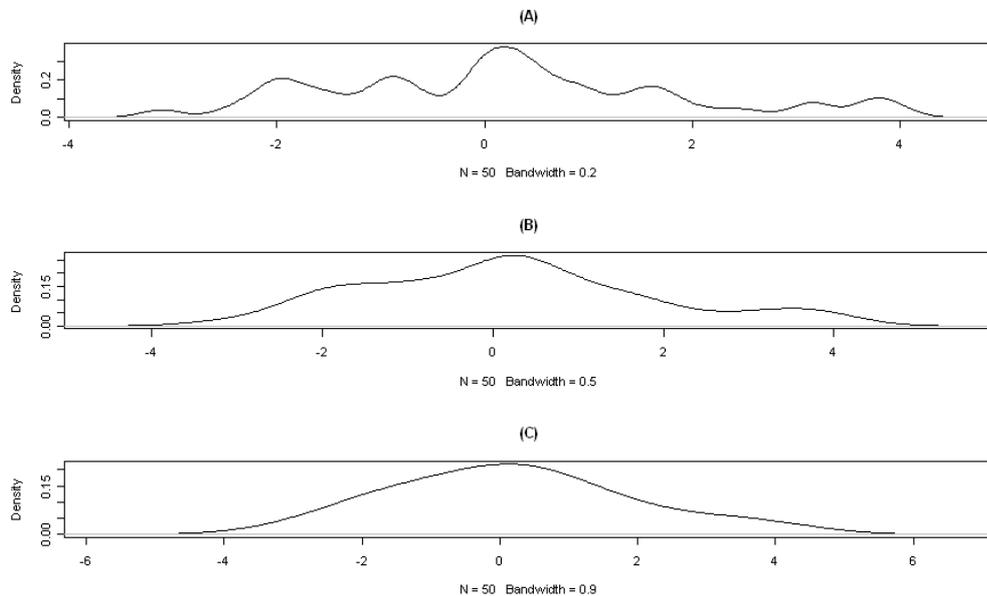


Figura 3.1.3: estimação de curva de densidades com diferentes parâmetros de suavização

Note que à medida que o parâmetro  $h$  vai ficando menor, a estimação torna-se gradativamente instável e errática, porém com a estimação da densidade mais próxima dos valores amostrados. Isto é, a variabilidade aumenta e o vício diminui. No entanto, à medida que  $h$  aumenta, as conclusões são análogas inversamente.

Os diferentes métodos de escolha do parâmetro  $h$  para estimação, assim como o método escolhido nesta monografia para escolha automática de  $h$ , serão discutidos posteriormente.

### 3.2 Estimação de Curvas de Regressão

Tratar-se-á agora da estimação de uma função de regressão. A fins de notação, o parâmetro de suavização  $h$  também poderá ser escrito como  $h_n$  indicando que seu valor pode depender do tamanho da amostra coletada  $n$ .

O objetivo principal é a modelagem não-paramétrica de expectativas condicionais, isto é:

$$Y = m(x) + \varepsilon = E(y | X = x) + \varepsilon \quad (3.2.1)$$

O objetivo é estimar  $m(x)$ . Por definição, pode-se escrever

$$m(x) = E(Y | X = x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y | x) dy = \int_{-\infty}^{+\infty} y \frac{f_{X,Y}(x, y)}{f_X(x)} dy \quad (3.2.2)$$

onde  $f_X(x)$ ,  $f_{X,Y}(x, y)$  e  $f_{Y|X}(y | x)$  são as densidades marginais de X, a densidade conjunta de X e Y e a densidade condicional de Y dado  $X = x$ , respectivamente. Estimando  $f_{X,Y}(x, y)$  e  $f_X(x)$  pode-se estimar  $m(x)$ . Esta foi a idéia que Nadaraya (1964) e Watson (1964) tiveram, o que gerou um estimador conhecido como Nadaraya-Watson. Uma estimativa núcleo de  $f_{X,Y}(x, y)$  é uma extensão bivariada de (3.1.2) e foi primeiramente estendida por Cacoullos (1996). Escreve-se esta estimativa por:

$$\hat{f}_{X,Y}(x, y) = \frac{1}{nh_{nx}h_{ny}} \sum_{i=1}^n K_x\left(\frac{x - x_i}{h_{nx}}\right) K_y\left(\frac{y - y_i}{h_{ny}}\right) \quad (3.2.1)$$

Usando (3.1.2) e (3.2.1), reparando que a integral da função ponderadora é um e tem esperança nula, supondo que  $h_{nx} = h_{ny} = h_n$  e que a função núcleo é simétrica, pode-se escrever o estimador de Nadaraya-Watson por:

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^n K_x\left(\frac{x - x_i}{h_n}\right) y_i}{\sum_{i=1}^n K_x\left(\frac{x - x_i}{h_n}\right)} \quad (3.2.2)$$

O estimador de Nadaraya-Watson pode ser visto como uma estimativa local de  $m(x)$  e um caso particular da classe dos estimadores polinomiais locais (Simonoff, 1996 e Pagan & Ullah, 1999). Atualmente, o estimador de Nadaraya-Watson é pouco utilizado, devido principalmente ao viés deste estimador nas fronteiras da distribuição das observações  $x_i$ .

Hastie e Tibshirani (1990) apresentam uma série de suavizadores que podem ser utilizados na estimação de curvas de regressão além do Nadaraya-Watson ou o simples média-móvel, como os “cubic smoothing splines” ou o “locally weighted running line

smoother (loess)”. Nesta monografia, usaremos o suavizador de Nadaraya-Watson e o caso linear da regressão polinomial local. Este último suavizador é detalhado a seguir.

### 3.2.1 Regressão Polinomial Local

Inicialmente por Stone (1977) e Cleveland (1979) os estimadores polinomiais locais estimam a função de regressão em um ponto particular através de um ajuste local de um polinômio de ordem  $p$ . O processo consiste em ajustar localmente este polinômio aos dados através de mínimos quadrados ponderados com auxílio da função *kernel* escolhida.

Note que no modelo (3.1) a função  $m(x)$  pode ser aproximada em um ponto específico através da expansão de Taylor utilizando-se de suas derivadas. A idéia por trás da regressão polinomial local é estimar uma função polinomial obtida através do truncamento desta expansão em cada ponto específico, ponderando a estimação de acordo com a vizinhança

Sejam  $m^{(1)}(x)$ ,  $m^{(2)}(x)$ , ...,  $m^{(p)}(x)$  as derivadas da função de regressão  $m(x)$  onde  $x$  é vizinho de  $x_0$ . A função  $m(x)$  pode ser escrita como

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2}(x - x_0)^2 + \dots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p \quad (3.2.1.1)$$

A função acima indica que podemos aproximar a função de regressão desconhecida  $m(x)$  usando esta expansão de Taylor. Então, se nós considerarmos um ponto de interesse  $x$  no domínio da variável aleatória  $X$ , pode-se definir o estimador polinomial local de  $m(x)$  como  $\hat{m}_p(x) = \hat{\beta}_0$ , com  $\hat{\beta}_0$  dado pela solução do seguinte problema de mínimos quadrados:

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \arg \min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n \{Y_i - \sum_{u=0}^p \beta_u (X_i - x)^u\}^2 K\left(\frac{X_i - x}{h}\right) \quad (3.2.1.2)$$

onde  $K(\cdot)$  é a função ponderadora *kernel* escolhida e  $h > 0$  é o parâmetro de suavização, ambos já citados anteriormente.

É importante salientar que o estimador polinomial local fornece as estimativas das derivadas da função  $g(x)$  como subproduto e que, diferentemente dos modelos de

mínimos quadrados paramétricos, os valores de  $\hat{\beta}_i$ , estimados a partir de (3.2.1.2), dependem do ponto específico  $x$  onde está se realizando a estimação local. Note também que, em cada ponto específico, estamos fazendo a estimação:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \hat{m}(x_0) \\ \hat{m}'(x_0) \\ \cdot \\ \cdot \\ \cdot \\ \hat{m}^{(p)}(x_0) \end{pmatrix}$$

Logo, se estamos interessados na estimação de  $m(x)$  apenas (para representação gráfica, por exemplo), estaríamos utilizando apenas o intercepto  $\hat{\beta}_0$  da estimação.

Além disto, se truncarmos a função  $m(x)$  na expansão (3.2.1.1) antes da primeira derivada, não estaríamos aproximando  $m(x)$  através de um polinômio, mas apenas estaríamos fazendo uma média ponderada de  $Y_i$  na vizinhança de  $x_i$  e recairíamos no caso do estimador de Nadaraya-Watson.

Diversos aspectos neste processo de suavização influenciam na estimativa não-paramétrica de  $m(x)$ . Determinar o parâmetro de suavização  $h$  que irá influenciar na estimação da curva num particular ponto, o grau do polinômio que será utilizado para aproximação de  $m(x)$ , assim como a escolha da função de ponderação dos pontos amostrais através do *kernel* também aqui se tornam indispensáveis.

Para mais detalhes sobre as propriedades dos estimadores de regressões polinomiais locais, ver Fan (1992).

### 3.2.2 Nadaraya-Watson e Regressão Linear Local

O grau  $p$  do polinômio determina parcialmente a sua complexidade. Porém, esta não é uma decisão crítica, pois é o parâmetro de suavização  $h$  que desempenha o papel mais importante no ajuste.

A escolha de uma ordem elevada para o polinômio pode reduzir o viés da estimação. No entanto, isso tende, simultaneamente, a aumentar a variância do modelo devido ao aumento do número de parâmetros. Desta maneira, é coerente escolher polinômios de baixa ordem.

Fan e Gijbels (1996) sugerem, com base em considerações teóricas e práticas, que o polinômio tenha grau  $p = r + 1$ , onde  $r$  é a ordem da derivada da função  $m(\cdot)$  a ser estimada. Consequentemente, é coerente escolhermos o caso particular em que  $p = 1$ , o

caso da regressão linear local, pois estamos interessados apenas na estimativa de  $m(x)$  e não nas suas derivadas. Ou seja:

$$m(x) = m(x_0) + m'(x_0)(x - x_0) + R \quad (3.2.2.1)$$

onde o termo  $R$ , considerado desprezível, depende de termos que envolvem a segunda derivada e derivadas de ordens maiores de  $m(x)$ . No caso local quadrático,

$$m(x) = m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2}(x - x_0)^2 + R \quad (3.2.2.2)$$

o termo  $R$  despreza os termos que envolvem a terceira derivada e derivadas de ordens maiores de  $m(x)$ , e assim sucessivamente.

Nesta monografia, serão utilizados o suavizador de Nadaraya-Watson e a regressão linear local.

Outra característica importante de regressão polinomial local, é que à medida que o parâmetro de suavização  $h$  aumenta, isto é, a ponderação entre os dados amostrais vai tornando-se cada vez mais homogênea, a estimação não-paramétrica se aproxima da estimação paramétrica linear onde as bases de  $X$  são termos que envolvem até o grau da estimação polinomial local utilizada. No limite em que  $h \rightarrow \infty$  a estimação polinomial local e a estimação linear paramétrica de mesma ordem são iguais.

Para ilustrar esta mínima diferença na estimação de  $m(x)$  entre diferentes graus de polinômio e esta relação com o caso paramétrico, foi gerada uma amostra aleatória de tamanho 100 a partir de uma equação quadrática:

$$Y_i = 4 + 1,32X_i - 3X_i^2 + \varepsilon_i \quad \varepsilon_i \sim N(0,6) \quad (3.2.2.3)$$

onde  $X$  é uma sequência de valores entre -3 e 3. Neste caso, a função  $m(x)$  a ser estimada é  $m(x) = 4 + 1,32x - 3x^2$ .

Foram adotados graus de polinômio  $p = 1$  e  $p = 2$  e função *kernel* gaussiana para a estimação de  $m(x)$  e as estimações  $\hat{m}(x)$ , representadas através da curva em preto, podem ser vistas nos gráficos a seguir:

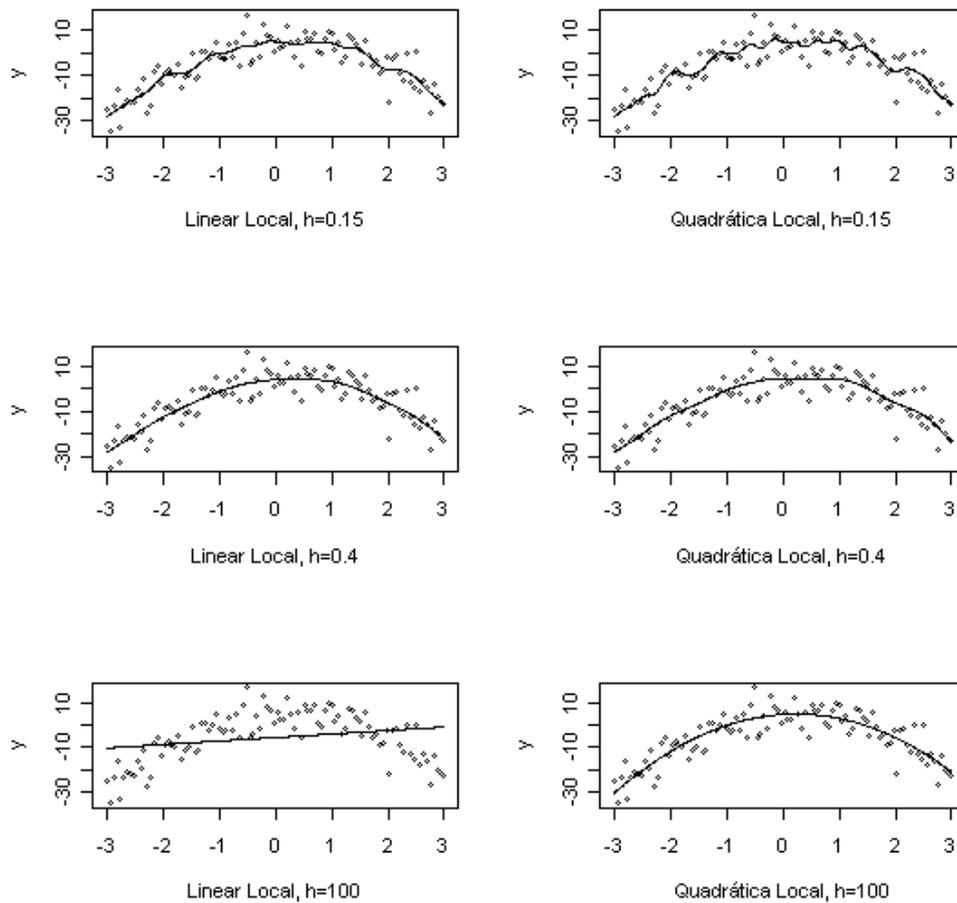


Figura 3.2.2.1: Comparação entre diferentes bandwidths e diferentes graus de polinômio

Como podemos analisar nos quatro gráficos superiores, não notamos diferença substancial quando comparados lado-a-lado (isto é, com mesmo parâmetro de suavização), ressaltando apenas um comportamento menos viciado, isto é, mais variável da estimação quadrática local quando  $h = 0,15$ . Esta diferença já não fica tão clara quando aumentamos para  $h = 0,4$  o parâmetro de suavização da estimação, notamos que as estimativas de  $\hat{m}(x)$  são praticamente equivalentes. No entanto, quando o parâmetro de suavização tende ao infinito, aproximado com  $h = 100$ , as estimações entre a regressão linear local e quadrática local são distintas entre si. De acordo com o que foi comentada em um parágrafo anterior, esta estimação local com  $h$  tendendo ao infinito e ponderando homogeneamente as observações, as estimativas coincidem com a estimação linear paramétrica polinomial.

Esta característica é ilustrada a seguir:

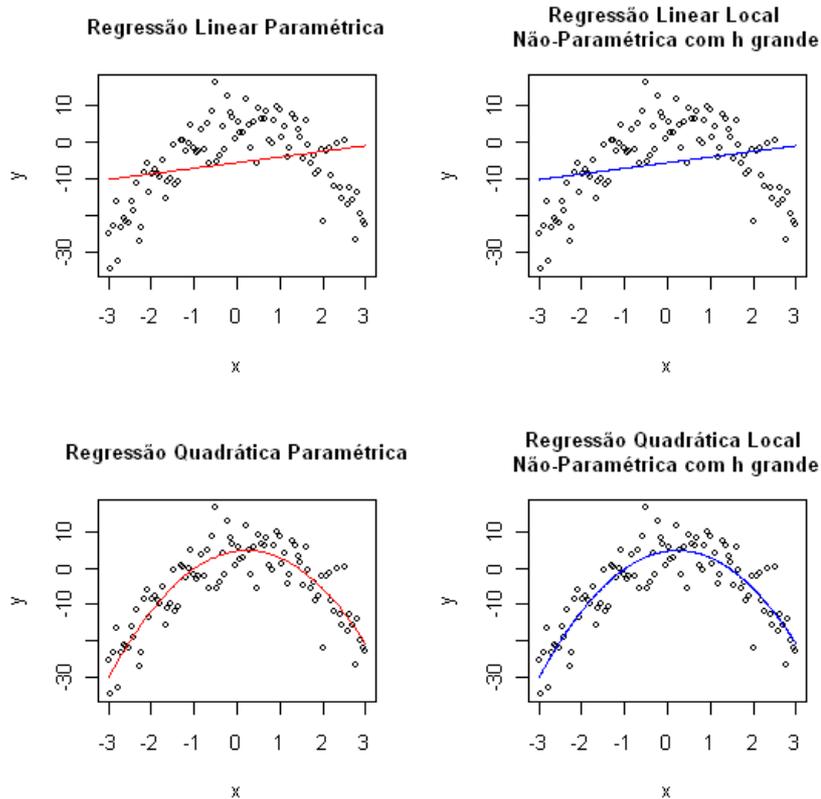


Figura 3.2.2.2: Comparação regressão paramétrica e polinomial não-paramétrica com  $h \rightarrow \infty$

As estimativas paramétricas são apresentadas nos gráficos à esquerda e as curvas estimadas ilustradas em vermelho. Note que a estimação paramétrica coincide com a estimação não-paramétrica, ilustrada em azul, quando  $h \rightarrow \infty$ . Devido à inevitável sobreposição entre estimação paramétrica e não-paramétrica e com intuito de visualização gráfica, as linhas foram plotadas em gráficos distintos.

A fim de ilustrar também o comportamento da comparação entre estimação polinomial local de  $\hat{m}(x)$  quando  $p = 1$  e  $p = 0$ , isto é, quando utilizamos o estimador linear local e de Nadaraya-Watson respectivamente, utilizamos o mesmo modelo descrito em (3.2.2.3), mantemos fixo tanto o *kernel* gaussiano, quanto o parâmetro de suavização e plotamos o gráfico abaixo:

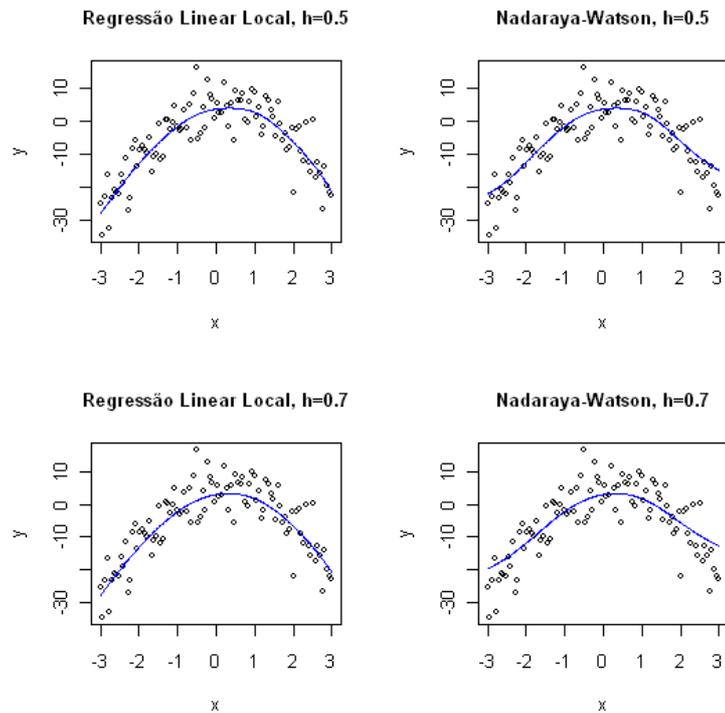


Figura 3.2.2.2: Comparação linear local e Nadaraya-Watson

Note no gráfico que o estimador de Nadaraya-Watson degenera nas fronteiras aumentando o vício de sua estimativa, característica que não acontece no caso linear local.

Wand e Jones (1995) afirmam que a estimação de pontos através de constantes, como faz o estimador de Nadaraya-Watson, é afetada por empecilhos como vieses nas fronteiras e nas regiões em que as derivadas são elevadas. Para mais detalhes, veja Simonoff (1996).

### 3.3 Problemas da Regressão Não-Paramétrica

Contrastando com a abordagem paramétrica, não é exigido dos estimadores não-paramétricos da função de regressão que a forma funcional de  $m(x)$  pertença, necessariamente, a uma classe limitada de funções. Ao contrário, a estimação não-paramétrica é bastante flexível, adaptando-se a amostra de maneira mais efetiva do que a sua contrapartida paramétrica. Entretanto, existem alguns problemas práticos e também teóricos que surgem quando se busca estimar um modelo multivariado irrestrito. Infelizmente, a convergência dos estimadores não-paramétricos torna-se mais lenta à medida que a dimensão do vetor de regressores aumenta. Esta situação faz com

que o problema de estimação necessite uma grande quantidade de observações quando temos muitos regressores, o que é frequentemente inviável na prática.

Outro problema surge quando se define vizinhanças em dimensões maiores ou iguais a dois para a ponderação local. Neste caso, existe a necessidade de assumir algum tipo de métrica que é difícil de justificar quando as variáveis são medidas em diferentes unidades ou são altamente correlacionadas (Buja, Hastie & Tibshirani, 1989). Além disto, de uma perspectiva prática, suavizadores multivariados são extremamente caros de computar, e a habilidade para visualizar e/ou interpretar a relação quando temos mais de dois regressores é virtualmente perdida.

O principal problema, que será detalhado a seguir, acerca de modelos não-paramétricos irrestritos é a característica de que a taxa de convergência dos estimadores é inversamente proporcional a adição de regressores no modelo, problema denominado *curse of dimensionality* (traduzido por “maldição da dimensionalidade”).

### 3.3.1 *Curse of Dimensionality*

Este problema identificado por Friedman & Stuetzle (1981) consiste no problema causado pelo aumento exponencial do número de variáveis independentes que se associa com a adição de dimensões em um espaço métrico. Isso quer dizer que estimativas de curvas e superfícies em pontos em vizinhanças com poucos pontos se tornarão menos locais quanto maior a dimensão do espaço.

Demonstra-se (Ibragimov e Hasminskii (1983) e Stone (1982)) que a melhor taxa de convergência é  $n^{(q/2q+d)}$ , onde  $d$  é a dimensão de  $X$  e a função a ser minimizada é continuamente diferenciável  $q$  vezes. Note que:

$$n^{(q/2q+d)} \propto \frac{1}{d}$$

e, portanto, no caso multivariado irrestrito, a característica da estimação local se perde com o aumento de  $d$ .

Para exemplificar, supomos que 100 pontos igualmente espaçados são suficientes para preencher um intervalo unitário com distâncias não superiores a 0,01 entre os pontos. Uma amostra de um cubo unitário equivale, num espaço de 10 dimensões, com espaçamento entre pontos adjacentes não superiores a 0,01, exigiria uma amostra de  $10^{20}$  pontos. Isto nos mostra que o conceito de “local” falha em grandes dimensões.

Pontos, que em um plano, parecem próximos, podem se afastar muito quando incluída uma terceira variável. Logo, estimar uma regressão local utilizando pontos muito

distantes se tornaria muito complicado. Isto porque estes pontos não guardam informações suficientes para estimar uma função que descreva de forma “eficiente” o comportamento geral da amostra.

A fim de ilustração gráfica, foi gerada uma amostra aleatória de tamanho 100 de três variáveis aleatórias e plotada em gráficos tridimensionais para inspecionar visualmente a distância entre os pontos. Um modelo contendo apenas uma variável regressora desta simulação da forma  $Y = X_1$  é plotado abaixo em diversos ângulos:

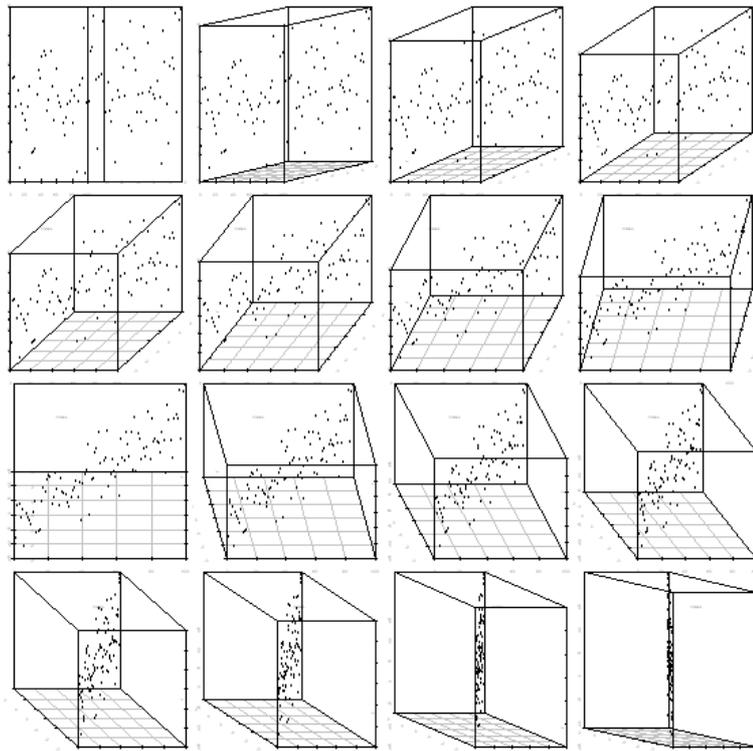


Figura 3.2.4.1: Gráficos de dispersão tridimensional de um modelo  $Y = X_1$

Note que os pontos estão próximos quando observados sobre a perspectiva bidimensional vista nos ângulos dos últimos gráficos. No entanto, quando adicionamos uma nova covariável no modelo e plotamos os “**mesmos**” pontos sob os **mesmos** ângulos, a característica de proximidade é perdida devido à inclusão da nova variável e, portanto, uma nova dimensão. Esta ilustração segue:

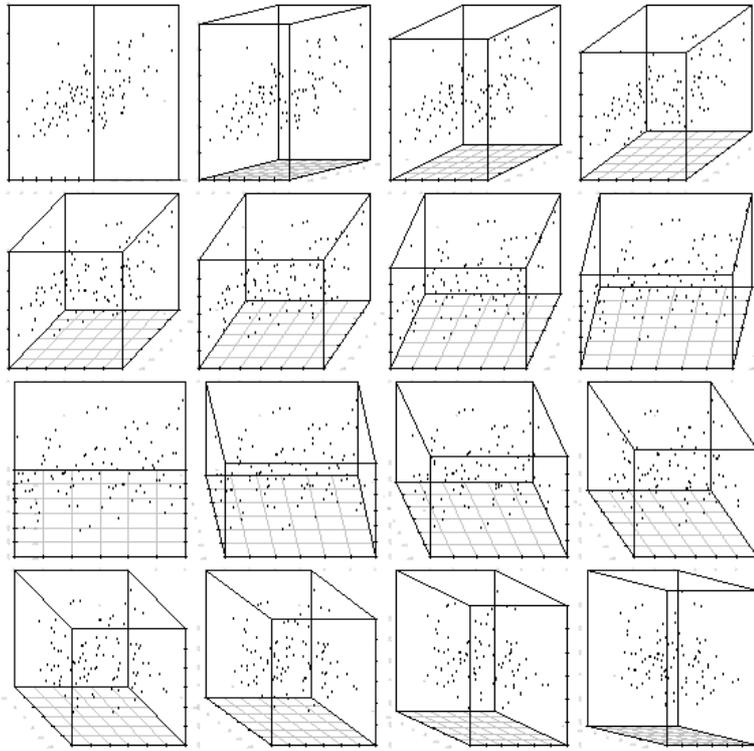


Figura 3.2.4.2: Gráficos de dispersão tridimensional de um modelo  $Y = X_1 + X_2$

A proximidade entre os pontos observada nos últimos gráficos na figura anterior agora é afetada por pontos mais afastados entre si e comprometendo a estimação de característica local. Para o leitor que não conseguiu visualizar de que maneira a adição de uma covariável afetou a característica de “proximidade amostral”, é recomendado que se rode no R, com o pacote `scatterplot3d`, a seguinte sequência de comandos para visualização em diversos ângulos (o gráfico de cima é o gráfico bivariado):

```
library(scatterplot3d)
x <- seq(-3, 3, l=50)
y <- x + rnorm(50, 0, sd=2)
u <- rnorm(50, 0, sd=6)
z <- x + y + u
par(mfrow=c(2,1))
repeat {
  for (i in 1:360) {
    scatterplot3d(x,y,z, cex.axis=0.2, cex.lab=0.2, mar=rep(0,4), type="p", xlim=c(-3,3),
    ylim=c(-15,15), zlim=c(-25,25), angle=i)
    scatterplot3d(x,y, cex.axis=0.2, cex.lab=0.2, mar=rep(0,4), type="p", angle=i)
  }
}
```

Ilustramos um exemplo prático e gráfico da maldição da dimensionalidade, porém tal problema pode ser remediado de algumas maneiras restringindo o espaço em que irá se executar a estimação, isto é, fazendo alguma suposição a priori sobre a natureza da expectativa condicional. Uma destas maneiras é discutida no próximo capítulo.

A estimação de regressões não-paramétricas possui diversos problemas como já sabidos da literatura. O principal deles é a maldição da dimensionalidade que é causada pela adição de múltiplos regressores no modelo que acarreta a exigência de uma amostra suficientemente grande para caracterizar a estimação local devido à baixa convergência dos estimadores. Para remediar tal problema, são propostos, para a estimação de modelos de regressão não-paramétricos múltiplos, os modelos aditivos (Hastie e Tibshirani, 1990) que são uma generalização do modelo de regressão linear. Por este motivo, foram citadas limitações desta última técnica no capítulo 2.

Os modelos aditivos são modelos semi-paramétricos que podem ser utilizados em qualquer cenário no qual uma abordagem distinta poderia ser tipicamente utilizada. Em sua versão generalizada (*modelos aditivos generalizados*), a variável resposta  $Y$  pode assumir diferentes distribuições de probabilidade, além da gaussiana, como dados de contagens, dados binários e categóricos, tanto ordinais quanto não ordinais.

Diferentes estimadores são viáveis a citar: o estimador *Backfitting Clássico (CBF)*, proposto por Buja et. al. (1989); o estimador de *Integração Marginal*, proposto por Newey (1994) e Linton & Nielsen (1995); o estimador de *2 estágios (2E)*, também conhecido como estimador *Híbrido*, proposto por Linton (1997) e Kim et. al. (1999), estimador *Smooth Backfitting (SBF)*, proposto por Mammen et. al. (1999), estimador *Spline-Backfitted*, proposto por Wang & Yang (2007).

Na presente monografia é abordada a estimação de modelos aditivos com métodos baseados em *kernels* através dos estimadores *Backfitting Clássico (CBF)*, *Smooth Backfitting (SBF)* e estimador *de 2 estágios (2E)*.

## 4.1 Definição

Problemas de regressão linear múltipla onde  $p > 1$  regressores são incluídos no modelo pode enfocar uma importante característica que se fez tão popular na inferência estatística: o modelo linear é *aditivo* no efeito dos regressores. Uma vez que é ajustado um modelo desta classe, podemos examinar os efeitos individuais de cada regressor separadamente, na ausência de interações. Modelos aditivos mantêm esta importante característica: eles são aditivos nos efeitos dos preditores.

Um modelo aditivo é definido por:

$$Y = \alpha + \sum_{d=1}^p m_d(x_d) + \varepsilon, \quad (4.1.1)$$

onde os erros podem assumir as suposições clássicas que podem ser relaxadas e as funções  $m_d$ 's são funções univariadas, obtidas através de algum processo de suavização, arbitrárias que são comumente chamadas de *componentes aditivos*.

Está implícita no modelo, para que se previna ambigüidade, a suposição de que

$$E(m_d(x_d)) = 0 \forall d, \quad (4.1.2),$$

uma vez que, caso não seja estabelecida, teríamos constantes livres em cada função univariada nos preditores. Esta condição é conhecida como *condição de identificação*. Isso nos leva a conduzir que  $E(Y) = 0$  quando se assume que a média já tenha sido subtraída do modelo.

Cada componente aditivo pode ser estimado separadamente dos demais, evitando assim a maldição da dimensionalidade, através desta estimação univariada. Foi mostrado por Stone (1985) que, se o modelo possui a estrutura aditiva, cada componente pode ser estimado com a taxa ótima de  $n^{q/(2q+1)}$  a qual não depende da dimensão  $d$ , mas apenas do grau de suavidade estabelecido  $q$ .

As funções estimadas de um modelo aditivo podem ser análogas aos coeficientes da regressão linear. No entanto, todas as “armadilhas” que são encontradas nos modelos lineares se aplicam nos modelos aditivos, e, de fato, podem ser mais severas. Não é aconselhável interpretar funções de variáveis que são insignificantes e deixá-las afetando outras mais importantes. Uma estratégia é selecionar e retirar as variáveis de maneira passo-à-passo, até o momento em que apenas variáveis importantes estão incluídas no modelo. Somente neste estágio é recomendado que se examine as funções ajustadas, sabendo que cada uma é necessária para a construção da superfície múltipla (Hastie e Tibshirani, 1990).

Hastie e Tibshirani (1990) explicam que, no momento em que cada variável é representada separadamente independente das demais, podem-se examinar as  $p$  funções, a fim de verificar os padrões parciais de comportamento na modelagem da resposta. Porém, tal facilidade implica que os modelos aditivos são aproximações generalizadas da verdadeira superfície de regressão. O que se espera é que seja estabelecida uma boa aproximação do verdadeiro modelo e que seja possível descobrir as variáveis explicativas mais importantes e suas relações com a resposta através dessa aproximação.

## 4.2 Estimação

Neste capítulo são apresentados os diferentes métodos de estimação de modelos aditivos estudados nesta monografia.

Para conveniência computacional, notação e exposição, um modelo bivariado é utilizado nesta monografia, mas as conclusões podem ser estendidas para mais dimensões. Seja  $(Y, X_1, X_2)$  um vetor aleatório com função de densidade conjunta  $f(y, x_1, x_2)$  tal que  $E(Y | X_1 = x_1, X_2 = x_2) = m(x_1, x_2) = \alpha + m_1(x_1) + m_2(x_2)$ , com  $E(m_1(x_1)) = E(m_2(x_2)) = 0$  e  $\text{Var}(Y | X_1 = x_1, X_2 = x_2) = \sigma^2$ . O principal interesse é na estimação de  $m(x_1, x_2) = \alpha + m_1(x_1) + m_2(x_2)$ . Seja  $\bar{y} = (y_1, \dots, y_n)'$ , e similarmente  $\bar{x}_1$  e  $\bar{x}_2$ . Em adição, seja  $\bar{m}_1(\bar{x}_1) = (m_1(x_{11}), \dots, m_1(x_{1n}))'$  e similarmente  $\bar{m}_2(\bar{x}_2)$ .

Uma vez que será utilizado um suavizador linear local nos estimadores sob estudo, iniciaremos introduzindo algumas notações. Seja  $K_{h_j}(\cdot) = K(\cdot/h_j)/h_j$  onde  $K(\cdot)$  é uma função kernel univariada e  $h_j, j=1,2$  são os parâmetros de suavização utilizados para a estimação de  $m_1$  e  $m_2$ , respectivamente. As matrizes suavizadoras linear local com respeito à  $\bar{x}_1$  e  $\bar{x}_2$  são definidas por

$$S_1 = \begin{pmatrix} s_1(x_{11}) \\ \cdot \\ \cdot \\ \cdot \\ s_1(x_{1n}) \end{pmatrix}, \text{ e } S_2 = \begin{pmatrix} s_2(x_{21}) \\ \cdot \\ \cdot \\ \cdot \\ s_2(x_{2n}) \end{pmatrix},$$

onde  $s_1(x_1), s_2(x_2) : \mathfrak{R} \rightarrow \mathfrak{R}^n$  são dadas por

$$s_1(x_1) = e(R_{X_1}(x_1)'W_{X_1}(x_1)R_{X_1}(x_1))^{-1}R_{X_1}(x_1)'W_{X_1}(x_1),$$

$$s_2(x_2) = e(R_{X_2}(x_2)'W_{X_2}(x_2)R_{X_2}(x_2))^{-1}R_{X_2}(x_2)'W_{X_2}(x_2)$$

onde  $e = (1, 0)$ ,  $W_{X_1}(x_1) = \text{diag}\{K_{h_1}(x_{1i} - x_1)\}_{i=1}^n$ ,  $W_{X_2}(x_2) = \text{diag}\{K_{h_2}(x_{2i} - x_2)\}_{i=1}^n$ ,

$R_{X_1}(x_1) = (\bar{1}_n, \bar{x}_1 - \bar{1}_n x_1)$ ,  $R_{X_2}(x_2) = (\bar{1}_n, \bar{x}_2 - \bar{1}_n x_2)$  e  $\bar{1}_n$  é um vetor de tamanho  $n$ .

Dado um estimador não-paramétrico bivariado  $\hat{m}(x_1, x_2)$  para  $m(x_1, x_2)$ , podemos definir os estimadores  $m_1(\cdot), m_2(\cdot)$  e  $\alpha$  como soluções do seguinte problema de minimização:

$$\iint \{ \hat{m}(x_1, x_2) - m_1(x_1) - m_2(x_2) - \alpha \}^2 dP(x_1, x_2) \quad (4.2.1)$$

sujeito a  $m_1 \in H_1, m_2 \in H_2, \alpha \in \mathfrak{R}$ ,

onde  $P(\cdot, \cdot)$  é uma medida conjunta, enquanto  $H_1$  e  $H_2$  são classes de funções cujos membros satisfazem as condições de identificação  $\iint m_1(x_1) dP(x_1, x_2) = 0$  e  $\iint m_2(x_2) dP(x_1, x_2) = 0$ . Note que, dadas as condições de identificação, um estimador adequado para  $\alpha$  é  $\bar{y}$ .

Após esta breve introdução sobre modelos aditivos, iremos agora apresentar os estimadores sob estudo da monografia.

### 4.3 *Backfitting* Clássico

As estimativas das funções  $m_d$  podem ser calculadas através de suavizadores lineares, já citados anteriormente. O modelo (4.1.1) imediatamente sugere um processo iterativo para estimar as funções  $m_d$ , para qualquer que seja o número  $k$  de variáveis. Podemos isolar cada componente individualmente e suavizá-lo com o resíduo parcial do lado esquerdo da equação.

Esta estimação será obtida através do processo iterativo descrito abaixo, em que as funções  $m_j$ , são vetores da forma  $\{m_j(x_{1j}), \dots, m_j(x_{nj})\}$  e  $D$  é o número de regressores:

(i) Início:  $m_j = m_j^{(0)}, j = 1, \dots, d$

(ii) Ciclo:  $j = 1, \dots, d, 1, \dots, d, \dots$

$$m_j = S_j \left( y - \alpha - \sum_{k \neq j} m_k(X_k) \mid X_j \right)$$

(iii) até que  $\sum_{i=1}^n \left( y_i - \sum_{d=1}^D m_d^{(m+1)}(x_d) \right)^2 - \sum_{i=1}^n \left( y_i - \sum_{d=1}^D m_d^{(m)}(x_d) \right)^2$  não mude, ou seja

menor que um nível de tolerância previamente especificado.

No entanto, no caso bivariado em que temos duas variáveis regressoras e ajustado por uma regressão linear local, o algoritmo *backfitting* converge para uma solução explícita. No problema não-paramétrico descrito em (4.2.1) se tomarmos a medida  $P(x_1, x_2)$  para ser a medida de probabilidade conjunta de  $X_1$  e  $X_2$ , isto é,

$dP(x_1, x_2) = f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$ , com  $f_{X_1, X_2}(x_1, x_2)$  sendo a densidade conjunta de  $X_1$  e  $X_2$ , a solução do problema de minimização deve satisfazer as seguintes equações,

$$m_1(x_1) = \int \hat{m}(x_1, x_2) \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 - \int m_2(x_2) \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 - \bar{y} \quad (4.3.1)$$

$$m_2(x_2) = \int \hat{m}(x_1, x_2) \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)} dx_1 - \int m_1(x_1) \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)} dx_1 - \bar{y} \quad (4.3.2)$$

onde  $f_{X_1}(\cdot)$  e  $f_{X_2}(\cdot)$  são densidades marginais de  $X_1$  e  $X_2$ , respectivamente. Substituindo a expectativa condicional que aparece em (4.3.1) e (4.3.2) com o suavizador linear local apropriado, o estimador *Backfitting* Clássico pode ser expresso como a solução para

$$\begin{pmatrix} I_n & S_1^* \\ S_2^* & I_n \end{pmatrix} \begin{pmatrix} \hat{m}_1^{CB}(\bar{x}_1) \\ \hat{m}_2^{CB}(\bar{x}_2) \end{pmatrix} = \begin{pmatrix} S_1^* \\ S_2^* \end{pmatrix} \bar{y}, \quad (4.3.4)$$

onde  $S_d^* = (I_n - 1/n \bar{1}_n \bar{1}_n')$   $S_d \equiv D_n S_d$  e  $d = 1, 2$  e  $I_n$  é uma matriz identidade. Na prática o sistema (4.3.4) é resolvido usando o algoritmo *backfitting*, porém, como comentado anteriormente, no caso bivariado, quando se usa o estimador linear local, o algoritmo de *backfitting* converge para uma solução explícita para  $\bar{m}_1(X_1)$  e  $\bar{m}_2(X_2)$  dadas por

$$\bar{m}_1^{CB}(X_1) = (I_n - (I_n - S_1^* S_2^*)^{-1} (I_n - S_1^*)) Y$$

e

$$\bar{m}_2^{CB}(X_2) = (I_n - (I_n - S_2^* S_1^*)^{-1} (I_n - S_2^*)) Y$$

Quando se usa o estimador linear local, Opsomer & Ruppert (1997, 1998) derivam uma série de resultados (para grandes amostras) que serão descritos abaixo. No nosso caso uma solução existe se:

A1: A função núcleo  $K$  é limitada, contínua, com suporte compacto e sua primeira derivada tem um número finito de mudanças de sinal sobre o seu suporte. Além disso,

$$\mu_j(K) = \int u_j K(u) du = 0 \text{ para todo } j \text{ ímpar e } \mu_2(K) \neq 0.$$

A2: As densidades  $f(x_1, x_2)$ ,  $f_{x_1}(x_1)$  e  $f_{x_2}(x_2)$  são limitadas, contínuas e tem suporte compacto, e suas derivadas tem um número finito de mudanças de sinal sobre o seu suporte. Além disso,  $f_{x_1}(x_1) > 0$  e  $f_{x_2}(x_2) > 0$  para todo  $(x_1, x_2) \in \text{supp}(f)$  e

$$\sup \left| \frac{f(x_1, x_2)}{f_{x_1}(x_1)f_{x_2}(x_2)} - 1 \right| < 1.$$

A3: Quando  $n \rightarrow \infty$ ,  $h_{1n}, h_{2n} \rightarrow 0$  e  $nh_{1n} / \log(n), nh_{2n} / \log(n) \rightarrow \infty$ .

A4: As segundas derivadas de  $m_1$  e  $m_2$  existem e são limitadas e contínuas.

#### 4.4 *Smooth Backfitting*

O estimador linear local *Smooth Backfitting* é motivado pelo seguinte problema de minimização

$$\iint \sum_{i=1}^n \{Y_i - \alpha - m_1(x_1) - m_2(x_2) - m_1^{(1)}(x_1)(x_{1i} - x_1) - m_2^{(1)}(x_2)(x_{2i} - x_2)\}^2 K_{h_1}(x_{1i} - x_1) K_{h_2}(x_{2i} - x_2) dx_1 dx_2 \quad (4.4.1)$$

sujeito às condições de identificação

$$\int \sum_{i=1}^n m_1(x_1) K_{h_1}(x_{1i} - x_1) dx_1 = \int \sum_{i=1}^n m_2(x_2) K_{h_2}(x_{2i} - x_2) dx_2 = 0 \quad (4.4.2)$$

Note que a minimização é com respeito à  $\alpha$ ,  $m_1(x_1)$ ,  $m_2(x_2)$  e suas primeiras derivadas. O valor de  $\alpha$  pode simplesmente ser estimado através de  $\bar{y}$ , então as condições de primeira ordem da minimização descrita acima com respeito à  $m_1(x_1)$  e  $m_1^{(1)}(x_1)$  são dadas por

$$\begin{pmatrix} m_1^{SB}(x_1) \\ m_1^{(1),SB}(x_1) \end{pmatrix} = \begin{pmatrix} \tilde{m}_1(x_1) \\ \tilde{m}_1^{(1)}(x_1) \end{pmatrix} - \hat{M}_{x_1}(x_1)^{-1} \int \hat{S}_{x_1, x_2}(x_1, x_2) \begin{pmatrix} m_2^{SB}(x_2) \\ m_2^{(1),SB}(x_2) \end{pmatrix} dx_2 \quad (4.4.3)$$

onde  $\begin{pmatrix} \tilde{m}_1(x_1) \\ \tilde{m}^{(1)}(x_1) \end{pmatrix}$  é a projeção linear local de  $(\bar{y} - \bar{1}_n \bar{y})$  no subespaço de  $\mathfrak{R}^n$  onde  $\bar{x}$  admite valores e

$$\hat{M}_{X_1}(x_1) = \begin{pmatrix} \hat{f}_{X_1}(x_1) & \hat{f}_{X_1}^{X_1}(x_1) \\ \hat{f}_{X_1}^{X_1}(x_1) & \hat{f}_{X_1}^{X_1 X_1}(x_1) \end{pmatrix}, \quad \hat{S}_{X_1 X_2}(x_1, x_2) = \begin{pmatrix} \hat{f}_{X_1 X_2}(x_1, x_2) & \hat{f}_{X_1 X_2}^{X_2}(x_1, x_2) \\ \hat{f}_{X_1 X_2}^{X_1}(x_1, x_2) & \hat{f}_{X_1 X_2}^{X_1 X_2}(x_1, x_2) \end{pmatrix}$$

com

$$\hat{f}_{X_1}(x_1) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)$$

$$\hat{f}_{X_1}^{X_1}(x_1) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)(x_{1i} - x_1)$$

$$\hat{f}_{X_1}^{X_1 X_2}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)(x_{1i} - x_1)(x_{2i} - x_2)$$

$$\hat{f}_{X_1 X_2}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)K_{h_2}(x_{2i} - x_2)$$

$$\hat{f}_{X_1 X_2}^{X_2}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)K_{h_2}(x_{2i} - x_2)(x_{2i} - x_2)$$

$$\hat{f}_{X_1 X_2}^{X_1 X_2} = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_{1i} - x_1)K_{h_2}(x_{2i} - x_2)(x_{1i} - x_1)(x_{2i} - x_2)$$

Condições de primeira ordem similares com as de (4.4.3) podem ser definidas para  $m_2(x_2)$  e  $m_2^{(1)}(x_2)$ . Com valores iniciais dados por  $m_j^{SB}(\cdot), m_j^{(1), SB}(\cdot), \tilde{m}_j(\cdot), m_j^{(1)}(\cdot)$  para  $j=1,2$ , o estimador *Smooth Backfitting* é obtido através de um procedimento iterativo de (4.4.3) e seu análogo com respeito a  $X_2$ , até que  $m_j^{SB}(\cdot)$ ,  $j=1,2$  convirja de acordo com um critério previamente estabelecido.

#### 4.5 Estimador de 2 estágios (Estimador Híbrido)

O estimador de dois estágios busca resolver os problemas associados a regressões multivariadas irrestritas do estimador de Integração Marginal (Linton & Nielsen, 1995 e Silva, 2001) levando em consideração a dependência entre  $X_1$  e  $X_2$  na resolução das equações (4.3.1) e (4.3.2). Isto é alcançado estimando

$$\int m_2(x_2) \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 \equiv E(m_2(x_2) | X_1 = x_1) \quad \text{através} \quad s_2(x_2) \bar{m}_1^{MI}(\bar{x}_1), \quad \text{onde}$$

$$\bar{m}_1^{MI}(\bar{x}_1) = (m_1^{MI}(\bar{x}_{11}), \dots, m_1^{MI}(\bar{x}_{1n})), \bar{m}_2^{MI}(\bar{x}_2) = (m_2^{MI}(\bar{x}_{21}), \dots, m_2^{MI}(\bar{x}_{2n})).$$

Em adiç3o,

$$\int \hat{m}(x_1, x_2) \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_1}(x_1)} dx_2 \equiv E(\hat{m}(x_1, x_2) | X_1 = x_1) \quad (4.5.1)$$

e

$$\int \hat{m}(x_1, x_2) \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_2}(x_2)} dx_1 \equiv E(\hat{m}(x_1, x_2) | X_2 = x_2) \quad (4.5.2)$$

s3o estimados respectivamente por  $s_1(x_1)\bar{y}$  e  $s_2(x_2)\bar{y}$ . Kim et. al. (1999) consideram o caso onde  $m_1^{MI}$  e  $m_2^{MI}$  s3o baseados em uma estimativa bivariada chamada de *Nadaraya-Watson internalizada* para  $m(x_1, x_2)$  e definem o estimador de 2 est3gios como,

$$\bar{m}_1^{2S}(x_1) = s_1(x_1)(\bar{y} - \bar{m}_2^{MI}(\bar{x}_2) - 1_n \bar{y}) = s_1(x_1)(\bar{y} - \bar{\gamma}_2^P(\bar{x}_2)) \quad (4.5.3)$$

e

$$\bar{m}_2^{2S}(x_2) = s_2(x_2)(\bar{y} - \bar{m}_1^{MI}(\bar{x}_1) - 1_n \bar{y}) = s_2(x_2)(\bar{y} - \bar{\gamma}_1^P(\bar{x}_1)) \quad (4.5.4)$$

onde  $\bar{\gamma}_1^P(\bar{x}_1) = (\gamma_1^P(x_{11}), \dots, \gamma_1^P(x_{1n}))'$  e  $\bar{\gamma}_2^P(\bar{x}_2)$  3e similarmente definido por

$$\gamma_1^P(x_{1i}) = \frac{1}{n} \sum_{j=1}^n K_{g_1}(x_{1j} - x_{1i}) \frac{\hat{f}_{X_2}(x_{2j})}{\hat{f}_{X_1 X_2}(x_{1j}, x_{2j})} y_j \quad (4.5.5)$$

$$\gamma_2^P(x_{2i}) = \frac{1}{n} \sum_{j=1}^n K_{g_2}(x_{2j} - x_{2i}) \frac{\hat{f}_{X_1}(x_{1j})}{\hat{f}_{X_1 X_2}(x_{1j}, x_{2j})} y_j \quad (4.5.6)$$

onde  $\hat{f}_{X_1}(x_{1i}), \hat{f}_{X_2}(x_{2i}), f_{X_1 X_2}(x_{1i}, x_{2i})$  s3o estimativas de densidades baseadas em *kernel* com par3metros de suavizaç3o  $g_1$  e  $g_2$  associadas com  $X_1$  e  $X_2$ , respectivamente.

## 5 Escolha do Parâmetro de Suavização

Como comentado anteriormente, um dos passos mais importantes na estimação de modelos de regressão não-paramétrica é a escolha dos parâmetros de suavização  $h_n$  (também chamado de *janela*). Em essência, uma vez escolhido o suavizador a ser utilizado, a escolha de  $h_n$  é basicamente a escolha do estimador, pois as estimativas para  $m_i$  variam consideravelmente em função de  $h_n$ .

Um critério apropriado (veja Ruppert & Wand, 1994 e Ruppert, Sheater & Wand, 1995) é o MISE (*mean integrated squared error*) ponderado condicional dado por:

$$MISE(\hat{m}_p(\cdot; h_n) | X_1, \dots, X_n) = E \int [\{\hat{m}_p(x; h_n) - m(x)\}^2 | X_1, \dots, X_n] f_X(x) dx \quad (5.1)$$

onde  $f_X(x)$  representa a densidade de  $X$  com suporte  $[a, b]$ . Assuma que a variância dos erros é  $\sigma^2$ . Para  $p$  ímpar Ruppert & Wand (1994) mostram que

$$MISE(\hat{m}_p(\cdot; h_n) | X_1, \dots, X_n) = \left[ \frac{h_n^{p+1} \mu_{p+1}(K_{(p)})}{(p+1)!} \right]^2 \int m^{(p+1)}(x)^2 f_X(x) dx + \frac{R(K_{(p)}) \sigma^2 (b-a)}{nh_n} + o_p[h_n^{2p+2} + (nh_n)^{-1}] \quad (5.2)$$

onde  $\mu_j(K) = \int u^j K(u) du$ ,  $K_{(p)}(u) = \{ | M_p(u) | | | N_p | \} K(u)$ ,  $N_p$  é uma matriz  $(p+1) \times (p+1)$  com o  $(i,j)$  éximo elemento igual a  $\mu_{i+j+2}$ ,  $M_p(u)$  é o mesmo que  $N_p$  mas com a primeira coluna trocada por  $(1, u, u^2, \dots, u^p)'$  e  $R(K_{(p)}) = \mu_0(K_{(p)}^2)$ .

O minimizador de (5.2) é assintoticamente

$$\tilde{h}_n = \left[ \frac{(p+1)(p!)^2 R(K_{(p)}) \sigma^2 (b-a)}{2n \mu_{p+1}(K_{(p)})^2 \int m^{(p+1)}(u)^2 f_X(u) du} \right]^{1/(2p+3)} \quad (5.3)$$

se  $\int m^{(p+1)}(u)^2 f_X(u) du$  for diferente de zero. Um critério conveniente que utiliza apenas valores ajustados aos pontos observados é o MASE (*mean averaged squared*

*error*) condicional, discutido em Härdle, Hall & Marron (1988). No caso univariado, o MASE de  $m$  pode ser escrito como

$$\begin{aligned} MASE(\hat{m}_p(\cdot; h_n) | X_1, \dots, X_n) &= \frac{1}{n} \sum_{i=1}^n E\{(\hat{m}_p(x_i; h_n) - m(x)) | X_1, \dots, X_n\}^2 \\ &= \frac{1}{n} \sum_{i=1}^n E(\hat{m}_p(x_i; h_n | X_1, \dots, X_n) - m_p(x_i; h_n))^2 + \frac{1}{n} \sum_{i=1}^n V(\hat{m}_p(x_i; h_n | X_1, \dots, X_n)) \end{aligned} \quad (5.4)$$

Note que (5.4) é uma aproximação discreta a (5.1). Além disto, o primeiro termo da expansão representa o vício e o segundo a variância deste critério.

O critério escolhido na presente monografia para a escolha do parâmetro de suavização  $h_n$  para os estimadores é obtida pelo corolário 4.2 de Opsomer & Ruppert (1997), que é a aproximação assintótica para o MASE condicional quando o modelo aditivo é ajustado pela regressão linear local. Denotada por AMASE, esta aproximação assintótica é dada por:

$$\begin{aligned} AMASE(h_{1n}, h_{2n} | X_1, X_2) &= \frac{\mu_2(K_{(1)})^2}{4} (h_{1n}^4 \theta_{11} + h_{1n}^2 h_{2n}^2 \theta_{12} + h_{2n}^4 \theta_{22}) + \\ &\quad \sigma^2 R(K_{(1)}) \left( \frac{b_{x_1} - a_{x_1}}{nh_{1n}} + \frac{b_{x_2} - a_{x_2}}{nh_{2n}} \right) \end{aligned} \quad (5.5)$$

A escolha de cada  $h_n$  para os diferentes estimadores é detalhada a seguir.

## 5.1 Escolha de $h_n$ para *Backfitting* Clássico e *Smooth Backfitting*

O método de seleção de  $h_n$  considerado para todos os estimadores depende da obtenção de aproximações assintóticas adequadas para o critério MASE. No caso do estimador do *Backfitting* Clássico, Opsomer e Ruppert (1997) mostraram que quando

$\frac{nh_1}{\log n}, \frac{nh_2}{\log n} \rightarrow \infty$  é possível obter aproximações assintóticas para o vício e variância

condicionais de  $m_1^{BC}(x_1)$  e  $m_1^{BC}(x_2)$ . Em geral, não existem expressões analíticas para os parâmetros  $h_1$  e  $h_2$ . No caso especial onde  $X_1$  e  $X_2$  são independentes, expressões para o  $h_1$  e  $h_2$  ótimos, no sentido que minimizam o critério AMASE, e que serão utilizadas nesta monografia, para o *Backfitting* Clássico são dados por

$$h_1^{CBF} = \left( \frac{\sigma^2 R_K}{n \mu_2^2 \theta_{11}} n^{-1} \sum_{i=1}^n f_{X_1}(x_{1i})^{-1} \right)^{1/5} \text{ e } h^{CBF} = \left( \frac{\sigma^2 R_K}{n \mu_2^2 \theta_{22}} n^{-1} \sum_{i=1}^n f_{X_2}(x_{2i})^{-1} \right)^{1/5} \quad (5.1.1)$$

onde,

$$\theta_{11} = n^{-1} \sum_{i=1}^n (m_1^{(2)}(x_{1i}) - E(m_1^{(2)}(X_1)))^2, \theta_{22} = n^{-1} \sum_{i=1}^n (m_1^{(2)}(x_{2i}) - E(m_1^{(2)}(X_2)))^2 \text{ com } m_1^{(2)}(.) \text{ e } m_2^{(2)}(.).$$

As expressões das janelas para o *Backfitting* Clássico e para *Smooth Backfitting* são equivalentes no sentido da expressão analítica que apresentam. Para maiores detalhes ver Martins & Yang (2007).

## 5.2 Escolha de $h_n$ para o Estimador de 2 estágios

Para a escolha do parâmetro de suavização para o estimador de 2 estágios baseados na minimização do critério AMASE, necessitamos garantir certas condições sobre a velocidade das taxas de convergência das janelas para zero, que, se satisfeitas, devem satisfazer o seguinte sistema de equações:

$$(h_1^{2S})^5 \mu_2^2 \phi_{11} + (h_2^{2S})^2 (h_1^{2S})^3 \mu_2 \phi_{12} = \sigma^2 R_K \left( \frac{1}{n} \sum_{i=1}^n f_{X_1}(x_{1i})^{-1} \right) \quad (5.2.1)$$

$$(h_2^{2S})^5 \mu_2^2 \phi_{22} + (h_1^{2S})^2 (h_2^{2S})^3 \mu_2 \phi_{12} = \sigma^2 R_K \left( \frac{1}{n} \sum_{i=1}^n f_{X_2}(x_{2i})^{-1} \right) \quad (5.2.2)$$

onde

$$\phi_{11} = \frac{1}{n} \sum_{i=1}^n \left( m_1^{(2)}(x_{1i}) - E(m_1^{(2)}(X_1)) + \int f_{X_1}^{(2)}(v) \gamma_1(v) dv \right)^2$$

$$\phi_{22} = \frac{1}{n} \sum_{i=1}^n \left( m_2^{(2)}(x_{2i}) - E(m_2^{(2)}(X_2)) + \int f_{X_2}^{(2)}(v) \gamma_2(v) dv \right)^2$$

$$\phi_{12} = \frac{1}{n} \sum_{i=1}^n \left( m_1^{(2)}(x_{1i}) - E(m_1^{(2)}(X_1)) + \int f_{X_1}^{(2)}(v) \gamma_1(v) dv \right) \left( m_2^{(2)}(x_{2i}) - E(m_2^{(2)}(X_2)) + \int f_{X_2}^{(2)}(v) \gamma_2(v) dv \right)$$

Para maiores detalhes ver Martins & Yang (2007).

### 5.3 Janelas Estimadas

Uma das maneiras de se estimar o parâmetro de suavização  $h_n$  é utilizar os métodos *plug-in* para *estimar* o verdadeiro valor da janela. O princípio básico dos métodos *plug-in* é a substituição direta de estimativas de parâmetros desconhecidos na expressão analítica  $h_n$ , após ter sido minimizado o critério a ser utilizado na estimação não-paramétrica.

No caso do estimador *Backfitting* Clássico e *Smooth Backfitting*, a estratégia de estimação utilizada para obter estimativas para  $\sigma^2$ ,  $\theta_{11}$  e  $\theta_{22}$  e substituí-los diretamente em (5.1.1) foi a que segue em Opsomer & Ruppert (1998). Assumimos que  $f_{X_1}$  e  $f_{X_2}$  são densidades uniforme sobre um suporte compacto e os termos  $n^{-1} \sum_{i=1}^n f_{X_1}(x_{1i})^{-1}$  e  $n^{-1} \sum_{i=1}^n f_{X_2}(x_{2i})^{-1}$  são estimados por  $\max_i(x_{1i}) - \min_i(x_{1i})$  e  $\max_i(x_{2i}) - \min_i(x_{2i})$ , respectivamente. Ou seja, pelas amplitudes amostrais geradas.

Uma vez que o *Smooth Backfitting* compartilha a mesma solução analítica do que *Backfitting* Clássico, as mesmas janelas foram usadas para ambos.

Os parâmetros de suavização do estimador de 2 estágios são obtidos pela solução do sistema não-linear descrito em (5.2.1) e (5.2.2). As quantidades  $\phi_{d\gamma}$  devem ser estimadas conjuntamente com  $f_{X_1}, f_{X_2}$  e  $\sigma^2$ . A estimação de  $\phi_{d\gamma}$  depende da estimação de duas partes:  $m_d^{(2)}(x_{1i}) - E(m_d^{(2)})$  e  $\int f_{X_1}^{(2)} \gamma_1(v) dv$  (análogo para  $X_2$ ). O primeiro termo é estimado através do *backfitting* clássico, o segundo termo pode ser

interpretado como  $E\left(\gamma_1(v) \frac{f_{X_1}^{(2)}(v)}{f_{X_1}(v)}\right)$  que é estimado por  $\frac{1}{n} \sum_{i=1}^n \hat{\gamma}_1(x_{1i}) \frac{\hat{f}_{X_1}^{(2)}(x_{1i})}{\hat{f}_{X_1}(x_{1i})}$ , onde

$\hat{\gamma}_1$  vem de um CBF preliminar e  $\hat{f}$  é estimada por um estimador de densidade com parâmetro de suavização “regra de bolso” de Silverman. O parâmetro  $\sigma^2$  é estimado como no caso do CBF.

### 5.3 Janelas Verdadeiras

As janelas verdadeiras consistem na eliminação do ruído causado pelos métodos *plug-in* de estimação. Este ruído é gerado pela estimação dos parâmetros nas expressões das janelas ótimas.

Uma vez que estamos trabalhando com um exercício de simulação de Monte Carlo podemos saber os valores exatos destas quantidades, pois eles são obtidos pelo processo gerador dos dados.

## 6 Processo gerador dos dados

Neste capítulo explicaremos o processo gerador dos dados utilizados neste estudo, bem como as especificações do modelo aditivo bivariado proposto.

Foi realizado um estudo de Monte Carlo para avaliar o desempenho dos estimadores *Backfitting* Clássico, *Smooth Backfitting* e Estimador de 2 Estágios em amostras finitas para uma regressão aditiva bivariada. Os suavizadores utilizados foram os estimadores de Nadaraya-Watson e linear local.

Assumiu-se que  $x_{1i}$  e  $x_{2i}$ ,  $i=1, \dots, n$  formam uma sequência de realizações de um vetor aleatório com distribuição normal bivariada com densidade conjunta dada por:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \stackrel{iid}{\sim} N \left( \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix}, \begin{pmatrix} 1/9 & c/9 \\ c/9 & 1/9 \end{pmatrix} \right)$$

onde  $c=0, \frac{1}{4}, \frac{3}{4}$  fornece a desejada correlação entre as variáveis aleatórias  $x_1$  e  $x_2$ . Estudaremos diferentes correlações, pois um dos objetivos é avaliar se a dependência entre os regressores afeta o desempenho das janelas, bem como dos estimadores sob estudo. Uma das suposições requeridas para garantir as expressões para média e variância condicional dos estimadores é de que  $f_{x_1, x_2}$  possuem suporte compacto. Para satisfazê-la, descartamos as simulações fora do intervalo  $[0,1]$  e reamostramos para garantir a presença dentro deste intervalo. A variável resposta  $y_i$  foi gerada a partir do seguinte modelo bivariado:

$$Y_i = m_1(X_{1i}) + m_2(X_{2i}) + \varepsilon_i$$

onde

$$m_1(X_1) = \frac{16}{3}X_1^3 - 8X_1^2 + \frac{11}{3}X_1$$
$$m_2(X_2) = \frac{1}{4}\text{sen}(4\pi X_2)$$

A fim de ilustração, os formatos das funções utilizadas no processo de geração dos dados assim como a superfície usada seguem a seguir:

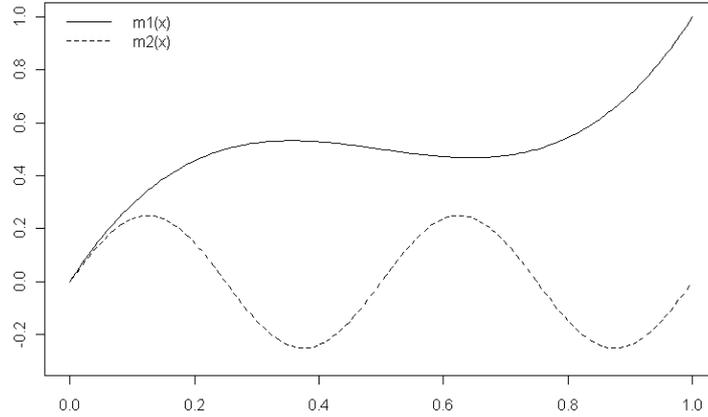


Figura 6.1: formato das funções simuladas dos dois componentes  $m_1(x)$  e  $m_2(x)$

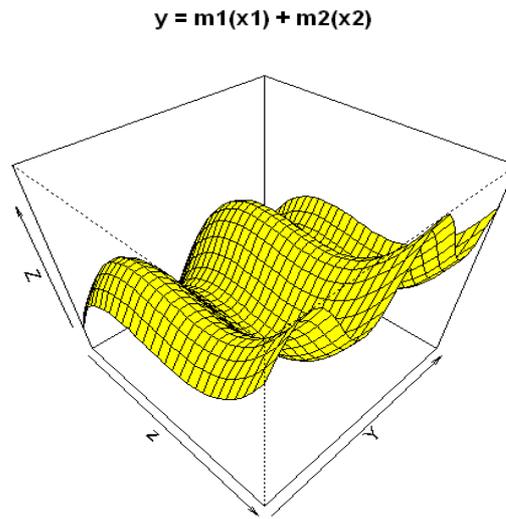


Figura 6.2: superfície utilizada no processo gerador dos dados

Serão geradas amostras de tamanho  $n=100$ ,  $250$  e  $400$  e o número de replicações será de  $1000$ . Além disto, para as expressões das janelas estimadas, para a função núcleo gaussiana obtemos:  $\mu_1(K) = 0$ ,  $\mu_2(K) = 1$  e  $R(K_{(1)}) = (\sqrt{2\pi})^{-1}$ .

O objetivo principal do trabalho é comparar o desempenho dos estimadores em amostras finitas. Para isto, nós computaremos o erro quadrático médio (*Average Squared Error*) da  $r$ -ésima replicação como sendo:

$$ASE^r(\hat{m}) = \frac{1}{n} \sum_{i=1}^n (y_i - (\bar{y}^r + \hat{m}_1^r(x_{1i}) + \hat{m}_2^r(x_{2i})))^2.$$

Após isso, calcularemos a média ao longo das replicações para estimar a média dos erros quadráticos médios (*Mean Average Squared Error*).

## 7 Resultados e Conclusões

Um exercício de simulação foi conduzido para avaliar o desempenho de diferentes maneiras de estimar o modelo aditivo proposto. Cabe salientar que tal estudo é estritamente restritivo devido à infinidade de possibilidades de combinações que montam os cenários do Monte Carlo tais como a seleção dos métodos de suavização, o tamanho da amostra, a correlação entre os regressores, a função *kernel* escolhida, escolha do parâmetro de suavização, dentre inúmeros outros fatores. Além disto, cabe enfatizar também, que o modelo possui um formato que apresenta diversos máximos e mínimos locais dentro do suporte como pode ser visto na Figura 6.2, o que pode caracterizar uma superfície mais difícil de estimar.

Como comentado no capítulo anterior, analisamos três tamanhos diferentes de amostra e três níveis de correlação e com essas combinações construímos nove cenários distintos:

- 1° Cenário)  $n = 100$  e  $\rho = 0$ ;
- 2° Cenário)  $n = 100$  e  $\rho = 0,25$ ;
- 3° Cenário)  $n = 100$  e  $\rho = 0,75$ ;
- 4° Cenário)  $n = 250$  e  $\rho = 0$ ;
- 5° Cenário)  $n = 250$  e  $\rho = 0,25$ ;
- 6° Cenário)  $n = 250$  e  $\rho = 0,75$ ;
- 7° Cenário)  $n = 400$  e  $\rho = 0$ ;
- 8° Cenário)  $n = 400$  e  $\rho = 0,25$ ;
- 9° Cenário)  $n = 400$  e  $\rho = 0,75$ .

O interesse principal da monografia é a comparação do *Backfitting* Clássico, *Smooth Backfitting* e Estimador de 2 estágios (2E) com diferentes métodos de suavização em amostras finitas. Para isto, escolhemos dois métodos de suavização: linear local e Nadaraya-Watson. Escolhemos como parâmetro de suavização utilizando  $h_n$  que minimiza o AMASE utilizando o estimador linear local em cada iteração. Foi de interesse avaliar o desempenho de estimadores com a suavização de Nadaraya-Watson utilizando a janela ótima supondo suavização linear local.

Ao todo, foram estudados quatro diferentes maneiras de estimação de modelos aditivos com janelas estimadas e verdadeiras: *backfitting* clássico com suavização linear local e com nadaraya-watson, *smooth backfitting* com suavização de nadaraya-watson e 2 estágios com suavização linear local.

As simulações foram rodadas software R-2.9.0 utilizando um computador com um processador Intel Core 2 Duo e o número de replicações foi de 1000. Devido a maior demanda computacional, estimativas que envolveram o estimador *smooth backfitting* foram replicadas 100 vezes. Além disto, no Anexo I, são apresentados diversos gráficos de estimação de densidade (estimadas através de Sheater & Jones, 1991) do logaritmo da razão entre as janelas verdadeiras e estimadas para as diferentes componentes. Densidades que estão localizadas a direita de zero indicam uma subsuavização da janela estimada. Em contrapartida, densidades a esquerda de zero indicam uma sobresuavização da mesma. Estimativas com diferentes níveis de correlação são plotados no mesmo gráfico para verificar o impacto nas janelas.

A tabela 7.1 que apresenta a demanda computacional média é apresentada abaixo:

Tempo Médio (em segundos) da Demanda Computacional por Estimador												
n	$\rho=0$				$\rho=0,25$				$\rho=0,75$			
	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E
100	0.081	0.061	2.442	0.218	0.082	0.063	2.422	0.208	0.089	0.066	2.439	0.216
250	0.250	0.239	3.248	1.334	0.250	0.239	3.258	1.327	0.264	0.244	3.302	1.332
400	0.522	0.518	4.089	5.397	0.523	0.518	4.224	5.430	0.533	0.522	4.244	5.440

Tabela 7.1: demanda computacional por estimador

A demanda computacional varia consideravelmente entre os estimadores. Note que o estimador *smooth backfitting* possui uma demanda maior do que o *backfitting* clássico devido ao termo extra na atualização da equação (4.4.3). Notamos também que para todos os estimadores o tempo de demanda computacional aumenta à medida que o tamanho de amostra cresce. Verificamos também que esta característica é salientada sob a perspectiva do estimador de 2 estágios devido a estimação preliminar do estimador de Integração Marginal (IM) que possui um número maior de operações do que os demais estimadores. A demanda computacional deste estimador para estimar os componentes aditivos em um modelo bivariado usando regressão linear local é maior devido a necessidade de calcular  $n^2$  regressões suavizadoras nos pontos  $(x_{1i}, x_{2i})$ ,  $i, j=1, \dots, n$ , as quais requerem operações da ordem  $O(nh_n)$  quando o *kernel* possui suporte compacto. Sendo assim  $\{\hat{\gamma}_1(X_{1i})\}_{i=1}^n$  demanda  $O(n^3h_n)$  operações para computar (Kim, Linton & Hengartner, 1999).

A tabela 7.2 que apresenta os valores dos MASE das simulações utilizando janelas verdadeiras segue:

Média dos Erros Quadráticos Médios usando Janelas Verdadeiras (MASE's)												
n	$\rho = 0$				$\rho = 0,25$				$\rho = 0,75$			
	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E
100	0.077	0.059	0.116	0.233	0.074	0.057	0.110	0.180	0.077	0.056	0.100	0.244
250	0.038	0.031	0.088	0.182	0.036	0.029	0.085	0.125	0.034	0.027	0.077	0.216
400	0.025	0.023	0.080	0.121	0.024	0.022	0.074	0.101	0.025	0.021	0.071	0.220

Tabela 7.2: média dos erros quadráticos médios por estimador, usando janelas verdadeiras

Ao analisar a tabela 7.2, como era esperado, notamos uma redução dos valores dos MASE ao longo de todos os cenários e todos os estimadores à medida que o tamanho de amostra aumenta, de um modo geral. Além disto, verificamos que os menores valores resultantes, indicando um melhor desempenho na estimação dentro do cenário construído, foi o caso da estimação utilizando *backfitting* clássico com suavizador de Nadaraya-Watson independente do tamanho da amostra e do nível de correlação. Este resultado foi um tanto quanto surpreendente, tanto devido ao processo de estimação da função de regressão univariada de Nadaraya-Watson ter propriedades inferiores as da regressão linear local, quanto para critério utilizado para a escolha do parâmetro de suavização, que foi o processo de minimização do AMASE supondo suavização linear local.

O método de estimação do modelo proposto utilizando *backfitting* clássico com suavização linear local e respectiva janela minimizadora do AMASE foi o método de estimação que obteve o segundo melhor desempenho. Independente do tamanho de amostra e de nível de correlação, o *backfitting* clássico se mostrou superior tanto ao *smooth backfitting*, quanto ao estimador proposto por Kim et.al. no Monte Carlo realizado.

Em terceiro lugar, obtivemos o estimador proposto por Mammen com suavizador de Nadaraya-Watson e parâmetro de suavização ótimo supondo suavização linear local.

O estimador proposto por Kim que busca resolver os problemas associados ao estimador de IM foi o que obteve os valores do MASE maiores na simulação realizada. Em todos os níveis de correlação o desempenho foi inferior em comparação com os outros estimadores sob estudo. Além disto, sua performance pareceu ter sido a que mais foi afetada entre as diferentes correlações entre os regressores.

Os efeitos da variação da correlação entre os estimadores são bem diferentes. Uma característica que não era esperada foi a de que as estimativas do MASE, quando

estabelecido um nível de correlação “baixa” ( $\rho = 0,25$ ), diminuíram em todos os estimadores e no caso do estimador de 2 estágios, esse valor diminuiu, em magnitude, mais ainda quando comparado com os outros. No entanto, com o aumento da correlação de 0,25 para 0,75, os valores do MASE dos estimadores CBF e SBF parecem não serem afetados e não mudam significativamente. Em contrapartida, o estimador de 2 estágios é altamente afetado pela correlação e apresenta valores de 0,244, 0,216 e 0,220 para  $n = 100, 250$  e  $400$ , respectivamente. Silva (2001) mostrou, sob específicas condições, que a performance do estimador de IM é inferior ao *backfitting* clássico. Aparentemente, a proposta de Kim ainda não consegue remediar totalmente o problema da instabilidade do desempenho do IM entre diferentes correlações.

Foi estudado também o desempenho destes estimadores utilizando as janelas estimadas. A tabela 7.3 resume os resultados obtidos:

Média dos Erros Quadráticos Médios usando Janelas Estimadas (MASE's)												
n	$\rho=0$				$\rho=0,25$				$\rho=0,75$			
	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E	CBFLL	CBFNW	SBFNW	2E
100	0.116	0.095	0.175	0.311	0.109	0.088	0.160	0.196	0.108	0.083	0.144	0.251
250	0.050	0.042	0.107	0.192	0.047	0.039	0.101	0.133	0.044	0.037	0.091	0.218
400	0.033	0.029	0.091	0.124	0.031	0.028	0.083	0.101	0.033	0.028	0.084	0.221

Tabela 7.2: média dos erros quadráticos médios por estimador, usando janelas estimadas

Ao analisar a tabela 7.3, notamos que as conclusões obtidas para os valores do MASE estimados usando as janelas verdadeiras são análogas utilizando as estimadas. Verificamos, no entanto, que os valores obtidos das médias dos erros quadráticos médios são maiores do que os valores obtidos usando as janelas verdadeiras. Característica que já era esperada, pois neste caso não estamos mais utilizando os valores verdadeiros, obtidos pelo processo gerador dos dados, nas expressões analíticas ótimas de minimização do AMASE da escolha do parâmetro de suavização, mas sim os estimando a partir dos dados simulados.

Cabe salientar que Martins & Yang (2007) propõem um método de 2 estágios alternativo uma vez que o suavizador de *Nadaraya-Watson internalizado* não produz um vetor equivalente de kernel que soma um. O estimador proposto substitui os valores  $y_j$  das equações (4.5.5) e (4.5.6) pelos valores  $(y_j - \bar{y})$ . Aparentemente, esta modificação melhora o desempenho do estimador de 2 estágios. Para maiores detalhes, ver Martins & Yang (2007).

A fim de ilustrar melhor o desempenho dos estimadores estudados, foram gerados gráficos de box-plots dos erros quadráticos médios das iterações de cada estimador nos

diferentes cenários construídos. As conclusões descritas na página anterior são mais fáceis de serem visualizadas. Os estimadores, da esquerda para a direita, são: CBF com suavização linear local com janela estimada (vermelho) e verdadeira (verde), CBF com suavização de Nadaraya-Watson com janela estimada (azul forte) e verdadeira (azul fraco), SBF com janela estimada (violeta) e verdadeira (amarelo), 2E com janela estimada (cinza) e verdadeira (laranja). Os gráficos seguem:

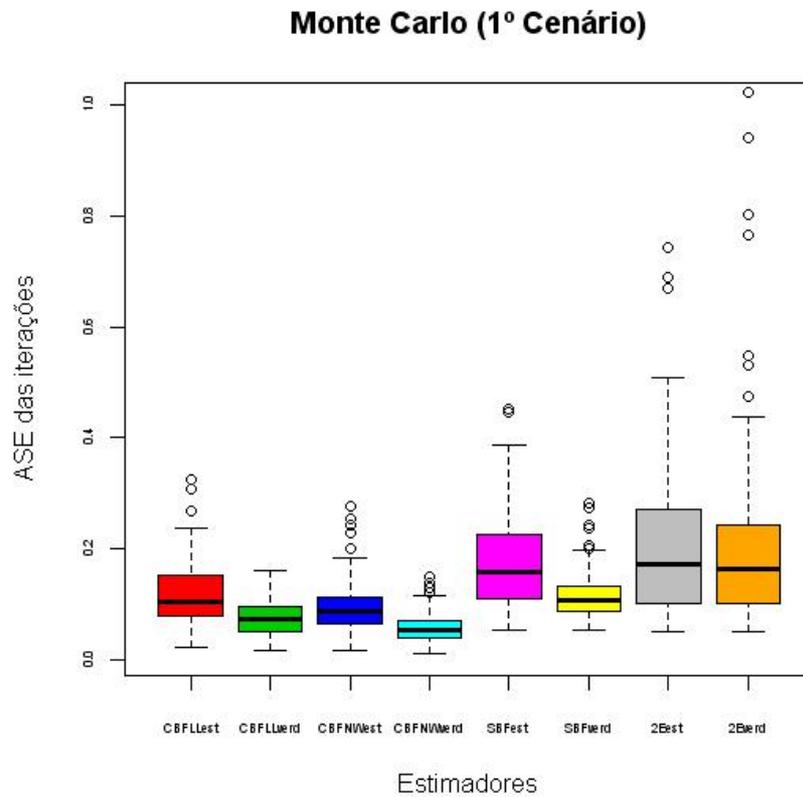


Figura 7.1: box-plot dos erros quadráticos médios das iterações do primeiro cenário

### Monte Carlo (2º Cenário)

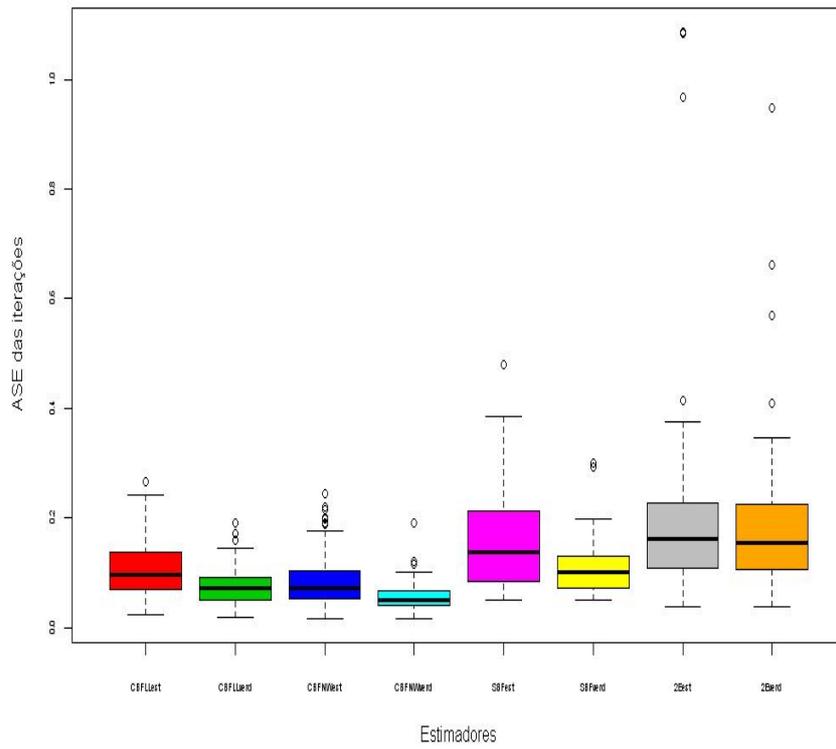


Figura 7.2: box-plot dos erros quadráticos médios das iterações do segundo cenário

### Monte Carlo (3º Cenário)

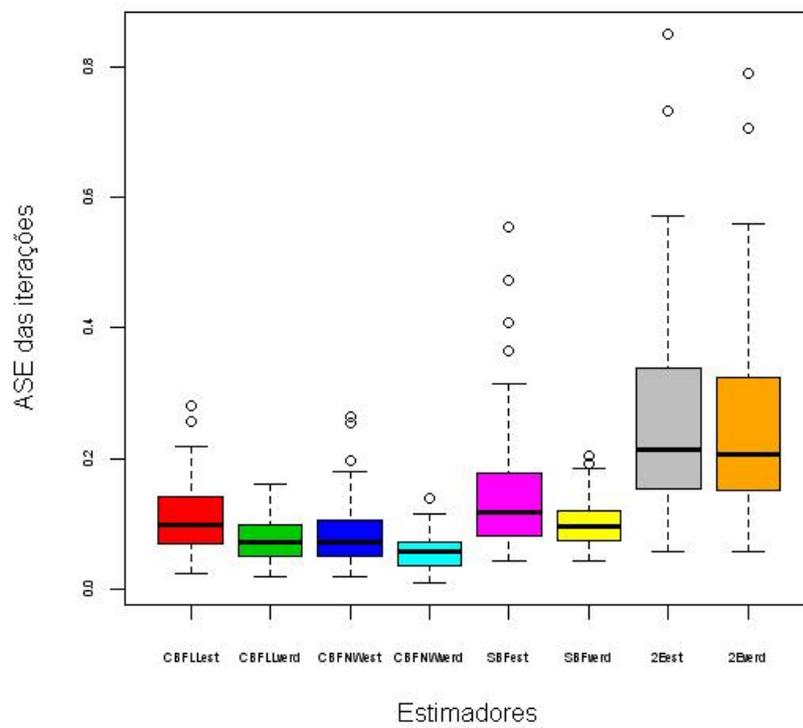


Figura 7.3: box-plot dos erros quadráticos médios das iterações do terceiro cenário

### Monte Carlo (4º Cenário)

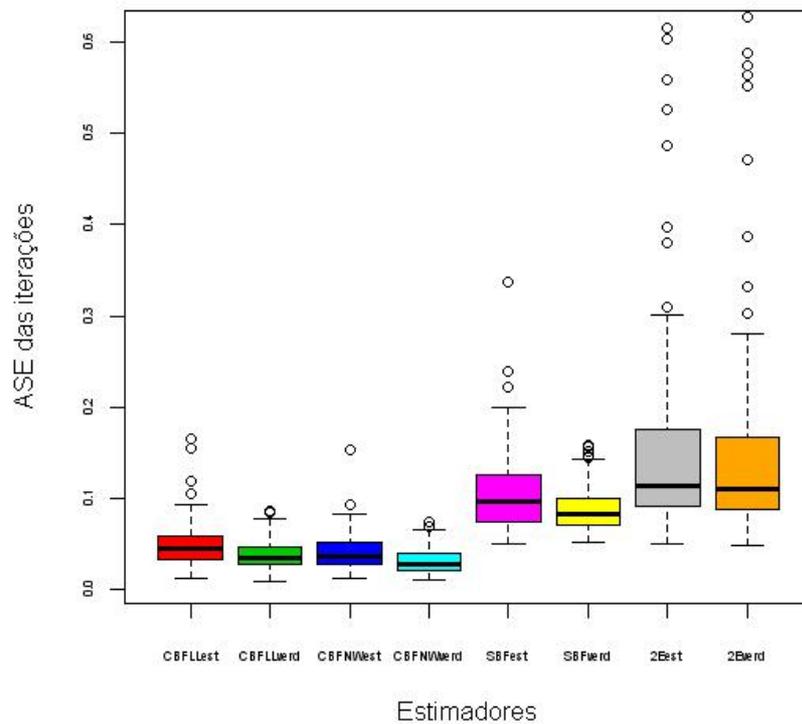


Figura 7.4: box-plot dos erros quadráticos médios das iterações do quarto cenário

### Monte Carlo (5º Cenário)

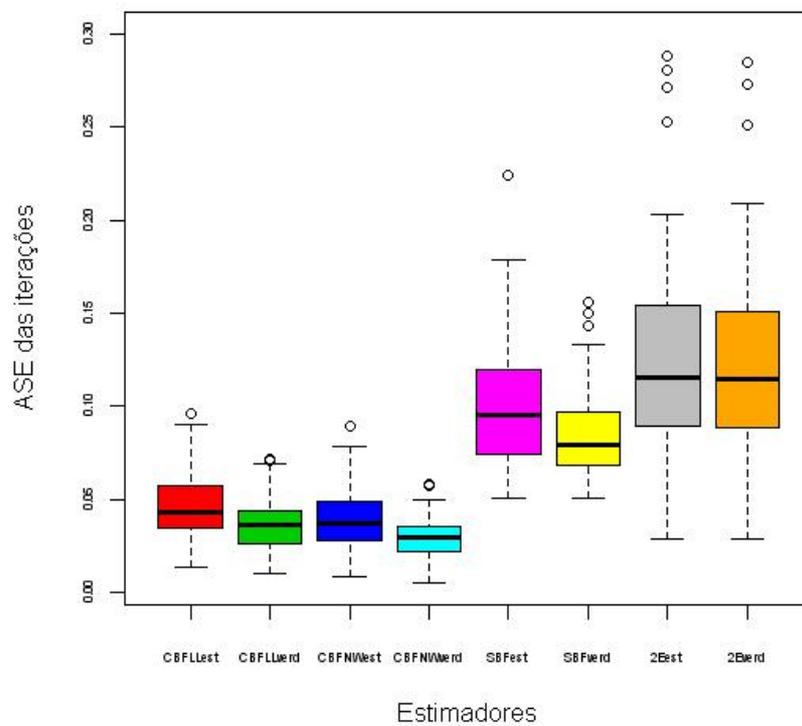


Figura 7.5: box-plot dos erros quadráticos médios das iterações do quinto cenário

### Monte Carlo (6º Cenário)

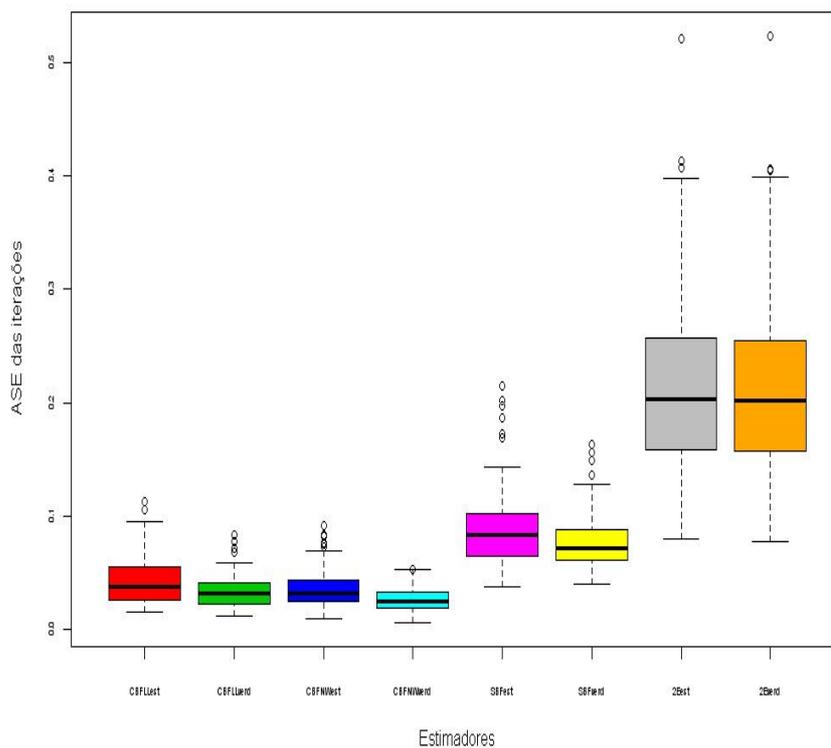


Figura 7.6: box-plot dos erros quadráticos médios das iterações do sexto cenário

### Monte Carlo (7º Cenário)

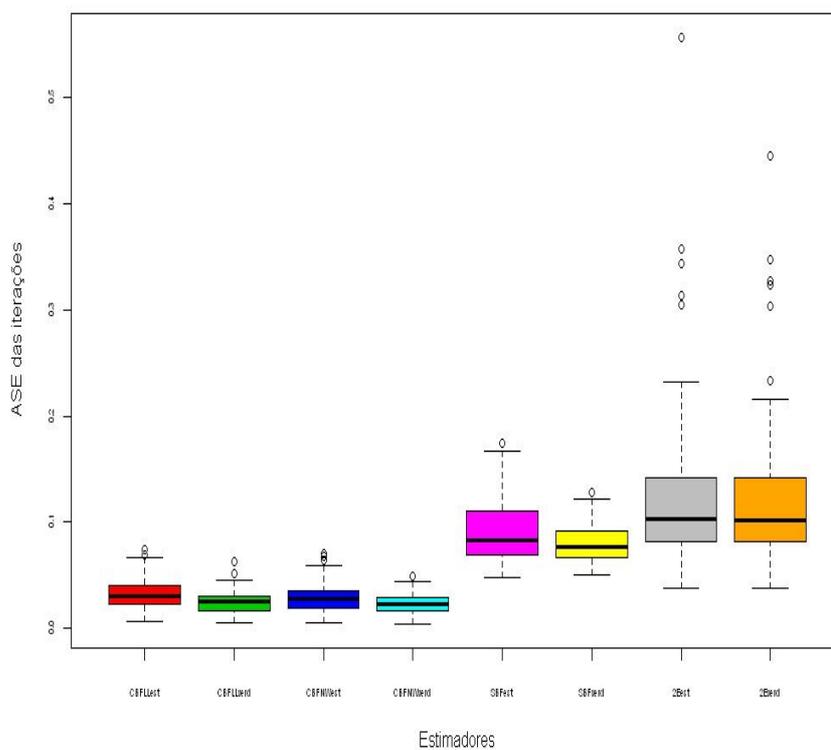


Figura 7.7: box-plot dos erros quadráticos médios das iterações do sétimo cenário

### Monte Carlo (8º Cenário)

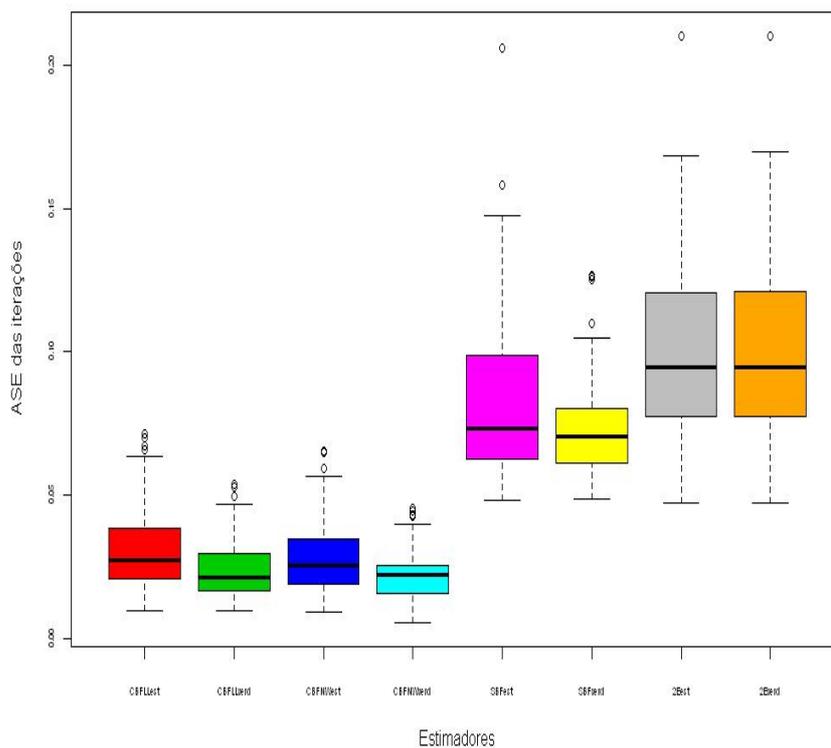


Figura 7.8: box-plot dos erros quadráticos médios das iterações do oitavo cenário

### Monte Carlo (9º Cenário)

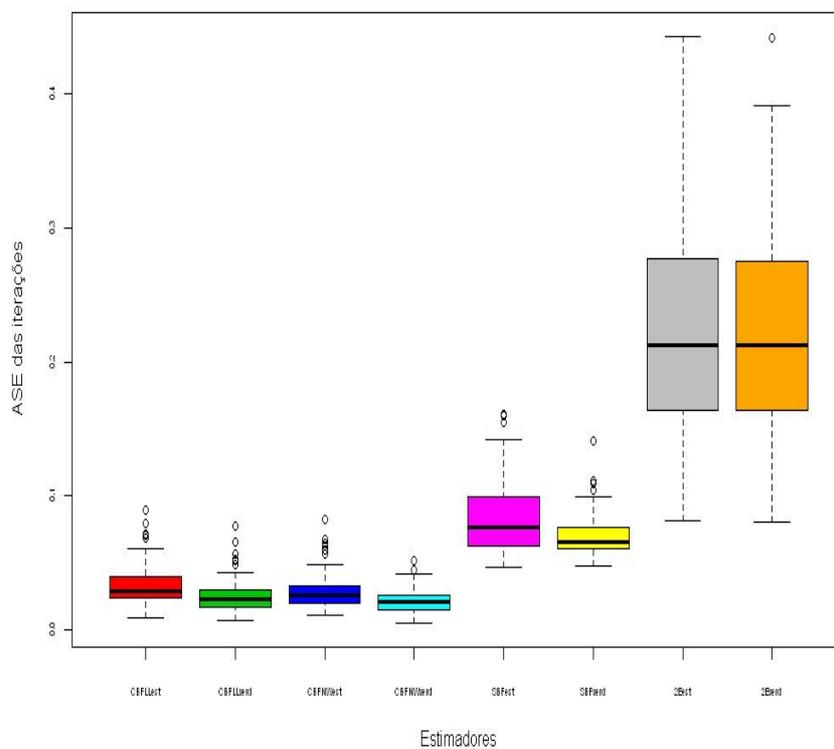


Figura 7.9: box-plot dos erros quadráticos médios das iterações do nono cenário

Analisando as figura anteriores, constatamos o comportamento já comentado anteriormente. Os valores dos erros quadráticos médios dos estimadores SBF e 2E (as quatro caixas à direita) são maiores do que os do estimador CBF (quatro caixas à esquerda). Além disto, é nítido o comportamento mais variado do estimador de 2 estágios, quando comparado com os demais, devido ao tamanho maior da caixa e a presença de diversos outliers.

O objetivo da monografia era comparar três métodos atuais para a estimação de uma regressão não-paramétrica aditiva. Dentre os principais achados, podemos enumerar:

1) A quebra de suposição de independência entre os regressores não afeta estimação feita via algoritmo de *backfitting* e *smooth backfitting*. O mesmo não ocorre quando se utiliza o estimador de 2 estágios no modelo proposto.

2) Uma diferença interessante entre duas maneiras de estimação usando o algoritmo *backfitting* foi constatada. Aparentemente, a estimação do modelo aditivo usando *backfitting* clássico com suavização de Nadaraya-Watson é superior do que com a suavização linear local nos cenários construídos. Mesmo não utilizando a janela ótima para ambas as estimações.

3) O estimador *smooth backfitting*, apesar de ter propriedades teóricas mais desejáveis do que o *backfitting* clássico, apresentou desempenho inferior a este último dentro das especificações do estudo realizado.

4) O estimador *smooth backfitting* apresenta uma demanda computacional mais elevada do que a do clássico devido a inclusão de um termo extra no processo iterativo. Além disto, a demanda computacional do estimador de 2 estágios aumenta cubicamente à medida que o tamanho da amostra cresce.

# Referências Bibliográficas

- Barbian, M. H., 2006, Regressão Semi-Paramétrica Utilizando Modelos de Índice Único, Monografia de Bacharelado em Estatística.
- Buja, A., Hastie, T. J. And Tibshirani, R. J., 1989, Linear smoothers and additive models. *The Annals of Statistics*, 17, 453-555.
- Cacoullos, T., 1996, Estimation of a Multivariate Density, *Annals of the Institute of Statistical Mathematics*, 18, 178-189.
- Cleveland, W., 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of American Statistical Association*, 83, 596-610.
- Conceição, G. M., Saldiva, P. H., Singer, J. M., 2001, Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo, *Rev. Bras. Epidemiol.*, Vol. 4, 203-219.
- Conceição, G. M. et. al., 2001, Air pollution and child mortality: a time series study in São Paulo, Brazil. *Environmental Health Perspectives*, Vol. 109, Sup. 3, 347-350.
- Efromovich, S., 1999, *Nonparametric Curve Estimation: Methods, Theory and Applications*, New York: Springer-Verlag.
- Epanechnikov, V. A., 1969, Nonparametric Estimates of a Multivariate Probability Density, *Theory of Probability and Applications*, 14, 153-158.
- Fan, J., 1992, Design-adaptive nonparametric regression, *Journal of the American Statistics Association*, 87, 998-1004.
- Fan, J., Gasser, T., Gijbels, I., Brockmann, M., Engel, J., 1993, Local Polynomial Fitting: a Standard for Nonparametric Regression, *Departament of Statistics*, UNC.
- Friedman, J. H. and Stuetzle, W., 1981, Projection Pursuit Regression, *Journal of the American Statistical Association*, 76, 817-823.
- Green, P. J. e Silverman, B. W., 1994, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.
- Greene, W. H., 1999, *Econometric Analysis*, 4ª Edição, New York: Prentice-Hall
- Gujarati, D., 2000, *Econometria Básica*, Makron Books.
- Hastie, T.J. and R.J. Tibshirani, 1986, Generalized additive models, *Statistical Science*, 1, 297-318.
- Hastie, T. J. and Tibshirani, R. J., 1990, *Generalized Additive Models*. London: Chapman and Hall.
- Hastie, T., Tibshirani, R., Friedman, J., 2001, *The elements of Statistical Learning: Data Mining, Inference and prediction*, Springer Series in Statistics.

- Ibragimov, I. A. and Hasminskii, R. Z., 1983, Estimation of Distribution Density, *J. Sov. Math.*, 25, 40-57.
- Johnson, J. and Dinardo, J., 1996, *Econometric Methods*, 4ª Edição, Irvine: McGraw-Hill
- Kim, W., Linton, O. B. and Hengartner, N. W., 1999, A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals, *Journal of Computational and Graphical Statistics*, 9, 2, 279-297.
- Linton, O. and Nielsen, J. P., 1995, A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*, 82, 1, 93-100.
- Linton, O. and W. Härdle, 1996, Estimation of additive regression models with known links, *Biometrika*, 83, 529-540.
- Linton, O., Nielsen, J. P. And Van de Geer, S., 2000, Estimating Multiplicative and Additive Marker Dependent Hazard Functions by Backfitting with the assistance of Marginal Integration, *Working Paper*, LSE, Department of Economics.
- Mammen, E., Linton, O. and Nielsen J., 1999, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *The Annals of Statistics*, 27, 1443-1490.
- Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P., 2001, A general projection framework for constrained smoothing, *Statistical Science*, 16, No. 3, 232-248.
- Martins-Filho, Carlos and Yang, Ke, 2007, Finite sample performance of kernel-based regression methods for nonparametric additive models under common bandwidth selection criterion, *Journal of Nonparametric Statistics*, 19, 23-62.
- Montgomery, D. and Peck, E, 1982, *Introduction to Linear Regression Analysis*. John Wiley & Sons.
- Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W., 1996, *Applied Linear Statistical Models*, 4ª Edição, MCGraw-Hill.
- Nielsen, J. P. and Sperlich, S., 2005, Smooth backfitting in practice, *Journal of Royal Statistical Society, Series B*, 67, 43-61
- Opsomer, J. and D. Ruppert, 1997, Fitting a bivariate additive model by local polynomial regression, *Annals of Statistics*, 25, 186-211.
- Opsomer, J. and D. Ruppert, 1998, A fully automated bandwidth selection method for fitting additive models, *Journal of the American Statistical Association*, 93, 605-619.
- Pagan, A. and A. Ullah, 1999, *Nonparametric econometrics*. Cambridge: Cambridge University Press.
- Ruppert, D., S.J. Sheather, and M.P. Wand, 1995, An effective bandwidth selector for least squares regression, *Journal of the American Statistical Association*, 90, 1257-1270.
- Scott, D. W., 1992, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.

Sheather, S.J. and M.C. Jones, 1991, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society B*, 53, 683-690.

Silva, F. A. B., 2001, Estimação de Regressões Aditivas via Backfitting e Integração Marginal: Performance em Amostras Finitas, Dissertação de Mestrado, Escola de Pós-Graduação em Economia, Fundação Getúlio Vargas.

Silverman, B.W., 1986, Density estimation for statistics and data analysis, Chapman and Hall: New York.

Simões, P., 2007, Modelos Aditivos Semi-Paramétricos: Aplicações em Séries Temporais, Monografia de Bacharelado em Estatística

Simonoff, J. S., 1996, Smoothing Methods in Statistics. New York: Springer-Verlag.

Stone, C., 1980, Optimal rates of convergence for nonparametric estimators, *Annals of Statistics*, 8, 1348-1360.

Stone, C. J., 1982, Optimal Global Rates of Convergence for Nonparametric Regression, *Annals of Statistics*, 10, 1040-1053.

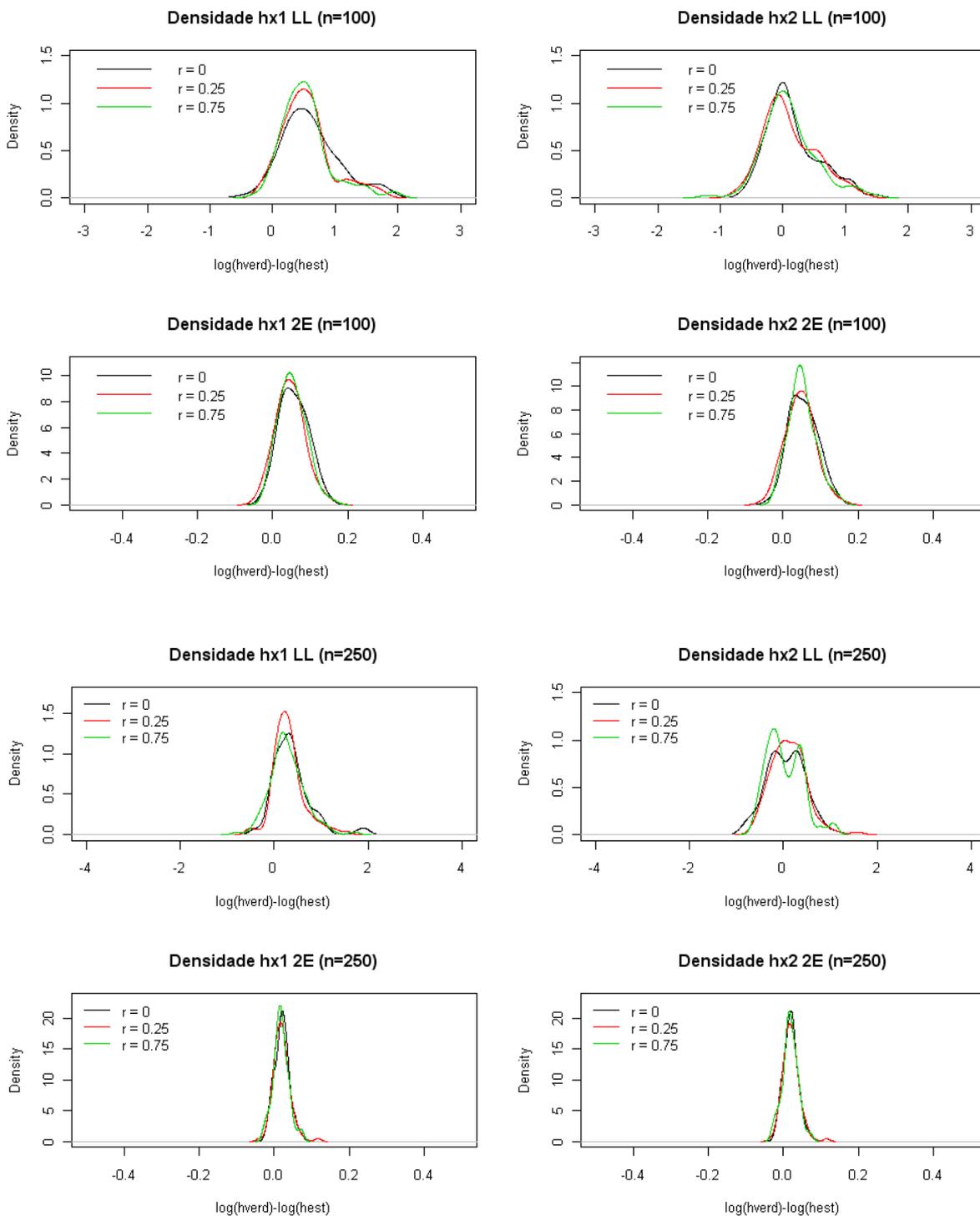
Stone, C., 1985, Additive regression and other nonparametric models, *Annals of Statistics*, 6, 689-705.

Wand, M. P. e Jones, M. C., 1995, Kernel Smoothing. London: Chapman and Hall.

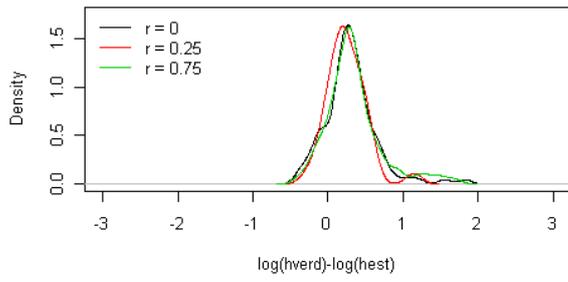
Wang, J. e Yang, L., 2007, Efficient and fast spline-backfitted kernel smoothing of additive models, *Annals of the Institute of Statistical Mathematics*, 61, 663-690

# Anexo I

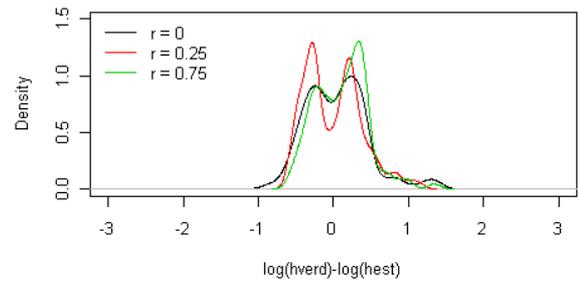
Gráficos de densidade dos logaritmos da razão entre janelas verdadeiras e estimadas utilizando o parâmetro de suavização de Sheater & Jones (1991):



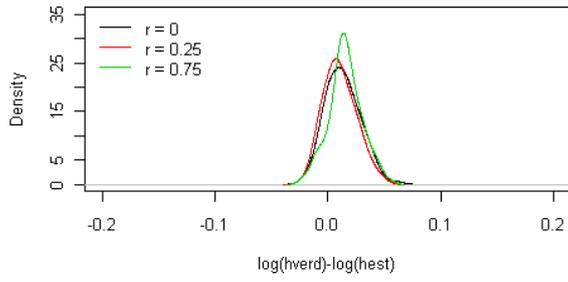
Densidade hx1 LL (n=400)



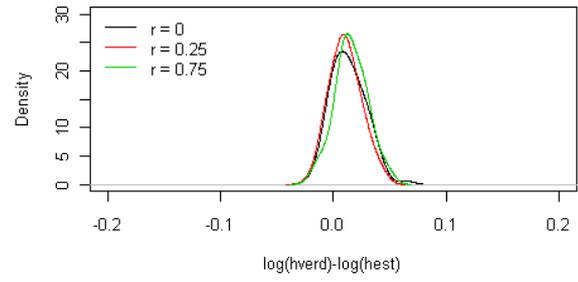
Densidade hx2 LL (n=400)



Densidade hx1 2E (n=400)



Densidade hx2 2E (n=400)



# Anexo II

## Funções utilizadas no R em ordem de aparição:

```
# Exemplo escolha de h no histograma
x <- rnorm(300, 10, 3)
x1 <- seq(0, 25, by=0.01)
x2 <- dnorm(x1, mean=10, sd=3)

par(mfrow=c(1,2))
hist(x, nclass=5, freq=FALSE, xlab="Sobresuavizado", main="", ylab="Densidade",
ylim=c(0,0.15))
lines(x1, x2, type="l", col="red", lwd=2)

hist(x, nclass=100, freq=FALSE, xlab="Subsuavizado", main="", ylab="Densidade")
lines(x1, x2, type="l", col="red", lwd=2)

# Variando o bandwidth na estimação da densidade
library(KernSmooth)
n <- 50
dp <- 2
e <- rnorm(n, 0, sd=dp)

par(mfrow=c(3,1))
plot(density(e, bw=0.2), main="(A)")
plot(density(e, bw=0.5), main="(B)")
plot(density(e, bw=0.9), main="(C)")

# Comparando bandwidths e grau do polinômio
library(KernSmooth)
n <- 100
dp <- 6
e <- rnorm(n, 0, sd=dp)
x <- seq(-3, 3, l=n)
y <- 4 + 1.32*x - 3*x^2 + e
par(mfrow=c(3,2))
plot(x, y, type="p", cex=0.6, xlab="Linear Local, h=0.15")
lines(locpoly(x, y, degree=1, band=0.15)) # Linear Local, h=0.15

plot(x, y, type="p", cex=0.6, xlab="Quadrática Local, h=0.15")
lines(locpoly(x, y, degree=2, band=0.15)) # Quadrática Local, h=0.15

plot(x, y, type="p", cex=0.6, xlab="Linear Local, h=0.4")
lines(locpoly(x, y, degree=1, band=0.4)) # Linear Local, h=0.4

plot(x, y, type="p", cex=0.6, xlab="Quadrática Local, h=0.4")
lines(locpoly(x, y, degree=2, band=0.4)) # Quadrática Local, h=0.4

plot(x, y, type="p", cex=0.6, xlab="Linear Local, h=100")
lines(locpoly(x, y, degree=1, band=100)) # Linear Local, h=100

plot(x, y, type="p", cex=0.6, xlab="Quadrática Local, h=100")
lines(locpoly(x, y, degree=2, band=100)) # Quadrática Local, h=100

# Comparando bandwidth grande e caso paramétrico
library(KernSmooth)
n <- 100
dp <- 6
e <- rnorm(n, 0, sd=dp)
x <- seq(-3, 3, l=n)
y <- 4 + 1.32*x - 3*x^2 + e
par(mfrow=c(2,2))
plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Linear Paramétrica", cex=1))
model_linear <- lm(y ~ x)
lines(x, predict(model_linear), col="red") # Linear Paramétrico

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Linear Local
Não-Paramétrica com h grande", cex=1))
lines(locpoly(x, y, degree=1, band=100), col="blue") # Linear Não-Paramétrico Local, h=100

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Quadrática Paramétrica", cex=1))
model_quadratico <- lm(y ~ x + I(x^2))
lines(x, predict(model_quadratico), col="red") # Quadrático Paramétrico
```

```

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Quadrática Local
Não-Paramétrica com h grande", cex=1))
lines(locpoly(x, y, degree=2, band=100), col="blue") # Quadrático Não-Paramétrico Local,
h=100

# Comparando Linear Local com Nadaraya-Watson
library(KernSmooth)
n <- 100
dp <- 6
e <- rnorm(n, 0, sd=dp)
x <- seq(-3, 3, l=n)
y <- 4 + 1.32*x - 3*x^2 + e
par(mfrow=c(2,2))
plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Linear Local, h=0.5", cex=1))
lines(locpoly(x, y, degree=1, band=0.5), col="blue") # Linear Local, h=0.5

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Nadaraya-Watson, h=0.5", cex=1))
lines(locpoly(x, y, degree=0, band=0.5), col="blue") # Nadaraya-Watson, h=0.5

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Regressão Linear Local, h=0.7", cex=1))
lines(locpoly(x, y, degree=1, band=0.7), col="blue") # Linear Local, h=0.7

plot(x, y, type="p", cex=0.6, xlab="x")
title(main = list("Nadaraya-Watson, h=0.7", cex=1))
lines(locpoly(x, y, degree=0, band=0.7), col="blue") # Nadaraya-Watson, h=0.7

# Efeito da adição de covariável na dimensionalidade irrestrita (gráfico repetitivo)
library(scatterplot3d)
x <- seq(-3, 3, l=50)
y <- x + rnorm(50, 0, sd=2)
u <- rnorm(50, 0, sd=6)
z <- x + y + u
par(mfrow=c(2,1))
repeat {
for (i in 1:360) {
scatterplot3d(x,y,z, cex.axis=0.2, cex.lab=0.2, mar=rep(0,4), type="p", xlim=c(-3,3),
ylim=c(-15,15), zlim=c(-25,25), angle=i)
scatterplot3d(x,y, cex.axis=0.2, cex.lab=0.2, mar=rep(0,4), type="p", angle=i)
}
}

# Gráfico das funções do processo de geração dos dados
m1 <- function(x) { (11/3)*x - 8*x^2 + (16/3)*x^3 } # Função Geradora do Primeiro
Componente
m2 <- function(x) { (1/4)*sin(4*pi*x) }
x <- seq(0,1,l=200)
y1 <- m1(x)
y2 <- m2(x)
plot(x, y1, type="l", ylim=c(-0.3, 1), ylab="", xlab="")
lines(x, y2, lty=2)
legend("topleft", c("m1(x)", "m2(x)"), lty=1:2, bty="n")

# Gráfico da superfície do processo gerador dos dados
x <- seq(0,1,l=30)
y <- seq(0,1,l=30)
m1 <- function(x) { (11/3)*x - 8*x^2 + (16/3)*x^3 }
m2 <- function(x) { (1/4)*sin(4*pi*x) }
k <- function(x,y) { m1(x) + m2(y) }
z <- outer(x,y, k)
persp(z, theta=45, phi=45, col=7, main="y = m1(x1) + m2(x2)")

# Código das simulações de um cenário como exemplo
library(MASS) # Função 'mvrnorm' Geradora de Normal Bivariada
library(KernSmooth) # Função 'locpoly' para estimação da Reg. Polinomial Local
library(sBF) # Função 'sBF' para estimação do Smooth Backfitting
library(nleqslv) # Função 'nleqslv' para resolução de sistemas não-lineares para h's do 2E

```

```

#####
##### PRIMEIRO CENÁRIO #####
#####

```

```

#####
# Funções Auxiliares #

llr <- function(x,y,h) {
n <- length(y)
mtle <- rep(0, n)
for (k in 1:n) {
s0tle <- (n^-1)*(sum((((x-x[k])^0)*(((2*pi)^-.5)*exp(-(((x-x[k])/h)^2)/2)/h))))
s1tle <- (n^-1)*(sum((((x-x[k])^1)*(((2*pi)^-.5)*exp(-(((x-x[k])/h)^2)/2)/h))))
s2tle <- (n^-1)*(sum((((x-x[k])^2)*(((2*pi)^-.5)*exp(-(((x-x[k])/h)^2)/2)/h))))
mtle[k] <- (n^-1)*(sum(((s2tle-s1tle*(x-x[k]))*(((2*pi)^-.5)*exp(-(((x-x[k])/h)^2)/2)/h)*y)/(s2tle*s0tle-(s1tle^2))))
}
mtle
}

hybr <- function(y,x1,x2,hx1,hx2,g1,g2) { # Função da estimação de 2 estágios

n <- length(y)
e <- c(1,0)
s1 <- matrix(0,n,n)
s2 <- matrix(0,n,n)

y <- y # Estimador 2E1

aux_x1 <- matrix(rep(x1,n), n)
direg1 <- aux_x1 - t(aux_x1)
aux_x2 <- matrix(rep(x2,n), n)
direg2 <- aux_x1 - t(aux_x2)

a1 <- (1/hx1) * direg1
a2 <- (1/hx2) * direg2
nk1 <- 1/(sqrt(2*pi)) * exp(-0.5*a1^2)
nk2 <- 1/(sqrt(2*pi)) * exp(-0.5*a2^2)
gal <- (1/n) * nk1 %*% (y * (nk2 %*% rep(1,n)) / ((nk1*nk2) %*% rep(1,n)))
ga2 <- (1/n) * nk2 %*% (y * (nk1 %*% rep(1,n)) / ((nk2*nk1) %*% rep(1,n)))

#####
m1 <- (nk1 %*% (y-ga2)) / (nk1 %*% rep(1,n)) # Primeiros Componente do IM de x1
m2 <- (nk2 %*% (y-gal)) / (nk2 %*% rep(1,n)) # Primeiros Componente do IM de x2
#####

yle <- y - ga2
y2e <- y - gal

ag1 <- (1/g1) * direg1
ag2 <- (1/g2) * direg2
kn1 <- 1/(sqrt(2*pi)) * exp(-0.5*ag1^2)
kn2 <- 1/(sqrt(2*pi)) * exp(-0.5*ag2^2)
aux1 <- matrix(0,n,2)
aux2 <- matrix(0,n,2)
j=1
while (j <= n) {
aux1 <- cbind(rep(1,n), direg1[,j])
aux2 <- cbind(rep(1,n), direg2[,j])
w1 <- (1/g1) * diag(kn1[,j],n,n)
w2 <- (1/g2) * diag(kn2[,j],n,n)
s1[j,] <- t(e) %*% solve(t(aux1) %*% w1 %*% aux1) %*% t(aux1) %*% w1
s2[j,] <- t(e) %*% solve(t(aux2) %*% w2 %*% aux2) %*% t(aux2) %*% w2
j <- j + 1
}

m1 <- s1 %*% yle
m2 <- s2 %*% y2e

return(list(m1=m1, m2=m2))

}

# Funções do DGP #

m1 <- function(x) { (11/3)*x - 8*x^2 + (16/3)*x^3 } # Função Geradora do Primeiro
Componente
m2 <- function(x) { (1/4)*sin(4*pi*x) } # Função Geradora do Segundo
Componente
d1m1 <- function(x) { (11/3) - 16*x + 16*x^2 } # Derivada do Primeiro Componente
d1m2 <- function(x) { (pi)*cos(4*pi*x) } # Derivada do Segundo Componente
d2m1 <- function(x) { -16 + 32*x } # Derivada 2ª do Primeiro Componente
d2m2 <- function(x) { 4*(pi^2)*-sin(4*pi*x) } # Derivada 2ª do Segundo Componente

```

```

### ESPECIFICAÇÕES DO CENÁRIO ###

r <- 1000 # Número de iterações
n <- 100 # Tamanho da amostra
c <- 0 # Correlação
medias <- c(0.5, 0.5) # Médias de x1 e x2
sigma <- matrix(c(1/9, c/9, c/9, 1/9), 2, 2) # Matriz de Covariâncias

### Objetos Auxiliares ###

ase_CBFLLest <- rep(0,r)
ase_CBFLLverd <- rep(0,r)
ase_CBFNWest <- rep(0,r)
ase_CBFNWverd <- rep(0,r)
ase_SBFest <- rep(0,r)
ase_SBFverd <- rep(0,r)
ase_2Eest <- rep(0,r)
ase_2Everd <- rep(0,r)

tempo_CBFLLest <- rep(0,r)
tempo_CBFLLverd <- rep(0,r)
tempo_CBFNWest <- rep(0,r)
tempo_CBFNWverd <- rep(0,r)
tempo_SBFest <- rep(0,r)
tempo_SBFverd <- rep(0,r)
tempo_2Eest <- rep(0,r)
tempo_2Everd <- rep(0,r)

janelas_ll_hx1 <- rep(0,r)
janelas_ll_hx2 <- rep(0,r)
janelas_TRUE_ll_hx1 <- rep(0,r)
janelas_TRUE_ll_hx2 <- rep(0,r)
janelas_2E_hx1 <- rep(0,r)
janelas_2E_hx2 <- rep(0,r)
janelas_TRUE_2E_hx1 <- rep(0,r)
janelas_TRUE_2E_hx2 <- rep(0,r)

valores_validos <- matrix(0,1,2) # Auxiliar para validar o loop de repetição, se
necessário.

### INÍCIO DAS SIMULAÇÕES ###
for (i in 1:r) { # Início da simulação

while (length(valores_validos[,1]) < n) {
simu <- mvrnorm(n = 3*n, medias, sigma) # Estabelecendo um valor confiável (3n) de que
teremos, no mínimo, n valores válidos. Caso contrário, o proceso repete.
aux <- rep(0, (3*n))
for (j in 1:(3*n)) {
aux[j] <- ifelse(0>=simu[j,1] || simu[j,1]>=1 || 0>=simu[j,2] || simu[j,2]>=1, 1,0)
}
valores_validos <- simu[which(aux==0),]
}

amostra <- valores_validos[1:n,] # Pega somente os valores dentro do suporte [0,1]

x1 <- amostra[,1]
x2 <- amostra[,2]
erros <- rnorm(n)
y <- m1(x1) + m2(x2) + erros

### BACKFITTING CLÁSSICO LOCAL LINEAR h EST ###

inicio <- Sys.time()

yb <- mean(y) # y barra
hx1 <- dpill(x1, y)
hx2 <- dpill(x2, y)
mtC1 <- llr(x1, (y-yb), hx1)
mtC2 <- llr(x2, (y-yb), hx2)
mr1 <- mtC1 - mean(mtC1)
mr2 <- mtC2 - mean(mtC2)
cc1 <- 1 # Erro Inicial do componente de x1
cc2 <- 1 # Erro Inicial do componente de x2

while (cc1 >= 0.0001 | cc2 >= 0.0001) { # Início do procedimento do Backfitting
Clássico

regm1 <- locpoly(x1, (y - yb - mtC2), bandwidth = hx1)
mtC1 <- approx(regm1$x, regm1$y, x1)$y
mtC1 <- mtC1 - mean(mtC1)

```

```

regm2 <- locpoly(x2, (y - yb - mtC1), bandwidth = hx2)
mtC2 <- approx(regm2$x, regm2$y, x2)$y
mtC2 <- mtC2 - mean(mtC2)

cc1 <- sum((mtC1 - mr1)^2)/(sum(mr1^2) + 0.0001)
cc2 <- sum((mtC2 - mr2)^2)/(sum(mr2^2) + 0.0001)

mr1 <- mtC1
mr2 <- mtC2

} # Final do procedimento do Backfitting Clássico

fim <- Sys.time()

ase_CBFLLest[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)
janelas_ll_hx1[i] <- hx1
janelas_ll_hx2[i] <- hx2

tempo_CBFLLest[i] <- fim - inicio

### 2 ESTÁGIOS LOCAL LINEAR h EST ###

inicio <- Sys.time()

range1 <- max(x1) - min(x1)
range2 <- max(x2) - min(x2)
R <- 1/sqrt(2*pi)
lambda1 <- as.numeric(quantile(x1, 0.625) - quantile(x1, 0.375))
lambda2 <- as.numeric(quantile(x2, 0.625) - quantile(x2, 0.375))
silh1 <- 0.79*lambda1*n^(-.2)
silh2 <- 0.79*lambda2*n^(-.2)
aux_x1 <- matrix(rep(x1,n), n)
direg1 <- aux_x1 - t(aux_x1)
a1 <- direg1/silh1
aux_x2 <- matrix(rep(x2,n), n)
direg2 <- aux_x2 - t(aux_x2)
a2 <- direg2/silh2
k1 <- dnorm(a1)
k2 <- dnorm(a2)
d2k1 <- 1 / sqrt(2*pi) * exp((-a1^2)/2) * (a1^2 - 1)
d2k2 <- 1 / sqrt(2*pi) * exp((-a2^2)/2) * (a2^2 - 1)
f1 <- 1/(n*silh1) * apply(k1,2,sum) # Soma nas colunas
f2 <- 1/(n*silh2) * apply(k2,2,sum) # Soma nas colunas
d2f1 <- 1/(n*silh1) * apply(d2k1,2,sum)
d2f2 <- 1/(n*silh2) * apply(d2k2,2,sum)

aux <- function(x) { d2m1(x)*dnorm(x,mean=medias[1],sd=sqrt(sigma[1,1])) }
ed2m1 <- integrate(aux, 0, 1)$value
theta11 <- (1/n) * sum((d2m1(x1) - ed2m1)^2)
aux <- function(x) { d2m2(x)*dnorm(x,mean=medias[2],sd=sqrt(sigma[2,2])) }
ed2m2 <- integrate(aux, 0, 1)$value
theta22 <- (1/n) * sum((d2m2(x2) - ed2m2)^2)

qui11 <- theta11 + (mean(d2f1*mr1/f1))^2 # Usa-se o componente final do CBF
qui22 <- theta22 + (mean(d2f2*mr2/f2))^2 # Usa-se o componente final do CBF
qui12 <- (1/(n^2)) * sum(d2f1*mr1/f1) * sum(d2f2*mr2/f2)

dfm <- diag(x=1,n,n) - ((rep(1,n) %*% t(rep(1,n))) * (1/n))
res <- dfm %*% y - mr1 - mr2
sig2 <- (1/n) * t(res) %*% res
range1 <- max(x1) - min(x1)
range2 <- max(x2) - min(x2)

h01 <- (R*sig2*range1 / (n*qui11))^(.2) # Chute Inicial Regra de Bolso de hx1
h02 <- (R*sig2*range2 / (n*qui22))^(.2) # Chute Inicial Regra de Bolso de hx2

sistema <- function(x) {
y <- numeric(2)
y[1] <- x[1]^5 * qui11 + x[2]^2 * x[1]^3 * qui12 - R*range1*sig2/n
y[2] <- x[2]^5 * qui22 + x[1]^2 * x[2]^3 * qui12 - R*range2*sig2/n
y
}

jacobiano <- function(x) {
y <- numeric(4)
y <- matrix(y,2,2)
y[1,1] <- 5*x[1]^4 * qui11 + x[2]^2 * 3* x[1]^2 * qui12
y[2,2] <- 5*x[2]^4 * qui22 + x[1]^2 * 3* x[2]^2 * qui12
y[1,2] <- 2 * x[2] * x[1]^3 * qui12
y[2,1] <- 2 * x[1] * x[2]^3 * qui12
}

```

```

y
}

### Resolvendo o Sistema ###

janelas_2E_est <- nleqslv(c(h01, h02), sistema, jac=jacobiano) # Janelas estimadas do 2E
janelas_2E_hx1[i] <- janelas_2E_est$x[1] # Janela estimada do 2E de x1
janelas_2E_hx2[i] <- janelas_2E_est$x[2] # Janela estimada do 2E de x2

### Ajustando o modelo dos 2E ###

modelo <-
hybr(y,x1,x2,janelas_2E_est$x[1],janelas_2E_est$x[2],janelas_2E_est$x[1],janelas_2E_est$x[2
])
mr1 <- modelo$m1
mr2 <- modelo$m2

ase_2Eest[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

fim <- Sys.time()

tempo_2Eest[i] <- fim - inicio

### BACKFITTING CLÁSSICO NADARAYA-WATSON COM h ESTIMADO ###

inicio <- Sys.time()

yb <- mean(y) # y barra
hx1 <- dpill(x1, y)
hx2 <- dpill(x2, y)
mtC1 <- llr(x1, (y-yb), hx1)
mtC2 <- llr(x2, (y-yb), hx2)
mr1 <- mtC1 - mean(mtC1)
mr2 <- mtC2 - mean(mtC2)
cc1 <- 1 # Erro Inicial do componente de x1
cc2 <- 1 # Erro Inicial do componente de x2

while (cc1 >= 0.0001 | cc2 >= 0.0001) { # Início do procedimento do Backfitting
Clássico

regm1 <- locpoly(x1, (y - yb - mtC2), bandwidth = hx1, degree=0)
mtC1 <- approx(regm1$x, regm1$y, x1)$y
mtC1 <- mtC1 - mean(mtC1)

regm2 <- locpoly(x2, (y - yb - mtC1), bandwidth = hx2, degree=0)
mtC2 <- approx(regm2$x, regm2$y, x2)$y
mtC2 <- mtC2 - mean(mtC2)

cc1 <- sum((mtC1 - mr1)^2)/(sum(mr1^2) + 0.0001)
cc2 <- sum((mtC2 - mr2)^2)/(sum(mr2^2) + 0.0001)

mr1 <- mtC1
mr2 <- mtC2

} # Final do procedimento do Backfitting Clássico

fim <- Sys.time()

ase_CBFNWest[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

tempo_CBFNWest[i] <- fim - inicio

### BACKFITTING CLÁSSICO LOCAL LINEAR COM h VERDADEIRO ###

inicio <- Sys.time()

yb <- mean(y) # y barra
aux <- function(x) { d2m1(x)*dnorm(x,mean=medias[1],sd=sqrt(sigma[1,1])) }
ed2m1 <- integrate(aux, 0, 1)$value
theta11 <- (1/n) * sum((d2m1(x1) - ed2m1)^2)
aux <- function(x) { d2m2(x)*dnorm(x,mean=medias[2],sd=sqrt(sigma[2,2])) }
ed2m2 <- integrate(aux, 0, 1)$value
theta22 <- (1/n) * sum((d2m2(x2) - ed2m2)^2)
v1 <- ((1/(2*sqrt(pi)*n*n*theta11)) *
sum(1/dnorm(x1,mean=medias[1],sd=sqrt(sigma[1,1]))))^(.2)
v2 <- ((1/(2*sqrt(pi)*n*n*theta22)) *
sum(1/dnorm(x2,mean=medias[2],sd=sqrt(sigma[1,1]))))^(.2)
mtC1 <- llr(x1, (y-yb), v1)
mtC2 <- llr(x2, (y-yb), v2)
mr1 <- mtC1 - mean(mtC1)
mr2 <- mtC2 - mean(mtC2)

```

```

cc1 <- 1 # Erro Inicial do componente de x1
cc2 <- 1 # Erro Inicial do componente de x2

while (cc1 >= 0.0001 | cc2 >= 0.0001) {      # Início do procedimento do Backfitting
Clássico

regm1 <- locpoly(x1, (y - yb - mtC2), bandwidth = v1)
mtC1 <- approx(regm1$x, regm1$y, x1)$y
mtC1 <- mtC1 - mean(mtC1)

regm2 <- locpoly(x2, (y - yb - mtC1), bandwidth = v2)
mtC2 <- approx(regm2$x, regm2$y, x2)$y
mtC2 <- mtC2 - mean(mtC2)

cc1 <- sum((mtC1 - mr1)^2)/(sum(mr1^2) + 0.0001)
cc2 <- sum((mtC2 - mr2)^2)/(sum(mr2^2) + 0.0001)

mr1 <- mtC1
mr2 <- mtC2

}                                             # Final do procedimento do Backfitting Clássico

fim <- Sys.time()

ase_CBFLLverd[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)
janelas_TRUE_ll_hx1[i] <- v1
janelas_TRUE_ll_hx2[i] <- v2

tempo_CBFLLverd[i] <- fim - inicio

### 2 ESTÁGIOS LOCAL LINEAR h VERDADEIRO ###

inicio <- Sys.time()

aux <- function(x) { d2m1(x)*dnorm(x,mean=medias[1],sd=sqrt(sigma[1,1])) }
ed2m1 <- integrate(aux, 0, 1)$value
thetall <- (1/n) * sum((d2m1(x1) - ed2m1)^2)
aux <- function(x) { d2m2(x)*dnorm(x,mean=medias[2],sd=sqrt(sigma[2,2])) }
ed2m2 <- integrate(aux, 0, 1)$value
theta22 <- (1/n) * sum((d2m2(x2) - ed2m2)^2)
aux <- function(x) { 1 / sqrt(2*pi) * exp((-x^2)/2) * (x^2 - 1) * m1(x) } # 2ª derivada da
norm padrao * m1(x)
a11 <- integrate(aux, 0, 1)$value
aux <- function(x) { 1 / sqrt(2*pi) * exp((-x^2)/2) * (x^2 - 1) * m2(x) } # 2ª derivada da
norm padrao * m2(x)
a22 <- integrate(aux, 0, 1)$value

phi11 <- thetall + a11^2
phi22 <- theta22 + a22^2
phi12 <- a11*a22
R <- 1/sqrt(2*pi)
h01 <- (R / (n*phi11))^(.2) # Chute Inicial Regra de Bolso de hx1
h02 <- (R / (n*phi22))^(.2) # Chute Inicial Regra de Bolso de hx2

sistema <- function(x) {
y <- numeric(2)
y[1] <- x[1]^5 * phi11 + x[2]^2 * x[1]^3 * phi12 - R/n
y[2] <- x[2]^5 * phi22 + x[1]^2 * x[2]^3 * phi12 - R/n
y
}

jacobiano <- function(x) {
y <- numeric(4)
y <- matrix(y,2,2)
y[1,1] <- 5*x[1]^4 * phi11 + x[2]^2 * 3* x[1]^2 * phi12
y[2,2] <- 5*x[2]^4 * phi22 + x[1]^2 * 3* x[2]^2 * phi12
y[1,2] <- 2 * x[2] * x[1]^3 * phi12
y[2,1] <- 2 * x[1] * x[2]^3 * phi12
y
}

### Resolvendo o Sistema ###

janelas_2E_verd <- nleqslv(c(h01, h02), sistema, jac=jacobiano) # Janelas verdadeiras do 2E
janelas_TRUE_2E_hx1[i] <- janelas_2E_verd$x[1] # Janela verdadeira do 2E de x1
janelas_TRUE_2E_hx2[i] <- janelas_2E_verd$x[2] # Janela verdadeira do 2E de x2

### Ajustando o modelo dos 2E ###

```

```

modelo <-
hybr(y,x1,x2,janelas_2E_verd$x[1],janelas_2E_verd$x[2],janelas_2E_verd$x[1],janelas_2E_verd
$x[2])
mr1 <- modelo$m1
mr2 <- modelo$m2

ase_2Everd[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

fim <- Sys.time()

tempo_2Everd[i] <- fim - inicio

### BACKFITTING CLÁSSICO NADARYA-WATSON COM h VERDADEIRO ###

inicio <- Sys.time()

yb <- mean(y) # y barra
aux <- function(x) { d2m1(x)*dnorm(x,mean=medias[1],sd=sqrt(sigma[1,1])) }
ed2m1 <- integrate(aux, 0, 1)$value
theta11 <- (1/n) * sum((d2m1(x1) - ed2m1)^2)
aux <- function(x) { d2m2(x)*dnorm(x,mean=medias[2],sd=sqrt(sigma[2,2])) }
ed2m2 <- integrate(aux, 0, 1)$value
theta22 <- (1/n) * sum((d2m2(x2) - ed2m2)^2)
v1 <- ((1/(2*sqrt(pi)*n*n*theta11)) *
sum(1/dnorm(x1,mean=medias[1],sd=sqrt(sigma[1,1]))))^(.2)
v2 <- ((1/(2*sqrt(pi)*n*n*theta22)) *
sum(1/dnorm(x2,mean=medias[2],sd=sqrt(sigma[1,1]))))^(.2)
mtC1 <- llr(x1, (y-yb), v1)
mtC2 <- llr(x2, (y-yb), v2)
mr1 <- mtC1 - mean(mtC1)
mr2 <- mtC2 - mean(mtC2)
cc1 <- 1 # Erro Inicial do componente de x1
cc2 <- 1 # Erro Inicial do componente de x2

while (cc1 >= 0.0001 | cc2 >= 0.0001) { # Início do procedimento do Backfitting
Clássico

regm1 <- locpoly(x1, (y - yb - mtC2), bandwidth = v1, degree=0)
mtC1 <- approx(regm1$x, regm1$y, x1)$y
mtC1 <- mtC1 - mean(mtC1)

regm2 <- locpoly(x2, (y - yb - mtC1), bandwidth = v2, degree=0)
mtC2 <- approx(regm2$x, regm2$y, x2)$y
mtC2 <- mtC2 - mean(mtC2)

cc1 <- sum((mtC1 - mr1)^2)/(sum(mr1^2) + 0.0001)
cc2 <- sum((mtC2 - mr2)^2)/(sum(mr2^2) + 0.0001)

mr1 <- mtC1
mr2 <- mtC2

} # Final do procedimento do Backfitting Clássico

fim <- Sys.time()

ase_CBFNWverd[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

tempo_CBFNWverd[i] <- fim - inicio

### SMOOTH BACKFITTING COM h ESTIMADA ###

inicio <- Sys.time()

hx1 <- dpill(x1,y)
hx2 <- dpill(x2,y)
dados <- cbind(y, x1, x2)
modelo <- sBF(dados, bw=c(hx1,hx2))
mr1 <- modelo$mxhat[,1] # Primeiro Componente
mr2 <- modelo$mxhat[,2] # Segundo Componente

ase_SBFest[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

fim <- Sys.time()

tempo_SBFest[i] <- fim - inicio

### SMOOTH BACKFITTING COM h VERDADEIRO ###

inicio <- Sys.time()

hx1 <- v1

```

```

hx2 <- v2
dados <- cbind(y, x1, x2)
modelo <- sBF(dados, bw=c(hx1,hx2))
mr1 <- modelo$mxhat[,1] # Primeiro Componente
mr2 <- modelo$mxhat[,2] # Segundo Componente

ase_SBFverd[i] <- (1/n) * sum((mr1 + mr2 - m1(x1) - m2(x2) + mean(y))^2)

fim <- Sys.time()

tempo_SBFverd[i] <- fim - inicio

valores_validos <- matrix(0,1,2) # Revalida o Procedimento 'while' de geração dos dados
}

tempomedio_CBFLLest <- mean(tempo_CBFLLest)
tempomedio_CBFLLverd <- mean(tempo_CBFLLverd)
tempomedio_CBFNWest <- mean(tempo_CBFNWest)
tempomedio_CBFNWverd <- mean(tempo_CBFNWverd)
tempomedio_SBFest <- mean(tempo_SBFest)
tempomedio_SBFverd <- mean(tempo_SBFverd)
tempomedio_2Eest <- mean(tempo_2Eest)
tempomedio_2Everd <- mean(tempo_2Everd)

mase_CBFLLest <- mean(ase_CBFLLest)
mase_CBFLLverd <- mean(ase_CBFLLverd)
mase_CBFNWest <- mean(ase_CBFNWest)
mase_CBFNWverd <- mean(ase_CBFNWverd)
mase_SBFest <- mean(ase_SBFest)
mase_SBFverd <- mean(ase_SBFverd)
mase_2Eest <- mean(ase_2Eest)
mase_2Everd <- mean(ase_2Everd)

aux <-
c(ase_CBFLLest,ase_CBFLLverd,ase_CBFNWest,ase_CBFNWverd,ase_SBFest,ase_SBFverd,ase_2Eest,as
e_2Everd)

boxplot(ase_CBFLLest,ase_CBFLLverd,ase_CBFNWest,ase_CBFNWverd,ase_SBFest,ase_SBFverd,ase_2E
est,ase_2Everd,main="Monte Carlo (1° Cenário)", col=c(2:8,"orange"),
names=c("CBFLLest","CBFLLverd","CBFNWest","CBFNWverd","SBFest","SBFverd","2Eest","2Everd"),
xlab="Estimadores", ylab="ASE das iterações",cex.axis=0.53,ylim=c(min(aux),max(aux)))
savePlot(filename="BoxPlot (1° Cenário)", type=c("jpeg"))

aux <- log(janelas_TRUE_ll_hx1) - log(janelas_ll_hx1)
h <- bw.SJ(aux) # Janela de Sheater & Jones
plot(density(aux, bw=h), main="Densidade hx1 LL (1° Cenário)", xlab="log(hverd)-log(hest)")
savePlot(filename="Densidade m1 LL (1° Cenário)", type=c("jpeg"))

aux <- log(janelas_TRUE_ll_hx2) - log(janelas_ll_hx2)
h <- bw.SJ(aux)
plot(density(aux, bw=h), main="Densidade hx2 LL (1° Cenário)", xlab="log(hverd)-log(hest)")
savePlot(filename="Densidade m2 LL (1° Cenário)", type=c("jpeg"))

aux <- log(janelas_TRUE_2E_hx1) - log(janelas_2E_hx1)
h <- bw.SJ(aux)
plot(density(aux, bw=h), main="Densidade hx1 2E (1° Cenário)", xlab="log(hverd)-log(hest)")
savePlot(filename="Densidade m1 2E (1° Cenário)", type=c("jpeg"))

aux <- log(janelas_TRUE_2E_hx2) - log(janelas_2E_hx2)
h <- bw.SJ(aux)
plot(density(aux, bw=h), main="Densidade hx2 2E (1° Cenário)", xlab="log(hverd)-log(hest)")
savePlot(filename="Densidade m2 2E (1° Cenário)", type=c("jpeg"))

save.image("Resultados (1° Cenário).Rdata")

rm(list=ls(all=TRUE)) # Remove todos os objetos da área de trabalho

```