Estimação de Regressões Aditivas via *Backfitting* e Integração Marginal: Performance em Amostras Finitas*

Autor: Fernando A. Boeira Sabino da Silva†
Orientador: Carlos Martins-Filho‡
Escola de Pós-Graduação em Economia - EPGE
Fundação Getúlio Vargas

Maio de 2001

Resumo

Nesta dissertação realizou-se um experimento de Monte Carlo para revelar algumas características das distribuições em amostras finitas dos estimadores *Backfitting*(B) e de Integração Marginal(MI) para uma regressão aditiva bivariada. Está-se particularmente interessado em fornecer alguma evidência de como os diferentes métodos de seleção da janela h_n, tais como os métodos *plug-in,* impactam as propriedades em pequenas amostras dos estimadores. Está-se interessado, também, em fornecer evidência do comportamento de diferentes estimadores de h_n relativamente a seqüência ótima de h_n que minimiza uma função perda escolhida. O impacto de ignorar a dependência entre os regressores na estimação da janela é também investigado. Esta é uma prática comum e deve ter impacto sobre o desempenho dos estimadores. Além disso, não há nenhuma rotina atualmente disponível nos pacotes estatísticos/econométricos para a estimação de regressões aditivas via os métodos de *Backfitting* e Integração Marginal. É um dos objetivos a criação de rotinas em Gauss para a implementação prática destes estimadores. Por fim, diferentemente do que ocorre atualmente, quando a utilização dos estimadores-B e MI é feita de maneira completamente *ad-hoc*, há o objetivo de fornecer a usuários informação que permita uma escolha mais objetiva de qual estimador usar quando se está trabalhando com uma amostra finita.

*O autor é responsável por quaisquer erros remanescentes.

†epsabino@fgv.br.
‡carlos.martins@orst.edu.

Abstract

In this thesis we conduct a Monte Carlo investigation to reveal some characteristics of the small sample distributions of the Backfitting (B) and Marginal Integration (MI) estimators for an additive bivariate regression. We are particularly interested in providing some evidence on how different data driven window width estimation procedures, such as some plug in methods impact the small sample properties of the MI and B estimators. We are also interested in providing evidence on the behavior of how the differente window widths estimators impact the optimal sequence of window widths that minimizes a chosen loss function. The impact of ignoring regressor dependency on window width estimation is also investigated. This is common practice and should impact estimators' performance. Besides, nowadays there no available statistical/econometrical packages that perform estimation of additive regression by Backfitting and Marginal Integration. It´s an objective of our dissertation the creation of routines in Gauss for the practical implementation of these estimators. Ultimately, differently from what occurs at the present time, when the utilization of the B e MI estimators is done in a way completely *ad-hoc*, our objective is to provide applied researches with information that allows for a more accurate comparison of these two competing alternatives in a finite sample setting.

Agradecimentos

Em primeiro lugar, gostaria de agradecer o prof. Carlos Martins-Filho que superou em muito as expectativas do que eu esperava de um orientador. A motivação que tive ao longo do trabalho deveu-se em grande parte a sua boa vontade ímpar, mesmo quando ele se encontrava fora do Brasil.

Agradeço também aos professores Jandyra M.G. Fachel, João Riboldi, Dinara W.X. Fernandez, , Jorge Ricardo Ducatti e Rudnei Dias da Cunha da Universidade Federal do Rio Grande do Sul pela cessão de seus laboratórios e equipamentos, além de Alexandre Rademaker, Viviane dos Santos, Marcos de Barros e Fernando Hepp Pulgatti pelo apoio técnico.

Agradeço muito a minha turma. Uma turma especial, que me proporcionou uma convivência social e intelectual inesquecível e da qual sentirei muitas saudades.

Em especial agradeço ao colega Alexandre Manoel Ângelo da Silva, aos alunos da EPGE Genaro Dueire Lins, Enestor da Rosa dos Santos Júnior, Gregório Silva Caetano e Pedro Alberto Chauffaile Saffi. Agradeço ainda a Martinho Roberto Lazzari.

Gostaria ainda de agradecer a Cristiano Augusto Coelho Fernandes e Marcelo Côrtes Neri pela leitura atenta da dissertação e pelos comentários.

Agradeço também ao prof. Marcos de Barros Lisboa pelo apoio durante o primeiro ano do curso.

Para finalizar eu agradeço as pessoas mais próximas a mim: minha família e Andréa. Estas pessoas foram a que tiveram maiores custos, devido ao tempo em que estive afastado deles trabalhando na dissertação.

Sumário

1 Introdução **5**

2 Estimação Não Paramétrica **9**

3 Especificação do Modelo e os Estimadores sob Estudo **17**
3.1 O Estimador *Backfitting* 20

3.2 O Estimador de Integração Marginal 22

4 Métodos de seleção automática da janela h_n **24**

5 O Processo de Geração dos Dados **27**

6 Resultados **28**

7 Conclusões **33**

1 Introdução

A estimação de regressões não paramétricas tem sido objeto de vários artigos recentes. Regressão não paramétrica tornou-se um campo de rápido desenvolvimento desde que se notou que regressão paramétrica não era adequada para muitos conjuntos de dados. Sem dúvida um dos problemas mais estudados em Estatística e Econometria está relacionado a estimação de funções paramétricas de regressão. A necessidade, bem como a adequação da modelagem de funções de regressão paramétricas são amplamente discutidas em um grande número de livros, ver, por exemplo, Draper & Smith(1998), Johnston & Dinardo(1996) e Greene(1999). Porém, as boas propriedades dos estimadores de máxima verossimilhanca e dos estimadores de mínimos quadrados, dependem da forma funcional paramétrica assumida estar corretamente especificada. Caso contrário, os estimadores de máxima verossimilhança e de mínimos quadrados não serão eficientes, assintoticamente normais e nem mesmo consistentes. Uma alternativa menos restrita para a estimação de uma função de regressão é o modelo de regressão não paramétrica $Y = E(Y \mid X = x) + \epsilon = m(x) + \epsilon$, onde $m(x)$ pertence a uma classe de funções bem mais abrangente do que a de modelos paramétricos. Considera-se neste cenário $d + 1 = p$ variáveis (Y, X) onde Y é o regressando e X é um vetor $(d \times 1)$ de regressores[1] e estas p variáveis são completamente caracterizadas pela sua função densidade de probabilidade conjunta $f(y; x_1, ..., x_d) = f(y, x)$ nos pontos y, x. Assume-se também que $\left\{(y_t, x_{1t}, ..., x_{dt})'\right\}_{t=1}^{n}$ formam uma seqüência de realizações de um vetor aleatório IID avaliado em $\Re^{d+1}$ e $\{\epsilon_t\}_{t=1}^{n}$ é uma seqüência de realizações de uma v.a. com $E(\epsilon_t \mid X_1, ..., X_d) = 0$ e $V(\epsilon_t \mid X_1, ..., X_d) = \sigma^2$ $\forall t$. Um grande número de procedimentos de estimação foi sugerido, porém a estimação da função $m = m(x)$ baseada em núcleo estimadores (*kernel*, em inglês) ganhou uma maior aceitação pelo fato de suas propriedades estatísticas estarem bem estabelecidas sob vários cenários de geração dos dados. A idéia em regressão não paramétrica é estimar a função $m(x)$ (isto é, obter $\widehat{m}(x)$) de maneira próxima a verdadeira função, dada a amostra. Analisando o caso em que X é uma variável aleatória contínua, m será estimada através da ponderação dos pontos y_t, correspondentes aos pontos x_t em um intervalo em torno do ponto onde está-se estimando x. Este intervalo conhecido como janela (ou largura da banda, ou ainda parâmetro de suavização) na literatura, determina quão local será a estimação. A janela[2], denotada por h_n, pode ser vista como um parâmetro que depende da amostra n e que deve convergir para zero quando $n \to \infty$.

[1] Quando os regressores são aleatórios, chama-se o modelo de desenho aleatório, em contraposição ao modelo de desenho fixo, onde os regressores x são fixos. Igualmente ao caso dos modelos de regressão usuais, em geral, os estimadores para os modelos de desenho aleatório e fixo coincidem assintoticamente, veja Brown, Cai, Low and Zhang(2002).

[2] Como comentar-se-á posteriormente, a escolha da janela é um dos elementos de maior importância em estimação não paramétrica.

Contrastando com a abordagem paramétrica, não é exigido dos estimadores não paramétricos da função de regressão que a forma funcional de m 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. Infelizmente, a convergência dos estimadores não paramétricos torna-se mais lenta a medida que o 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 freqüentemente impossível na prática. Pode-se verificar que o número de elementos na amostra para fazer-se a estimação cresce exponencialmente quando o número de regressores aumenta, se a janela for mantida constante. Demonstra-se[3] que a melhor taxa possível de convergência é $n^{-q/(2q+d)}$, onde d é a dimensão de X e m é uma função continuamente diferenciável q vezes. Este problema é conhecido na literatura como maldição da dimensionalidade (*curse of dimensionality*, em inglês) e foi identificada por Friedman & Stuetzle(1981). Outro problema surge quando se define vizinhanças em dimensões maiores ou iguais a 2 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 disso, de uma perspectiva prática, suavizadores multivariados são extremamente caros para computar, e a habilidade para visualizar e/ou interpretar a relação quando temos mais de três regressores é virtualmente perdida. Desta forma, desde que um dos objetivos de modelos econométricos/estatísticos é facilitar a interpretação e isolar a contribuição de um regressor, mantendo todas as outras variáveis fixas, o uso de uma regressão não paramétrica irrestrita é freqüentemente indesejado.

Um meio de evitar os problemas citados acima é restringir o espaço em que irá se executar a estimação, isto é, fazer alguma suposição a priori sobre a natureza da expectativa condicional. Quatro importantes métodos propostos na literatura são o modelo aditivo generalizado de Hastie & Tibshirani(1986, 1990), o método de Friedman & Tukey(1974), conhecido como *project pursuit regression* (PPR), o método de Breiman & Friedman(1985) e Friedman & Stuetzle(1981) conhecido por ACE e os procedimentos de redes neurais desenvolvidos por Kuan & White(1994). Todas estas suposições levam a taxas de convergência que são independentes do problema da dimensionalidade, apesar desta independência ser possivelmente artificial, devido a imposição de uma estrutura na função.

O modelo aditivo (AM), com variável resposta Y e variáveis explicativas $X \in \Re^d$ tem a forma

[3] Ibragimov e Hasminskii(1983), e Stone(1982) .

$$E(Y_t \mid X_{1t} = x_{1t}, X_{2t} = x_{2t}, ..., X_{dt} = x_{dt}) = m_0 + \sum_{d=1}^{D} m_d(x_{dt}), \qquad (1)$$

para $t = 1, ...n$, onde $E(Y) = m_0$ e as funções univariadas, também chamadas de componentes aditivos, m_d obedecem a condição $E[\ m_d(x_d)] = 0$, $\forall d$ para identificação. Uma possível extensão é o modelo aditivo generalizado (GAM). Modelos aditivos generalizados são da forma

$$E(Y_t \mid X_{1t} = x_{1t}, X_{2t} = x_{2t}, ..., X_{Dt} = x_{Dt}) = G\left\{ m_0 + \sum_{d=1}^{D} m_d(x_{dt}) \right\} \qquad (2)$$

onde G é uma função de ligação conhecida, m_d tem as mesmas condições vistas acima e $m_0 = G^{-1}\{E(Y)\}$. Casos especiais deste modelo são os modelos de regressão logit e probit. Pode-se notar que o modelo aditivo é uma generalização de um modelo linear e, portanto, retém uma característica importante deste modelo, no qual a natureza do efeito de uma variável em uma superfície de resposta não depende dos valores de qualquer outra variável, ou seja, podemos verificar a contribuição de cada variável em predizer a resposta individualmente. Como pode-se notar, modelos aditivos permitem a flexibilidade de uma regressão não paramétrica e garantem um comportamento assintótico razoável.

Atualmente dois estimadores são alternativas viáveis para a estimação de um modelo não paramétrico aditivo - o estimador *Backfitting* (estimador-B) e o estimador de Integração Marginal (estimador-MI). O estimador-B é baseado em Friedman & Stuetzle(1981), porém ganhou popularidade através dos trabalhos de Hastie & Tibshirani(1986,1990). Suas propriedades foram estudadas em Buja, Hastie & Tibshirani(1989) e Opsomer & Ruppert(1997). Atualmente, pouco se sabe sobre as propriedades estatísticas do estimador-B. Em geral, ainda não é possível construir intervalos de confiança válidos assintoticamente para a regressão estimada, mesmo quando $h_n \to 0$ a uma taxa desejada. O conhecimento das propriedades do estimador-B é ainda mais escasso, quando h_n é escolhida ao minimizar as funções critério mais utilizadas na literatura. Conseqüentemente, na prática, pouco se sabe sobre as propriedades assintóticas e em amostras finitas do estimador-B.

O estimador-MI foi introduzido através dos artigos seminais de Linton & Nielsen(1995) e Linton & Härdle(1996). Uma das propriedades mais atraentes do estimador-MI é que se pode mostrar que ele é assintoticamente normal quando a janela h_n converge a zero a uma taxa pré-especificada. Entretanto, a sua distribuição assintótica é ainda desconhecida quando h_n é escolhida pelos métodos disponíveis na literatura (dependentes dos dados), tais como validação cruzada,

e vários métodos *plug-in*, incluindo aqueles propostos por Silverman(1986) e Opsomer & Ruppert(1998). A dificuldade em estabelecer a normalidade assintótica tem duas fontes: primeiramente, tem-se seqüências estocásticas h_n que podem interagir de uma maneira perniciosa com os regressores e o regressando, o que cria uma dificuldade adicional em estabelecer a normalidade assintótica do estimador-MI. Segundo, a janela h_n é escolhida minimizando uma função critério (perda ou risco). Para a maioria das funções critério utilizadas, a seqüência ótima resultante h_n, não converge a zero a taxa necessária para obter a normalidade assintótica. Igualmente ao estimador-B, pouco se sabe, na prática, sobre as propriedades assintóticas e em pequenas amostras do estimador-MI.

Para tornar os resultados acerca da distribuição (assintótica) atualmente disponíveis úteis é necessário adaptá-los ao caso em que h_n é uma seqüência estocástica que depende dos dados. Uma alternativa é fornecer evidência experimental do desempenho dos estimadores baseados em vários métodos de seleção da janela h_n através de um estudo Monte Carlo. Em decorrência, nesta dissertação ir-se-á conduzir um estudo de Monte Carlo para revelar algumas características das distribuições em pequenas amostras dos estimadores-B e MI para uma regressão aditiva bivariada. Está-se particularmente interessado em fornecer alguma evidência de como os diferentes métodos de seleção da janela h_n, tais como os métodos *plug-in* impactam as propriedades em pequenas amostras dos estimadores. Está-se interessado, também, em fornecer evidência do comportamento de diferentes estimadores de h_n relativamente a seqüência ótima de h_n que minimiza uma função perda escolhida. O impacto de ignorar a dependência entre os regressores na estimação da janela é também investigado. Esta é uma prática comum e deve impactar a performance dos estimadores. Além disso, não há nenhuma rotina atualmente disponível nos pacotes estatísticos/econométricos para a estimação de regressões aditivas via os métodos de *Backfitting* e Integração Marginal. É um dos objetivos a criação de rotinas em Gauss para a implementação prática destes estimadores. Por fim, diferentemente do que ocorre atualmente, quando a utilização dos estimadores-B e MI é feita de maneira completamente *ad-hoc*, há o objetivo de fornecer a usuários informação que permita uma escolha mais criteriosa de qual estimador usar quando se está trabalhando com uma amostra finita. Além da introdução, esta dissertação tem mais seis seções. A seção 2 faz uma breve revisão dos métodos a serem utilizados. A seção 3 descreve a especificação do modelo e os dois estimadores sob estudo em um formato unificado. A seção 4 fornece um descrição dos métodos de seleção da janela h_n utilizados. A seção 5 apresenta o processo de geração dos dados a ser utilizado no estudo de Monte Carlo. Na seção 6 discute-se os resultados da análise. Por fim, na seção 7 faz-se as conclusões.

2 Estimação Não Paramétrica

A maneira mais comum de estimar-se uma função de regressão é optar por um modelo paramétrico linear ou não linear. Modelos paramétricos de regressão são ferramentas poderosas para modelar uma expectativa condicional, permitindo previsões para valores futuros de Y e a construção de intervalos de confiança e testes para os parâmetros.

Na abordagem paramétrica é possível especificar-se uma família de formas funcionais para m de maneira errada. Este problema, desastroso para a abordagem paramétrica, inexiste 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 paramétricos[4]. O objetivo da regressão não paramétrica é aproximar m. 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 caso paramétrico, quando o modelo paramétrico especificado for o correto, devido às taxas de convergência dos estimadores não paramétricos serem mais lentas do que a dos estimadores paramétricos.

Os dois problemas mais abordados na literatura não paramétrica são a estimação de densidades e a estimação de uma função de regressão. Um método amplamente utilizado para estes fins são os estimadores baseados em núcleo. Os estimadores núcleo são simplesmente um modo de ponderar as observações, de maneira que as mais próximas do ponto onde se deseja estimar recebam um peso maior e as mais afastadas um peso menor. A função núcleo, denotada na dissertação por $K(\cdot)$ será uma função real $K(\cdot) : \Re^d \longrightarrow \Re$, contínua, limitada, tal que $\int \Psi\Psi^T K(\Psi)d\Psi = \mu_2(K)\mathbf{I}$, onde $\mu_2(K)$ é um escalar e $\mathbf{I}$ é uma matriz identidade $d \times d$ - , com suporte compacto e simétrica, que integra a um, isto é:

$$\int_{-\infty}^{+\infty} K(\Psi)d\Psi = 1. \tag{3}$$

Além disso, todos os momentos de ordem ímpar de K são zero, ou seja, $\int \Psi_1^{l_1}\Psi_2^{l_2}...\Psi_d^{l_d}K(\Psi)d\Psi = 0$, $\forall l_1, l_2, ..., l_d$ tal que a sua soma seja ímpar.

O estimador de densidades univariadas mais comum é o histograma (Fix & Hodges, 1951). A idéia básica da estimação não paramétrica é simplesmente

[4] A classe de funções a que m pode pertencer é bastante vasta, sendo a única restrição imposta a de que m seja uma função continuamente diferenciável.

construir um histograma suave ajustado aos dados. Para isso, precisa-se de um parâmetro de suavização, a janela, que é a dimensão do intervalo que utilizar-se-á para conseguir a suavização. O histograma porém é uma medida um tanto rudimentar, pelo fato de poder apresentar descontinuidades e saltos, quando a função verdadeira for contínua e suave. Isto ocorre, devido ao fato do histograma usar a função indicadora como função de ponderação (veja Pagan & Ullah, 1999, p.7). Rosenblatt(1956) trocou a função indicadora por uma função núcleo satisfazendo as propriedades dadas acima. O estimador de uma função densidade f, pode ser escrito, então, da seguinte forma:

$$\widehat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\Psi_i\right), \tag{4}$$

onde x é o ponto no qual deseja-se estimar f , x_i são os elementos de uma amostra de tamanho n e h_n é o comprimento fixo do intervalo (janela) utilizado para obter-se a suavização desejada. Em (4) pode-se ver os dois elementos importantes em estimação de densidades não paramétricas: a escolha de uma função de ponderação e de um "parâmetro" - a função núcleo e a janela.

Uma maneira rudimentar de escolher h_n é por tentativa e erro baseado no uso de análises gráficas. Este método consiste em olhar vários gráficos de $\widehat{f}(x)$ contra x, quando $\widehat{f}(x)$ é calculado através de diferentes valores de h_n. Como os métodos gráficos disponíveis são bastante avançados hoje em dia, este método pode ser efetivo, embora seja vantajoso escolher h_n de uma maneira mais objetiva. A escolha de h_n é crucial na estimação, pois dependendo da maneira como se escolhe a janela, o estimador pode tomar formatos muito diferentes. A escolha da função núcleo, em geral, é feita em função do custo computacional e da velocidade de convergência dos estimadores. Comenta-se adiante os motivos.

Para fazer uma escolha automática de h_n e/ou $K(.)$ é necessário ter algum critério. O objetivo é simplesmente avaliar se $\widehat{f}(x)$ é um bom estimador para a $f(x)$. Os critérios mais populares envolvem a minimização de uma função perda ou risco quadrática, dentre as quais destacam-se: ISE (*Integrated Squared Error),* MISE (*Integrated Mean Squared Error*), AMISE (*Asymptotic Integrated Mean Squared Error*), ASE (*Averaged Squared Error*), MASE (*Mean Averaged Squared Error*), AMSE (*Asymptotic Mean Squared Error*) e AMASE (*Asymptotic Mean Averaged Squared Error*). Todos estes critérios podem ser escritos através do viés e da variância da $\widehat{f}$. As aproximações assintóticas permitem que se escreva o critério a ser minimizado de uma maneira mais simples. Isto é geralmente feito, devido ao fato de que expressões exatas para os critérios serem freqüentemente difíceis de serem obtidas. Isto decorre do fato de que o viés e a variância da $\widehat{f}$ serem funções complicadas de h_n. O leitor pode consultar Simonoff(1996), Pagan

& Ullah(1999) e Opsomer e Ruppert(1998) para ver exemplos de expressões destes critérios.

Como se pode notar pela discussão acima, é necessário que as preferências do usuário no que tange a viés e variância sejam refletidas de algum modo, quando ele escolhe a função critério a ser minimizada. A variância da $\widehat{f}$ diminui quando h_n torna-se maior e vice-versa. O viés da $\widehat{f}$ aumenta quando h_n aumenta e vice-versa, veja Simonoff(1996) e Pagan & Ullah(1999) para ver alguns exemplos. Logicamente ao escolher uma função critério a ser minimizada o pesquisador necessita fazer um *tradeoff* entre viés e variância. Se h_n for escolhido grande estar-se-á suavizando muito a função, o que gera o fenômeno conhecido na literatura como *oversmoothing*. O contrário, isto é, h_n muito pequeno, acarretará numa função subsuavizada, gerando o fenômeno conhecido por *undersmoothing*. O ideal é escolher h_n de maneira a não gerar nenhum dos dois fenômenos descritos. Note que h_n é fixo[5] na discussão feita até agora. Logicamente, não há a necessidade disso, podendo o valor da janela ser modificado, de acordo com o número de dados disponíveis nas regiões de estimação. Na dissertação o enfoque é baseado em h_n fixo dentro do suporte utilizado.

Mesmo quando um critério para a escolha de h_n e de $K(\cdot)$ tiver sido selecionado, ainda existem duas maneiras de proceder-se. Na primeira, conhecida como validação cruzada, termos desconhecidos na função critério são trocados por estimativas amostrais e então a minimização é feita com relação a h_n e $K(\cdot)$. Este método caracteriza-se por subsuavizar a função estimada. Na segunda alternativa, denominada de métodos *plug-in*, a função critério é minimizada analiticamente e uma solução para h_n é encontrada, e na seqüência, troca-se qualquer função populacional por uma estimativa amostral. Jones, Marron & Sheater(1996) comentam que os métodos *plug-in* são superiores aos métodos de validação cruzada, em estudos de simulação e assintoticamente[6]. Na seção 4, os métodos utilizados na dissertação são tratados de maneira específica.

A escolha da função núcleo $K(\cdot)$, apesar de desempenhar papel fundamental na estimação não paramétrica, é uma tarefa de menor dimensão que a escolha de h_n. Estudos demonstram que a função núcleo dita ótima, no sentido que minimiza as funções critério comumente utilizadas, leva a pequenas melhorias em relação a maioria das funções núcleo utilizadas. Conseqüentemente, a simplicidade, o custo computacional, e a velocidade de convergência dos estimadores freqüentemente

[5] Diz-se que o intervalo h_n é fixo quando seu tamanho é mantido constante independente da região de estimação.

[6] Os métodos *plug-in* demandam menos tempo computacional, não apresentam o problema de subsuavização do método de validação cruzada e a taxa de convergência de $\left(\widehat{h}_n - h_n\right) \to 0$ quando $\widehat{h}_n$ é escolhido por métodos *plug-in* é mais rápida do que a taxa de convergência de $\left(\widetilde{h}_n - h_n\right) \to 0$ quando $\widetilde{h}_n$ é obtido por validação cruzada.

determinam a escolha da função núcleo. A função núcleo ótima foi primeiramente sugerida por Epanechnikov(1969) e é por isso, freqüentemente, chamada de função núcleo Epanechnikov. A função núcleo Epanechnikov (para o caso univariado) é dada por:

$$K_e(\Psi) = \begin{cases} 0.75(1-\Psi^2), \text{ se } |\Psi| \leq 1, \Psi \in \Re \\ 0, \text{ caso contrário} \end{cases} \tag{5}$$

A função núcleo Epanechnikov[7] tem formato quadrático, e pode ser vista em Härdle(1990, p.26) e Silverman(1986, p.42). Outra função núcleo comumente utilizada e que não apresenta descontinuidades é a função núcleo gaussiana, isto é,

$$K(\Psi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\Psi^2\right\}, -\infty < \Psi < +\infty. \tag{6}$$

Simonoff(1996, p.44) e Silverman(1986, p.43) mostram uma tabela onde se comparam a eficiência das funções núcleo comumente utilizadas em relação a função Epanechnikov. Pode-se notar, que até a função núcleo mais "ingênua", isto é, a função uniforme[8] tem uma eficiência de aproximadamente 0.93 da função Epanechnikov. Na dissertação utilizar-se-á apenas o núcleo gaussiano.

Tratar-se-á agora da estimação de uma função de regressão. Utilizando a notação vista na página 5, o objetivo é a modelagem não paramétrica de expectativas condicionais, isto é, $Y = m(x) + \varepsilon = E(Y \mid X = x) + \varepsilon$. O objetivo é estimar a função $m(x)$. Por definição, pode-se escrever

$$\begin{aligned} m(x) &= E(Y \mid X = x) \\ &= \int_{-\infty}^{+\infty} y f_{Y|X}(y \mid x) dy \\ &= \int_{-\infty}^{+\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} dy \end{aligned} \tag{7}$$

onde $f_X(x)$, $f_{X,Y}(x,y)$, e $f_{Y|X}(y \mid x)$ 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

[7] Esta função pode apresentar problemas computacionais (*overflow*) se a amostra for muito pequena, quando se estiver utilizando determinados estimadores em modelos de regressão, como por exemplo o estimador linear local que será discutido adiante.

[8] É dita "ingênua", pois a ponderação que recebem as observações que estão mais próximas do ponto onde se deseja estimar é a mesma dada aos pontos mais afastados, isto é, a ponderação independe da distância.

como Nadaraya-Watson. Uma estimativa núcleo de $f_{X,Y}(x,y)$ é uma extensão bivariada de (4) e foi primeiramente extendida por Cacoullos(1966). Escreve-se esta estimativa por:

$$\widehat{f}_{X,Y}(x,y) = \frac{1}{nh_{nx}h_{ny}} \sum_{i=1}^{n} K_x \left(\frac{x - x_i}{h_{nx}} \right) K_y \left(\frac{y - y_i}{h_{ny}} \right). \tag{8}$$

Usando (4) e (8), reparando que $\int K_y(u)du = 1$ e $\int uK_y(u)du = 0$, supondo que $h_{nx} = h_{ny} = h_n$ e que a função núcleo é simétrica, pode-se escrever o estimador de Nadaraya-Watson por:

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_x \left(\frac{x - x_i}{h_n} \right) y_i}{\sum_{i=1}^{n} K_x \left(\frac{x - x_i}{h_n} \right)}. \tag{9}$$

O estimador de Nadaraya-Watson pode ser visto como uma estimativa local de $m(x)$. Dentro da classe dos estimadores polinomiais locais (veja Simonoff, 1996 e Pagan & Ullah, 1999) se eu quiser um polinômio de ordem zero, $\widehat{m}_{NW}$ otimiza uma distância em L^2 ponderado. No sentido de projeção este estimador será, então, ótimo, fornecendo, portanto, a melhor estimativa local da m. A idéia de ponderação local de m é útil para ajudar a perceber algumas características desejáveis que um estimador deve ter. Para valores de x_i que estão distantes de x, mantendo constante a janela h_n, Ψ_i será "grande" e conseqüentemente K_i será "pequeno". Desta forma, estas observações receberão pouca importância na determinação de m, tal como deveria ser. Desta forma, é necessário que $h_n \to 0$ quando $n \to \infty$, de maneira que as observações que não estão muito perto de x recebam pouco peso ou não participem[9] da estimação quando a amostra aumentar. Outra característica interessante, é que o formato da função núcleo não é muito importante dado que se dá um peso pequeno as observações x_i que estão distantes do ponto x. Para entender melhor esta última questão, é útil pensar na situação em que x_i é uma v.a.discreta (veja Pagan & Ullah, 1999). Seja um dos valores que x_i pode assumir denotado por x e sejam as observações correspondentes a este valor arranjadas como as primeiras n^*. Desta forma, uma maneira sensível de estimar $m(x)$ seria ponderar todos os y_i correspondentes aos x_i's que são iguais a x, isto é, fazer $\widehat{m} = (n^*)^{-1} \sum_{i=1}^{n*} y_i$. Dado que $\widehat{m} = m + (n^*)^{-1} \sum_{i=1}^{n^*} \epsilon_i$, e $E(\epsilon_i) = 0$. A aplicação de uma lei dos grandes números mostra que $\widehat{m}$ é um estimador consistente de m desde que $n^* \to \infty$ quando $n \to \infty$.

O estimador de Nadaraya-Watson tornou-se popular, pois assumindo-se algumas condições a respeito da janela h_n este estimador terá propriedades desejáveis. As condições mais importantes são as de convergência, isto é, $h_n \to 0$ quando

[9] Depende da função núcleo utilizada.

$n \to \infty$; sob determinadas condições na taxa de convergência de h_n para zero, $\widehat{m}$ é um estimador consistente para m e que $(nh_n)(\widehat{m}(x) - m(x))$ converge em distribuição para uma Normal. O leitor interessado pode consultar Simonoff(1996), Pagan & Ullah(1999), Härdle(1990) e Silverman(1996) para obter um tratamento mais formal deste estimador. A conclusão de Nadaraya-Watson é importante em termos teóricos, porém sua distribuição assintótica é questionável quando se escolhe h_n através da minimização de algum critério. A janela h_n ótima encontrada ao minimizar-se algum dos critérios já citados, em geral, não satisfaz a taxa de convergência que dá a normalidade assintótica deste estimador. Isto torna as inferências baseadas neste estimador bastante problemáticas.

Atualmente, o estimador de Nadaraya-Watson é pouco utilizado, devido principalmente ao viés deste estimador nas fronteiras da distribuição das observações x_i (veja Simonoff, 1996). Hoje em dia, os usuários comumente utilizam estimadores polinomiais locais[10] , tais como estimadores quadráticos ou linear local, que corrigem automaticamente o viés na fronteira. Na próxima seção, tratar-se-á dos estimadores polinomiais com mais detalhes. Por fim, nesta seção, vai-se comentar um pouco sobre regressão não paramétrica multivariada irrestrita, o modelo aditivo e as duas alternativas viáveis para estimar um modelo não paramétrico aditivo.

Como se comentou na introdução a estimação de uma regressão não paramétrica multivariada irrestrita é limitada pela característica multidimensional do problema. É difícil visualizar as superfícies de regressão quando temos mais de três regressores, assim como se torna uma tarefa difícil interpretar a estrutura complexa que pode surgir. Além disso, a maldição da dimensionalidade implica que quando o número de regressores aumenta, grandes quantidades de dados serão necessárias para conseguir-se uma estimação acurada. Uma maneira de evitar estas dificuldades é restringir o formato da função de regressão $m(x)$. Uma possibilidade natural é generalizar um modelo de regressão linear múltipla para permitir funções aditivas arbitrárias do tipo visto em (1). Suavizadores univariados podem ser utilizados para estimar as funções m_d, evitando assim a maldição da dimensionalidade[11], mas há um custo crucial: se a forma aditiva dada em (1) não

[10] Fan(1992), Fan, Gasser, Gijbels, Brockmann & Engel(1993) e Ruppert & Wand(1994) demonstram que os estimadores polinomiais locais possuem propriedades teóricas e práticas desejáveis em relação a outros estimadores, incluindo o estimador de Nadaraya-Watson.

[11] Como comentado em Hastie & Tibshirani(1986) sabe-se que suavizadores multivariados apresentam problemas em "grandes" dimensões (em especial a maldição da dimensionalidade). A variância de uma estimativa depende do número de pontos na vizinhança. Quando se está trabalhando em uma "grande" dimensão as estimativas não serão tão locais e podem ser muito viesadas. Esta é uma grande motivação para o uso de um modelo aditivo, pois, neste caso, usam-se suavizadores univariados. Desta forma, na prática, pode-se incluir pontos suficientes nas vizinhanças de forma a manter a variância das estimativas "pequena" (as estimativas serão mais locais em cada coordenada comparativamente ao caso em que se usam suavizadores mul-

for correta, o estimador $\widehat{m}$ não será nem mesmo consistente[12]. Porém, se m satisfizer (1), a taxa de convergência do estimador $\widehat{m}$ será igual a de um suavizador unidimensional (se não, a taxa de convergência será equivalente a aproximação aditiva mais próxima no sentido de MSE), veja Stone(1985, 1986). Como escrito na página 6, existem atualmente duas alternativas viáveis para estimar um modelo aditivo: o estimador de *Backfitting* (estimador-B) e o estimador de Integração Marginal (estimador-MI). Abaixo faz-se alguns comentários sobre estes dois estimadores, o que não dispensa a leitura dos artigos seminais pelos leitores mais interessados.

O estimador *Backfitting* projeta o problema de regressão multidimensional dentro do espaço de modelos aditivos. O método está sempre procurando pelo modelo aditivo que melhor ajusta a regressão, seja o modelo verdadeiro aditivo ou não. Se o modelo verdadeiro for um modelo aditivo como o descrito em (1), os funcionais estimados serão os componentes aditivos do modelo verdadeiro. Se o modelo verdadeiro não for aditivo, uma razão para o uso de *Backfitting* é o seu efeito de redução da dimensão em problemas de regressão multidimensionais. Mesmo que o modelo verdadeiro não seja de fato o modelo aditivo, o estimador de *Backfitting* freqüentemente leva a uma regressão razoável. Neste caso, os componentes aditivos não devem ser interpretados. Desde que o algoritmo para a estimação dos componentes aditivos é diretamente relacionado com toda a regressão, não é possível estimar cada componente separadamente. A implementação do procedimento iterativo é dada em Buja, Hastie & Tibshirani(1989), Hastie & Tibshirani(1990) e envolve a solução iterativa de equações não lineares. Sob algumas condições, é esperado que o algoritmo de *Backfitting* convirja para uma única solução. No caso em que se busca estimar não parametricamente uma regressão aditiva bivariada, o algoritmo de *Backfitting* converge para uma solução explícita, veja Opsomer & Ruppert(1997) e Opsomer & Ruppert(1998).

O estimador de Integração Marginal, como indicado pelo seu nome, estima a influência marginal de um regressor particular em um problema de regressão multidimensional. Se o modelo verdadeiro é um modelo aditivo, o funcional a ser

tivariados). É claro que o modelo aditivo pode gerar uma estimativa viesada da verdadeira superfície de regressão, porém, em geral, este viés será muito menor do que o produzido quando se usam suavizadores multivariados.

[12] Gozalo & Linton(2001) propõe testes para a suposição de aditividade dada em (1). A hipótese nula de aditividade entre os regressores é testada por quatro diferentes estatísticas de teste, implementadas por dois métodos diferentes: um baseado em um enfoque assintótico e outro em *bootstrap*. Um estudo de simulação é feito no artigo para comparar a performance das estatísticas de teste utilizadas. Interpreta-se a rejeição de H_0 como evidência de a interação entre os regressores ser forte o suficiente para rejeitar a aditividade. Termos de interação podem ser facilmente colocados em um modelo aditivo não paramétrico especificado, por exemplo, $E\left(y_t \mid x_{1t}, ..., x_{dt}\right) = m_0 + \sum_{d=1, d\neq h,l}^{D} m_d(x_{dt}) + m_{hl}(z_{lt}, z_{ht})$.

estimado será exatamente o componente aditivo. A idéia básica do procedimento de estimação é estimar um pré-estimador da superfície de regressão multidimensional e então integrar em relação as dimensões que não são de interesse, mantendo a direção de interesse fixa. Ele requer, portanto, a escolha de uma medida de integração. Linton & Härdle(1996) descrevem o procedimento em detalhes. Segundo eles a grande vantagem deste método é que se pode obter as suas propriedades assintóticas. Porém, igualmente ao já comentado na página 14 em relação ao estimador de Nadaraya-Watson, neste contexto, também, a taxa de convergência da janela h_n ótima encontrada ao minimizar os critérios citados, em geral, também não satisfaz a taxa de convergência que acarreta a normalidade assintótica deste estimador. Linton & Nielsen(1995) apontam ainda como vantagens do método, o fato dele ser mais simples de implementar do que o algoritmo de *Backfitting* e que devido a sua construção permitir a escolha entre um modelo aditivo e um multiplicativo, pois ambas as estruturas são consistemente estimadas pelo mesmo procedimento.[13] Porém, existem duas desvantagens principais do estimador-MI em relação ao estimador-B. Primeiramente, a demanda computacional deste estimador é superior a do estimador-B. Para estimar os componentes aditivos em um modelo bivariado usando regressão linear local, ele necessita computar n^2 regressões suavizadoras avaliadas nos pares (X_i, Z_j), $i, j = 1, ..., n$ cada uma das quais requer $O(nh_n)$ operações quando a função núcleo tem suporte compacto. Logo, será necessário computar $O(n^3 h_n)$ operações para estimar os componentes aditivos (Kim, Linton & Hengartner, 1999).[14]Para o estimador-B o tempo computacional é da ordem $O(n^2 h_n m)$ onde m é o número de iterações requeridas para alcançar am convergência (em geral requer menos operações). Uma segunda desvantagem é que o estimador-MI é estatisticamente ineficiente em relação ao estimador-B (Linton, Nielsen & Geer, 2000) de acordo com um critério MSE usual, o que talvez seja a principal motivação para a implementação da metodologia de *Backfitting*. Na próxima seção vai-se escrever os dois estimadores sob estudo em um formato unificado.

[13] Além disso, o método de integração marginal é, talvez, mais fácil de ser entendido por não estatísticos, pelo fato de envolver ponderações ao invés da solução iterativa de equações não lineares, como no algoritmo de *Backfitting*.

[14] O número de operações requeridas é desta ordem para outros estimadores, incluindo o estimador de Nadaraya-Watson.

3 Especificação do Modelo e os Estimadores sob Estudo

O modelo de regressão que é considerado aqui pode ser genericamente descrito por

$$Y_t = \sum_{d=1}^{D} m_d(X_{dt}) + \epsilon_t \tag{10}$$

para $t = 1, ...n$. Assume-se que $\left\{(y_t, x_{1t}, ..., x_{dt})'\right\}_{t=1}^{n}$ formam uma seqüência de realizações de um vetor aleatório IID $(Y, X_1, ..., X_d)$ avaliado em $\Re^{d+1}$ e $\{\epsilon_t\}_{t=1}^{n}$ é uma seqüência de realizações de uma v.a. com $E(\epsilon_t \mid X_1, ..., X_d) = 0$ e $V(\epsilon_t \mid X_1, ..., X_d) = \sigma^2 \ \forall t$. Decorre que $E(Y \mid X_1 = x_{1t}, X_2 = x_{2t}, ..., X_d = x_{dt}) = m_1(X_1) + m_2(X_2) + ... + m_d(X_d)$ e $V(Y \mid X_1 = x_{1t}, X_2 = x_{2t}, ..., X_d = x_{dt}) = \sigma^2$. Também, assume-se que $E(\epsilon_t^4) < \infty$. Requere-se ainda que as funções $m_d(X_d), \forall d$ sejam mensuráveis (ver Buja, Hastie & Tibshirani, 1989) com $E(m_d^2(X_d)) < \infty$ pertencendo a um espaço de Hilbert H_d com produto interno dado por $E(m_d(X_d)$ $m_d(X_d))\forall d$. Defini-se $H^a = \sum_{d=1}^{D} H_d \subset H$ onde H é o espaço de funções arbitrárias quadrado integráveis e centradas de $X_1, ..., X_D$. Busca-se uma função $\widehat{g} \in H^a$ que minimiza $E(Y - g(\mathbf{X}))^2$ com $g(X_1, ..., X_d) = \sum_{d=1}^{D} m_d(X_d)$. De fato, dado que H^a é um subespaço fechado de H, $\widehat{g}$ existe e é único. Buja, Hastie & Tibshirani (1989) mostram que o ótimo $\widehat{g}(X_1, ..., X_d) = \sum_{d=1}^{D} \widehat{m}_d(X_d)$ deve satisfazer

$$\widehat{m}_j(X_j) = E\left(Y - \sum_{d=1, d\neq j}^{D} \widehat{m}_d(X_d) \mid X_j\right) \tag{11}$$

para $j = 1, ..., D$. Uma versão equivalente é obtida ao trocar em (11) as variáveis aleatórias $(Y, X_1, X_2, ..., X_d)$ por suas realizações $(y_i; x_{1i}, x_{2i}, ..., x_{di})$ e as expectativas condicionais $E(\cdot \mid X_j)$ por suavizadores S_j em x_j, isto é,

$$\widetilde{m}_j(X_j) = S_j\left(Y - \sum_{d=1, d\neq j}^{D} \widetilde{m}_d(X_d) \mid X_j\right) \tag{12}$$

onde x_j é uma $n-$upla com os componentes x_{jt} e $S_j(\cdot \mid x_j)$ é uma matriz suavizadora $n \times n$ associada com uma regressão não paramétrica univariada, onde x_j é o regressor e o resíduo parcial $R_j = y - \sum_{d=1, d\neq j}^{D} \widetilde{m}_d(x_d)$ é o regressando (veja Buja, Hastie & Tibshirani, 1989 e Martins-Filho & Bin, 1999 para mais detalhes).

Na presente dissertação o modelo utilizado tem apenas dois regressores. Sendo assim, é conveniente para os propósitos da dissertação definir os seguintes vetores: $Y = (Y_1, ..., Y_n)'$, $X = (X_1, ..., X_n)'$, $Z = (Z_1, ..., Z_n)'$, $\overrightarrow{m}_1(X) = (m_1(X_1), ..., m_1(X_n))'$, $\overrightarrow{m}_2(Z) = (m_2(Z_1), ..., m_2(Z_n))'$, $e_t^k = (0, ..., 1, ..., 0)'$ onde e_t^k é um vetor de comprimento k, e o número um aparece na t-ésima posição do vetor, e para qualquer constante c, $\overrightarrow{c}_k = (c, ..., c)'$ é um vetor de comprimento k. Denota-se por $K_d : \Re^d \to \Re$ uma função núcleo simétrica $d-$variada com $d = 1{,}2$ e por h_{1n} e h_{2n} as janelas associadas com a estimação de m_1 e m_2 respectivamente. A atenção será dada a um estimador (suavizador) polinomial (linear) local.

Antes de especificar os dois estimadores em estudo em um formato unificado, vai-se tratar, brevemente, da estimação polinomial local. Utilizando cálculo básico, pode-se mostrar que o estimador de Nadaraya-Watson, $\widehat{m}_{NW}$, é a solução natural de um problema de mínimos quadrados ponderados, sendo o minimizador $\widehat{m}$ de $\sum_{i=1}^n \{y_i - m\}^2 K\left(\frac{x_i - x}{h_n}\right)$, levando à equação normal $\sum_{i=1}^n y_i K_i = \sum_{i=1}^n K_i m$, onde $K_i = K\left(\frac{x_i - x}{h_n}\right)$. Em decorrência,

$$\widehat{m}_{NW}(x) = \left(\sum_{i=1}^n K_i\right)^{-1} \sum_{i=1}^n K_i y_i = \overline{y}. \tag{13}$$

Alternativamente, $\widehat{m}_{NW}(x)$ é o estimador de mínimos quadrados do parâmetro m_{NW} na regressão ponderada de y_i na unidade com pesos $K_i^{1/2}$ (veja Pagan & Ullah, 1999). Logo, o estimador de m_{NW}, $\widehat{m}_{NW}$ aproxima localmente $m(x)$ por uma constante. Isto sugere a utilização de polinômios locais de ordem superior, desde que aproximações locais constantes só fazem sentido, em geral, em pequena vizinhanças. Stone(1977) e Cleveland(1979) sugeriram que se minimizasse $\sum_{i=1}^n \{y_i - m - (x_i - x)m_1\}^2 K\left(\frac{x - x_i}{h_n}\right)$ com respeito a m e m_1. Esta estimativa pode ser encontrada fazendo uma regressão de mínimos quadrados ponderados de y_i contra $z_i' = (1, (x_i - x))$ com pesos $K_i^{1/2}$. Logo, enquanto o estimador de Nadaraya-Watson aproxima $m(x)$ por uma constante, esta aproximação, chamada de linear local, ajusta uma reta.

Uma vantagem do estimador linear local é que ele pode ser analisado com técnicas padrões de regressão. Ele também tem as mesmas propriedades se x_i é estocástico ou não estocástico. A lógica da regressão linear local pode ser vista expandindo $m(x_i)$ em torno de x. Obtem-se

$$m(x_i) = m(x) + \frac{\partial m}{\partial x}(x^*)(x_i - x), \tag{14}$$

onde x^* está entre os valores de x_i e x. Isto pode ser expresso por:

$$m(x_i) = m(x) + m_1(x^*)(x_i - x). \tag{15}$$

Desde que $E(y_i \mid x_i) = m(x_i)$, a função objetivo $\sum_{i=1}^{n}(y_i - m(x_i))^2 K_i = \sum_{i=1}^{n}(y_i - m(x) - m_1(x^*)(x_i - x))^2 K_i$ é essencialmente a soma de quadrados dos resíduos de uma regressão que usa somente observações próximas a $x_i = x$. Note que desta forma, $m_1(x^*)$ será muito próximo de uma constante, pois x^* está entre x_i e x. Neste caso, melhorias podem ser feitas, expandindo $m(x_i)$ por um polinômio de ordem p em $(x_i - x)$. Um estimador de regressão polinomial local (p-ésima ordem) é o minimizador de

$$\sum_{i=1}^{n} \{y_i - m - ... - m_p (x - x_i)^p\}^2 K\left(\frac{x - x_i}{h_n}\right). \tag{16}$$

Definindo X_x como a matriz de delineamento

$$\begin{pmatrix} 1 & x - x_1 & ... & (x - x_1)^p \\ . & . & & . \\ . & . & ... & . \\ . & . & & . \\ 1 & x - x_n & ... & (x - x_n)^p \end{pmatrix}$$

e fazendo $W_x = h_n^{-1} diag\left[K\left(\frac{x-x_1}{h_n}\right), ..., K\left(\frac{x-x_n}{h_n}\right)\right]$ ser a matriz de ponderação; então, se $X_x' W_x X_x$ é invertível, tem-se:

$$\widehat{m} = (X_x' W_x X_x)^{-1} X_x' W_x y. \tag{17}$$

O estimador $\widehat{m}_p(x)$ é o termo de intercepto $\widehat{m}$, isto é,$\widehat{m}_p(x) = e_1^{p+1'} (X_x' W_x X_x)^{-1} X_x' W_x y$. Mais geralmente, $\widehat{m}^{(r)}(x) = r! e_{r+1}^{p+1'} (X_x' W_x X_x)^{-1} X_x' W_x y$ é uma estimativa da r-ésima derivada[15] de $m(x)$, ou seja, $m^{(r)}(x)$.

O estimador (suavizador) linear local é acessível computacionalmente, e é bastante fácil de interpretá-lo, desde que ele generaliza um dos métodos estatísticos mais comuns; a regressão linear. Para mais detalhes consulte Simonoff(1996) e Pagan & Ullah(1999).

[15] Uma das críticas ao estimador de Nadaraya-Watson é a dificuldade de estimar as suas derivadas, o que torna mais difícil a análise e a interpretação destes componentes.

3.1 O Estimador *Backfitting*

Um procedimento conveniente para obter uma solução para (12) é o algoritmo de *Backfitting*[16] . A implementação deste algoritmo, na prática, envolve a especificação de valores iniciais $(\widetilde{m}_1^0, \widetilde{m}_2^0, ..., \widetilde{m}_d^0)$ para as funções a serem estimadas e a definição de regressandos R_j (chamados de resíduos parciais) para a m-ésima iteração do algoritmo:

$$R_j^{(m)} = y - \sum_{d=1}^{j-1} \widetilde{m}_d^{(m)}(x_d) - \sum_{d=j+1}^{D} \widetilde{m}_d^{(m-1)}(x_d) \tag{18}$$

com $j = 1, ..., d$, $m = 1, 2, ...$e $\widetilde{m}_d^m$ é um estimador de uma regressão univariada de R_j em x_j, isto é, $\widetilde{m}_d^{(m)} = E_j\left(R_j^{(m)} \mid x_j\right)$. Caso esteja-se trabalhando com uma amostra finita trocam-se as expectativas condicionais no algoritmo de *Backfitting* por suas estimativas (os suavizadores), isto é, $\widetilde{m}_d^{(m)} = S_j\left(R_j^{(m)} \mid x_j\right)$. As iterações continuam até que

$$\sum_{t=1}^{n}\left(y_t - \sum_{d=1}^{D} \widetilde{m}_d^{(m+1)}(x_d)\right)^2 - \sum_{t=1}^{n}\left(y_t - \sum_{d=1}^{D} \widetilde{m}_d^{(m)}(x_d)\right)^2$$

não mude ou seja menor que um nível de tolerância previamente especificado (veja Hastie & Tibshirani, 1986). Breiman & Friedman(1985) provaram a convergência do algoritmo, como definido, para operadores projeção em um conjunto compacto. A avaliação de $\widetilde{m}_d$ requer a escolha de um suavizador a ser utilizado assim como o grau de suavização a ser usado em cada uma das regressões, normalmente representado por um parâmetro de suavização $h_{in}, i = 1, ..., D$. Como mencionado anteriormente, o suavizador polinomial local possui propriedades teóricas e práticas desejáveis em relação a outros suavizadores, incluindo o suavizador Nadaraya-Watson.

No presente contexto ir-se-á trabalhar com um modelo aditivo bivariado ajustado por uma regressão linear local (no capítulo 5 tratar-se-á em maiores detalhes o processo de geração dos dados). Utilizando a notação introduzida anteriormente, defina duas funções ponderadoras[17] por:

$$\mathbf{s}_1(x) : \Re \to \Re^n : \mathbf{s}_1(x) = e_1^{2'}\left(\mathbf{R}_X(x)'\mathbf{V}_X(x)\mathbf{R}_X(x)\right)^{-1}\mathbf{R}_X(x)'\mathbf{V}_X(x)$$

[16] Este algoritmo é uma aplicação do conhecido algoritmo de Gauss-Seidel, veja Buja, Hastie & Tibshirani, 1989.

[17] As funções $\mathbf{s}_1(x)$ e $\mathbf{s}_2(z)$ são suavizadores lineares locais avaliados nos pontos x e z.

e

$$\mathbf{s}_2(z) : \Re \rightarrow \Re^n : \mathbf{s}_2(z) = e_1^{2'} \left(\mathbf{R}_Z(z)'\mathbf{V}_Z(z)\mathbf{R}_Z(z)\right)^{-1} \mathbf{R}_Z(z)'\mathbf{V}_Z(z), \qquad (19)$$

onde (no caso de x) $\mathbf{V}_X(x) = diag\left\{\frac{1}{h_{1n}} K_1\left(\frac{X_t - x}{h_{1n}}\right)\right\}_{t=1}^{n}$, $\mathbf{R}_X(x) = \left(\overrightarrow{1}_n, X - \overrightarrow{x}\right)$.

Faça $\mathbf{S}_1$ e $\mathbf{S}_2$ representarem as matrizes cujas linhas são os suavizadores em $\mathbf{X}$ e $\mathbf{Z}$, isto é,

$$\mathbf{S}_1 = \begin{pmatrix} \mathbf{s}_1(X_1) \\ . \\ . \\ . \\ \mathbf{s}_1(X_n) \end{pmatrix} \text{ e } \mathbf{S}_2 = \begin{pmatrix} \mathbf{s}_2(Z_1) \\ . \\ . \\ . \\ \mathbf{s}_2(Z_n) \end{pmatrix}.$$

Defina o vetor de valores estimados nos pontos $X_1, ..., X_n$ por $\widehat{\mathbf{m}} = \widehat{\mathbf{m}}_1 + \widehat{\mathbf{m}}_2$, onde $\widehat{\mathbf{m}}_1$ e $\widehat{\mathbf{m}}_2$ são as soluções do seguinte sistema de equações:

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{m}}_1 \\ \widehat{\mathbf{m}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \end{bmatrix} \mathbf{Y} \qquad (20)$$

onde $\mathbf{I}_n$ é uma matriz identidade de tamanho n e $\mathbf{S}_d^* = \left(\mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n\right)\mathbf{S}_d$, $d = 1, 2$.[18] Na prática, o sistema é resolvido usando o algoritmo de *Backfitting*, porém no caso bivariado, quando se usa o estimador linear local, o algoritmo de *Backfitting* converge para uma solução explícita para $\overrightarrow{\mathbf{m}}_1(\mathbf{X})$ e $\overrightarrow{\mathbf{m}}_2(\mathbf{Z})$ dadas por

$$\overrightarrow{\mathbf{m}}_1^b(\mathbf{X}) = \left(\mathbf{I}_n - \left(\mathbf{I}_n - \mathbf{S}_1^*\mathbf{S}_2^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_1^*\right)\right)\mathbf{Y}$$

e

$$\overrightarrow{\mathbf{m}}_2^b(\mathbf{Z}) = \left(\mathbf{I}_n - \left(\mathbf{I}_n - \mathbf{S}_2^*\mathbf{S}_1^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_2^*\right)\right)\mathbf{Y} \qquad (21)$$

se as inversas existirem. Existência da solução do algoritmo de *Backfitting*, assim como as suas propriedades estocásticas ainda são em geral desconhecidas, no entanto quando se usa o estimador linear local, Opsomer & Ruppert(1997, 1998) derivaram uma série de resultados (para grandes amostras) que reproduzir-se-á 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, $\mu_j(K) = \int u^j K(u)du = 0$ para todo j ímpar e $\mu_2(K) \neq 0$.

A2: As densidades $f(x, z)$, $f_X(x)$ e $f_Z(z)$ são limitadas, contínuas e têm suporte compacto, e suas primeiras derivadas têm um número finito de mudanças

[18] Este ajuste dos suavizadores é necessário para assegurar unicidade das soluções (se elas existirem), veja Hastie & Tibshirani(1990).

de sinal sobre o seu suporte. Além disso, $f_X(x) > 0$ e $f_Z(z) > 0$ para todo (x, z) $\in$ suporte(f) e

$$\sup \left| \frac{f(x,z)}{f_X(x) f_Z(z)} - 1 \right| < 1.$$

A3: Quando $n \to \infty$, h_{1n}, $h_{2n} \to 0$ e $nh_{1n} \mid \log(n)$, $nh_{2n} \mid \log(n) \to \infty$.
A4: As segundas derivadas de m_1 e m_2 existem e são limitadas e contínuas.

3.2 O Estimador de Integração Marginal

Para definir o estimador MI, considere $\widehat{m}(x, z; h_{1n}, h_{2n})$ uma função real definida por $\widehat{m}(x, z; h_{1n}, h_{2n}) = e_1^{3'} \left(\mathbf{X}(x,z)' \mathbf{W}(x,z) \mathbf{X}(x,z)\right)^{-1} \mathbf{W}(x,z) \mathbf{Y}$, onde $\mathbf{X}(x,z) = \left(\overrightarrow{1}_n, X - \overrightarrow{x}, Z - \overrightarrow{z}\right)$ e

$$\mathbf{W}(x,z) = diag \left\{ \frac{1}{h_{1n} h_{2n}} K_2 \left(\frac{1}{h_{1n}} (X_t - x), \frac{1}{h_{2n}} (Z_t - z) \right) \right\}_{t=1}^{n}. \qquad (22)$$

Seja $\left\{(x,z)' \in \Re^2 : x = X_t, z = Z_\tau, \text{ com } t,\tau = 1, 2, ..., n\right\}$ e defina a matriz

$$\widehat{\mathbf{m}}(\mathbf{X}, \mathbf{Z}) = \begin{pmatrix} \widehat{m}(X_1, Z_1) & \widehat{m}(X_1, Z_2) & \cdots & \widehat{m}(X_1, Z_n) \\ \widehat{m}(X_2, Z_1) & \widehat{m}(X_2, Z_2) & \cdots & \widehat{m}(X_2, Z_n) \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{m}(X_n, Z_1) & \widehat{m}(X_n, Z_2) & \cdots & \widehat{m}(X_n, Z_n) \end{pmatrix}.$$

O estimador MI para $\overrightarrow{\mathbf{m}}_1(\mathbf{X})$ e $\overrightarrow{\mathbf{m}}_2(\mathbf{Z})$, utilizando a função identidade como função de ligação e sem considerar um intercepto (veja Linton & Nielsen, 1995), é dado respectivamente por, $\overrightarrow{\mathbf{m}}_1^{mi}(\mathbf{X}) = \frac{1}{n} \widehat{\mathbf{m}}(\mathbf{X}, \mathbf{Z}) \overrightarrow{1}_n$, e $\overrightarrow{\mathbf{m}}_2^{mi}(\mathbf{Z}) = \frac{1}{n} \widehat{\mathbf{m}}(\mathbf{X}, \mathbf{Z})' \overrightarrow{1}_n$. As funções de ponderação Q_1 e Q_2 (veja Linton & Nielsen, 1995) utilizadas na estimação foram as funções de distribuição empíricas $F_{xn}(x)$ e $F_{zn}(z)$ que convergem em distribuição para $F_X(x)$ e $F_Z(z)$ respectivamente. As apro-ximações dadas em Linton & Nielsen(1995, p.95) continuam válidas quando se escreve as funções empíricas no lugar de Q. Em particular, quando x e z são independentes, as funções empíricas serão as funções de ponderação ótimas no sentido que elas minimizarão as variâncias das aproximações assintóticas.

As definições dadas acima tomam h_{1n} e h_{2n} como seqüências não estocásticas conhecidas que convergem a zero a uma taxa especificada. Para o estimador-B, Opsomer & Ruppert(1997) mostram que quando, $n \to \infty$, h_{1n}, $h_{2n} \to 0$ e $\frac{nh_{1n}}{\log n}$, $\frac{nh_{2n}}{\log n} \to \infty$ é possível obter-se uma aproximação assintótica para o viés condicional

e a variância condicional de $\overrightarrow{m}_1^b(X_i)$ e $\overrightarrow{m}_2^b(Z_i)$, onde $\overrightarrow{m}_1^b(X_i)$ e $\overrightarrow{m}_2^b(Z_i)$ são os i-ésimos elementos de $\overrightarrow{\mathbf{m}}_1^b(\mathbf{X})$ e $\overrightarrow{\mathbf{m}}_2^b(\mathbf{Z})$, respectivamente.[19] Para o estimador-MI, Linton & Nielsen(1995) mostram que quando h_{1n}, $h_{2n} \to 0$ e $nh_{1n}h_{2n}^2$, $nh_{2n}h_{1n}^2 \to \infty$, então $\sqrt{nh_{1n}}\left(\overrightarrow{m}_1^{mi}(X_i) - E\left(\overrightarrow{m}_1^{mi}(X_i)\right)\right)$ e $\sqrt{nh_{2n}}\left(\overrightarrow{m}_2^{mi}(Z_i) - E\left(\overrightarrow{m}_2^{mi}(Z_i)\right)\right)$ são assintoticamente normais, onde $\overrightarrow{m}_1^{mi}(X_i)$ e $\overrightarrow{m}_2^{mi}(Z_i)$ são os i-ésimos elementos de $\overrightarrow{\mathbf{m}}_1^{mi}(\mathbf{X})$ e $\overrightarrow{\mathbf{m}}_2^{mi}(\mathbf{Z})$, respectivamente.[20]

[19] A aproximação é válida sob mais três suposições (veja Opsomer & Ruppert, 1997).

[20] Pode-se demonstrar que os métodos de seleção da janela dependentes dos dados atualmente utilizados na literatura, incluindo validação cruzada e vários métodos *plug-in* não produzem seqüências $\{h_{1n}\}$ e $\{h_{2n}\}$ que convirjam a zero às taxas requeridas acima. Para a prova veja Martins-Filho(2001).

4 Métodos de seleção automática da janela h_n

Como comentou-se no capítulo 2, um dos passos mais importantes na estimação de modelos de regressão não paramétrica é a escolha dos parâmetros de suavização h_n. 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. Na dissertação consideram-se dois diferentes métodos de seleção automática da janela h_n. Os dois são variantes dos métodos *plug-in*, que se utilizam de uma otimização analítica.

Um critério apropriado (veja Ruppert & Wand, 1994 e Ruppert, Sheater & Wand, 1995) é o MISE ponderado condicional dado por (no caso de $\mathbf{X}$)

$$MISE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right) = E\int\left[\left\{\widehat{m}_p(x;h_n)-m(x)\right\}^2 \mid X_1,...,X_n\right] f_X(x)dx \tag{23}$$

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

$$MISE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right) = \left[\frac{h_n^{p+1}\mu_{p+1}\left(K_{(p)}\right)}{(p+1)!}\right]^2\int m^{(p+1)}(x)^2 f_X(x)dx + \frac{R(K_{(p)})\sigma^2\left(b-a\right)}{nh_n} + o_p\left[h_n^{2p+2}+(nh_n)^{-1}\right] \tag{24}$$

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) ésimo elemento igual a μ_{i+j+2}, $M_p(u)$ é o mesmo que N_p mas com a primeira coluna trocada por $(1,u,u^2,...,u^p)'$ e $R(K_{(p)}) = \mu_0\left(K_{(p)}^2\right)$. O minimizador de (24) é assintoticamente

$$\widetilde{h}_n = \left[\frac{(p+1)\left(p!\right)^2 R\left(K_{(p)}\right)\sigma^2\left(b-a\right)}{2n\mu_{p+1}\left(K_{(p)}\right)^2\int m^{(p+1)}(u)^2 f_X(u)du}\right]^{1|(2p+3)} \tag{25}$$

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

$$MASE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\left(\widehat{m}_p\left(x_i;h_n\right)-m(x)\right)|X_1,...,X_n\right\}^2. \tag{26}$$

Note que (26) é uma aproximação discreta a (23).

O princípio básico dos métodos *plug-in* é a substituição direta de estimativas de σ^2 e dos funcionais que aparecem nas expressões que descrevem os parâmetros de suavização h_n, após ter sido minimizado o critério a ser utilizado na estimação não paramétrica.

O método *plug-in* proposto em Linton & Nielsen(1995) é baseado na seguinte "regra de bolso"(ROT):

$$h_{inROT} = \left\{ \frac{\widetilde{\sigma}^2 R(K_{(1)})(b_i - a_i)}{\mu_2(K_{(1)})^2 \left(\widehat{\theta}_1 + \widehat{\theta}_2\right)^2} \right\}^{1/5} n^{-1/5}, \tag{27}$$

onde $i = 1, 2$, b_i e a_i denotam o máximo e o mínimo amostral do regressor de interesse, $\widehat{\theta}_1$ e $\widehat{\theta}_2$ são os coeficientes de $x^2/2$ e $z^2/2$ obtidos de uma regressão de mínimos quadrados ordinários de y em uma constante, x, z, $x^2/2$, $z^2/2$ e xz e $\widetilde{\sigma}^2$ é obtido dos resíduos desta regressão. Esta regra é ótima com respeito ao critério AMISE (veja 25), quando $p = 1$, x e z são independentes e o modelo de regressão bivariado $m(x, z)$ é uma função quadrática. $\widehat{\theta}_1$ e $\widehat{\theta}_2$ são simplesmente aproximações para a derivada segunda que aparecerá em (25) quando $p = 1$.

Outro método *plug-in* que se utilizou foi proposto em Opsomer & Ruppert (1998). Está-se interessado em escolher $h_{1n}, h_{2n} \in \Re$ tal que

$$MASE(h_{1n}, h_{2n} \mid \mathbf{X}, \mathbf{Z}) = \frac{1}{n}\sum_{t=1}^{n} \left\{ E\left(\sum_{d=1}^{n} \widehat{m}(X_i, Z_i) - m(X_i, Z_i) \mid \mathbf{X}, \mathbf{Z} \right)^2 \right\}. \tag{28}$$

Do corolário 4.2 de Opsomer & Ruppert(1997), a aproximação assintótica para o $MASE$ condicional dado em (28), quando o modelo aditivo é ajustado por regressão linear local $(p = 1)$, denotada por $AMASE$, é dada por:

$$AMASE(h_{1n}, h_{2n} \mid \mathbf{X}, \mathbf{Z}) = \frac{\mu_2(K_{(1)})^2}{4} \left(h_{1n}^4 \theta_{11} + h_{1n}^2 h_{2n}^2 \theta_{12} + h_{2n}^4 \theta_{22} \right) + \sigma^2 R(K_{(1)}) \left(\frac{b_x - a_x}{n h_{1n}} + \frac{b_z - a_z}{n h_{2n}} \right) \tag{29}$$

onde

$$\theta_{11} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i' D^2 \mathbf{m}_1 + \mathbf{v}_i' E\left(m_1^{(2)}(X_i) \mid \mathbf{Z}\right)\right)^2,$$

$$\theta_{22} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_i' D^2 \mathbf{m}_2 + \mathbf{t}_i' E\left(m_2^{(2)}(Z_i) \mid \mathbf{X}\right)\right)^2,$$

$$\theta_{12} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i' D^2 \mathbf{m}_1 + \mathbf{v}_i' E\left(m_1^{(2)}(X_i) \mid \mathbf{Z}\right)\right)\left(\mathbf{v}_i' D^2 \mathbf{m}_2 + \mathbf{t}_i' E\left(m_2^{(2)}(Z_i) \mid \mathbf{X}\right)\right)^2$$

e $\mathbf{t}_i'$ e $\mathbf{v}_j$ representam a i-ésima linha e a j-ésima coluna de $(\mathbf{I} - \mathbf{T}_{12}^*)^{-1}$, desde que a inversa exista e $[\mathbf{T}_{12}^*]_{ij} = \frac{1}{n}\frac{f_{XZ}(X_i,Z_j)}{f_X(X_i)f_Z(Z_j)} - \frac{1}{n}$.

Denotando os valores das janelas que minimizam $AMASE$ por $h_{1nAMASE}$ e $h_{2nAMASE}$ e supondo $\mathbf{X}$ e $\mathbf{Z}$ independentes, pode-se escrever

$$h_{1nAMASE} = \left(\frac{R(K_{(1)})\sigma^2\,(b_x - a_x)}{n\mu_2(K_{(1)})^2\theta_{11}}\right)^{1|5}$$

e

$$h_{2nAMASE} = \left(\frac{R(K_{(1)})\sigma^2\,(b_z - a_z)}{n\mu_2(K_{(1)})^2\theta_{22}}\right)^{1|5}. \tag{30}$$

A estratégia de estimação utilizada envolve obter estimativas para σ^2 e θ_{ii}, $i = 1, 2$ e substituí-las diretamente em (30). A regra *plug-in* (PI) que utilizou-se foi a seguinte: σ^2 foi estimada através de $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{m}_1^b(X_i) - \widehat{m}_2^b(Z_i)\right)^2$ onde $\widehat{m}_1^b(X_i)$ e $\widehat{m}_2^b(Z_i)$ são as soluções do algoritmo de *Backfitting* dadas em (21) e $\widehat{\theta}_{11}$ e $\widehat{\theta}_{22}$ foram estimadas pela equação (10) de Opsomer & Ruppert(1998), isto é, $\widehat{\theta}_{11} = \frac{1}{n}Tr\mathbf{V}_1^{(2)*}\mathbf{Y}\mathbf{Y}^T\mathbf{V}_1^{(2)*\prime}$ e $\widehat{\theta}_{22} = \frac{1}{n}Tr\mathbf{V}_2^{(2)*}\mathbf{Y}\mathbf{Y}'\mathbf{V}_2^{(2)*\prime}$ onde

$$\begin{aligned}\mathbf{V}_1^{(2)} &= \mathbf{S}_1^{(2)}\left(\mathbf{I}_n - S_2^*S_1^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_2^*\right),\ \mathbf{V}_2^{(2)} = \mathbf{S}_2^{(2)}\left(\mathbf{I}_n - \mathbf{S}_1^*\mathbf{S}_2^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_1^*\right),\\ \mathbf{V}_1^{(2)*} &= \left(\mathbf{I} - \mathbf{1}\mathbf{1}'/n\right)\mathbf{V}_1^{(2)},\ \mathbf{V}_2^{(2)*} = \left(\mathbf{I} - \mathbf{1}\mathbf{1}'/n\right)\mathbf{V}_2^{(2)}\end{aligned} \tag{31}$$

e $\mathbf{S}_1^{(2)}$ e $\mathbf{S}_2^{(2)}$ representam as matrizes cujas linhas podem ser escritas por $\left(\mathbf{s}_{1,x}^{(2)}\right)' = 2!e_3^{4\prime}\left(\mathbf{R}_X(x)'\mathbf{V}_X(x)\mathbf{R}_X(x)\right)^{-1}\mathbf{R}_X(x)'\mathbf{V}_X(x)$ e $\left(\mathbf{s}_{2,z}^{(2)}\right)' = 2!e_3^{4\prime}\left(\mathbf{R}_Z(z)'\mathbf{V}_Z(z)\mathbf{R}_Z(z)\right)^{-1}\mathbf{R}_Z(z)'\mathbf{V}_Z(z)$. Usou-se a "regra de bolso"(ROT) descrita na página anterior para estimar as matrizes $\mathbf{V}_X(x)$ e $\mathbf{V}_Z(z)$ que aparecem em $\left(\mathbf{s}_{1,x}^{(2)}\right)'$ e $\left(\mathbf{s}_{2,z}^{(2)}\right)'$.

5 O Processo de Geração dos Dados

Os dados utilizados no estudo foram gerados por um modelo de regressão não paramétrica aditiva bivariada ajustada por regressão linear local com a correlação variando para avaliar a robustez a desvios de independência entre os regressores. Assume-se que $\left\{(y_t, x_t, z_t)'\right\}_{t=1}^{n}$ formam uma seqüência de realizações de um vetor aleatório $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ avaliado em $\Re^3$e $\{\epsilon_t\}_{t=1}^{n}$ é uma seqüência de realizações de uma v.a. com distribuição $N(0,1)$. O modelo utilizado é descrito por

$$Y_t = m_1(X_t) + m_2(Z_t) + \epsilon_t \tag{32}$$

onde $m_1(X_t) = -6X_t + 36X_t^2 - 53X_t^3 + 22X_t^5$, $m_2(Z_t) = sen(Z_t)$, $X_t = S_t$, $Z_t = 5\pi W_t$, com $\{W_t, S_t\}_{t=1}^{n}$ sendo gerados por uma função densidade conjunta com a correlação desejada, com marginais dadas por uma distribuição $N(1/2, 1/9)$. Utilizou-se três níveis de correlação para investigar a robustez: 0 (independência), .25 (correlação "baixa"), .75 (correlação "alta").

A existência de uma solução do algoritmo de *Backfitting* é, de maneira geral, desconhecida, mas no caso em que se utilizam estimadores de mínimos quadrados linear local, Opsomer & Ruppert(1997, 1998) derivaram uma série de condições suficientes que garantem a existência de uma única solução no caso bivariado (veja na página 21).

Devido a A2 rejeitaram-se todas as observações para o qual um dos regressores ultrapassou $\pm 1.5\sigma$ da média (ou equivalentemente fora do intervalo [0,1]) e neste caso trocaram-se por novas observações que estivessem fora destes limites. Consideraram-se amostras de tamanho 100, 150 e 200, cada uma replicada 800, 600 e 400 vezes, respectivamente.

A função núcleo utilizada no estudo foi a Gaussiana. Alguns resultados importantes dentro do contexto são dados a seguir. Para a função núcleo Gaussiana obtém-se: $\mu_1(K) = 0$, $\mu_2(K) = 1$ e $R(K_{(1)}) = (2\sqrt{\pi})^{-1}$.

6 Resultados

Conduziu-se um estudo de simulação para avaliar e comparar a performance dos estimadores-B e MI em amostras finitas para uma regressão aditiva bivariada. Tal estudo é necessariamente restritivo, pois existem muitas possibilidades para a escolha da função de regressão, da densidade dos regressores, da correlação entre eles, da densidade do erro, do tamanho da amostra, do tipo de regressão polinomial, da função núcleo, da janela escolhida, do tipo de função critério erro quadrático utilizado, etc...

Olhando-se a figura 2 de Opsomer & Ruppert(1997, p.191) pode-se notar que a correlação 0.75 está fora dos limites prescritos pela suposição A2 do artigo referido (p.190), no caso em que se usa uma distribuição Normal bivariada. Aparentemente isto não afeta a convergência. Isto suporta a idéia de que a correlação dentro dos limites pedidos por A2, embora suficiente, não é uma condição necessária para a convergência dos estimadores *Backfitting*. A função núcleo utilizada também não atende a condição A1 descrita na página 21. Da mesma forma, aparentemente, isto não afeta a aplicabilidade dos resultados derivados por Opsomer & Ruppert(1997).

O interesse primário da dissertação é comparar a performance dos estimadores-B e MI em amostras finitas. Para isto, computou-se o erro quadrático médio $ASE = \frac{1}{n}\sum_{t=1}^{n}\left(\widehat{m}_1(X_t) + \widehat{m}_2(Z_t) - m_1(X_t) - m_2\left(Z_t\right)\right)^2$ nos estudos de simulação. Em seguida calculou-se a média das replicações para estimar-se o $MASE$. Comparando-se os valores das Tabelas 1 e 2 pode-se notar que os estimadores-B apresentaram uma performance melhor do que os estimadores-MI.

Tabela 1. Estimativas de $MASE$ usando *Backfitting* com as janelas PI e as verdadeiras $AMASE$

	h_{PI}			h_{AMASE}		
	$n = 100$[21]	$n = 150$	$n = 200$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	0.606	0.470	0.397	0.327	0.284	0.261
$\rho = 0.25$	0.600	0.479	0.395	0.323	0.288	0.261
$\rho = 0.75$	0.596	0.470	0.399	0.321	0.279	0.260

[21] A janela h_{2nPI} apresentou problemas de *overflow* no estudo de simulação feito. O processo de geração dos dados foi repetido uma vez quando $\rho = 0$ e $\rho = 0.75$ e duas vezes quando $\rho = 0.25$.

Tabela 2. Estimativas de $MASE$ usando Integração Marginal com as janelas ROT e as verdadeiras $AMASE$

	h_{ROT}			h_{AMASE}		
	$n = 100$	$n = 150$	$n = 200$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	1.799	0.802	0.645	0.587	0.500	0.367
$\rho = 0.25$	5.133	1.036	3.173	0.694	0.460	0.385
$\rho = 0.75$	14.967	5.297	12.028	3.385	2.228	2.525

Analisando as Tabelas 1 e 2 é possível observar uma série de fatos importantes. Primeiramente, note que a quebra da suposição de independência entre os regressores não afeta a estimação feita via o algoritmo de *Backfitting*, seja a correlação entre os regressores baixa ($\rho = 0.25$) ou alta ($\rho = 0.75$). O mesmo não acontece quando se utiliza Integração Marginal. Neste caso, o impacto de ignorar a dependência altera em muito os resultados obtidos.

Note também que as janelas utilizadas neste estudo de Monte Carlo são escolhidas de modo a minimizar $MASE$. Desta forma a comparação entre os estimadores deve ser feita via o critério $MASE$. Cabe, no entanto, comentar que se for utilizado a mediana das replicações para comparar os estimadores os resultados apresentam diferenças perceptíveis. Os resultados obtidos eram de certa forma esperados. Opsomer & Ruppert(1997, p.198) comentam que uma diferença interessante entre os dois estimadores ocorre quando $\mathbf{X}$ e $\mathbf{Z}$ são independentes. Nesse caso é natural esperar que o viés assintótico de estimadores de um modelo aditivo para estimar uma das funções componentes não dependa do comportamento da outra função. Opsomer & Ruppert(1997) demonstram que o estimador-B tem essa propriedade, enquanto que o estimador MI não. A não ser que os efeitos de viés das funções componentes cancelem-se entre si, isto provavelmente resultará em um viés maior do estimador-MI em relação ao estimador-B. A comparação entre as variâncias assintóticas é mais direta, devido ao formato semelhante das expressões para os dois estimadores. Neste caso, pode-se demonstrar que a variância assintótica dos estimadores-B é sempre menor do que a dos estimadores-MI, a não ser que $\mathbf{X}$ e $\mathbf{Z}$ sejam independentes.

A comparação entre os dois estimadores fica mais clara quando se utilizam as verdadeiras janelas $h_{1nAMASE}$ e $h_{2nAMASE}$. Em um estudo de simulação como este nada é desconhecido em (30), ou seja, não existirá nenhum "ruído" inerente ao processo de estimação quando se comparar os dois estimadores. Nesse caso, pode-se notar que existem fortes indícios da superioriedade dos estimadores-B.

Na tentativa de clarificar a superioriedade dos estimadores-B calculou-se o $MASE$ destes estimadores usando as janelas h_{inROT}, $i = 1, 2$, diretamente. Estas janelas foram construídas em um formato apropriado para a estimação via Integração Marginal. Suspeitava-se que mesmo usando uma regra apropriada

para estimar via estimadores-MI a performance dos estimadores-B ainda seria superior, o que realmente ficou evidenciado. Porém, um fato surpreendente aconteceu, como se pode reparar olhando a Tabela 3 comparativamente a Tabela 1 . Aparentemente, a estimação da segunda derivada feita em Opsomer & Ruppert(1998) em vez de melhorar a performance dos estimadores-B a deteriora em pequenas amostras. Apesar de inesperado, o resultado é interessante, desde que pouco se sabe sobre as propriedades em amostras finitas deste estimador.

Tabela 3. Estimativas de $MASE$ usando *Backfitting* com a janela ROT.

	h_{nROT}		
	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	0.440	0.376	0.334
$\rho = 0.25$	0.438	0.365	0.329
$\rho = 0.75$	0.438	0.364	0.329

As figuras 1 e 2 mostram as densidades[22] de $\log(h_{inAMASE}) - \log(h_{inPI})$ e $\log(h_{inAMASE}) - \log(h_{inROT})$, $i = 1, 2$ para os níveis de correlação utilizados e as amostras de tamanho 100, 150 e 200 cada uma replicada 800, 600 e 400 vezes, respectivamente. Como se pode notar as densidades para os diferentes níveis de correlação estão bastante próximas. Aparentemente o grau de correlação entre as covariáveis tem pouco efeito nas janelas estimadas, o que justifica o uso de independência no cômputo de h_{nPI} e h_{nROT}. O estimador h_{1nPI} apresenta um pequeno viés (*undersmoothing)* na estimação de m_1 (polinômio de baixo grau) enquanto o estimador h_{1nROT} apresenta um viés mais acentuado, causando um *oversmoothing* na estimação de m_1. Neste caso, os estimadores apresentam uma variabilidade similar. Os dois estimadores apresentam um grande viés na estimação de m_2 (*undersmoothing*)[23], porém a magnitude do viés dos estimadores h_{2nPI} é maior, além de serem mais variáveis. Os estimadores h_{nROT} apresentam uma variabilidade similar na estimação de m_1 e m_2[24].

[22] Estimadas através da janela de Sheather-Jones(1991).

[23] Isto provavelmente deve-se ao fato de m_2 ser uma função seno (portanto menos a afeita a aproximações de primeira ordem do que m_1).

[24] Note, ainda, que a performance de ambos os estimadores melhora quando o tamanho da amostra aumenta.

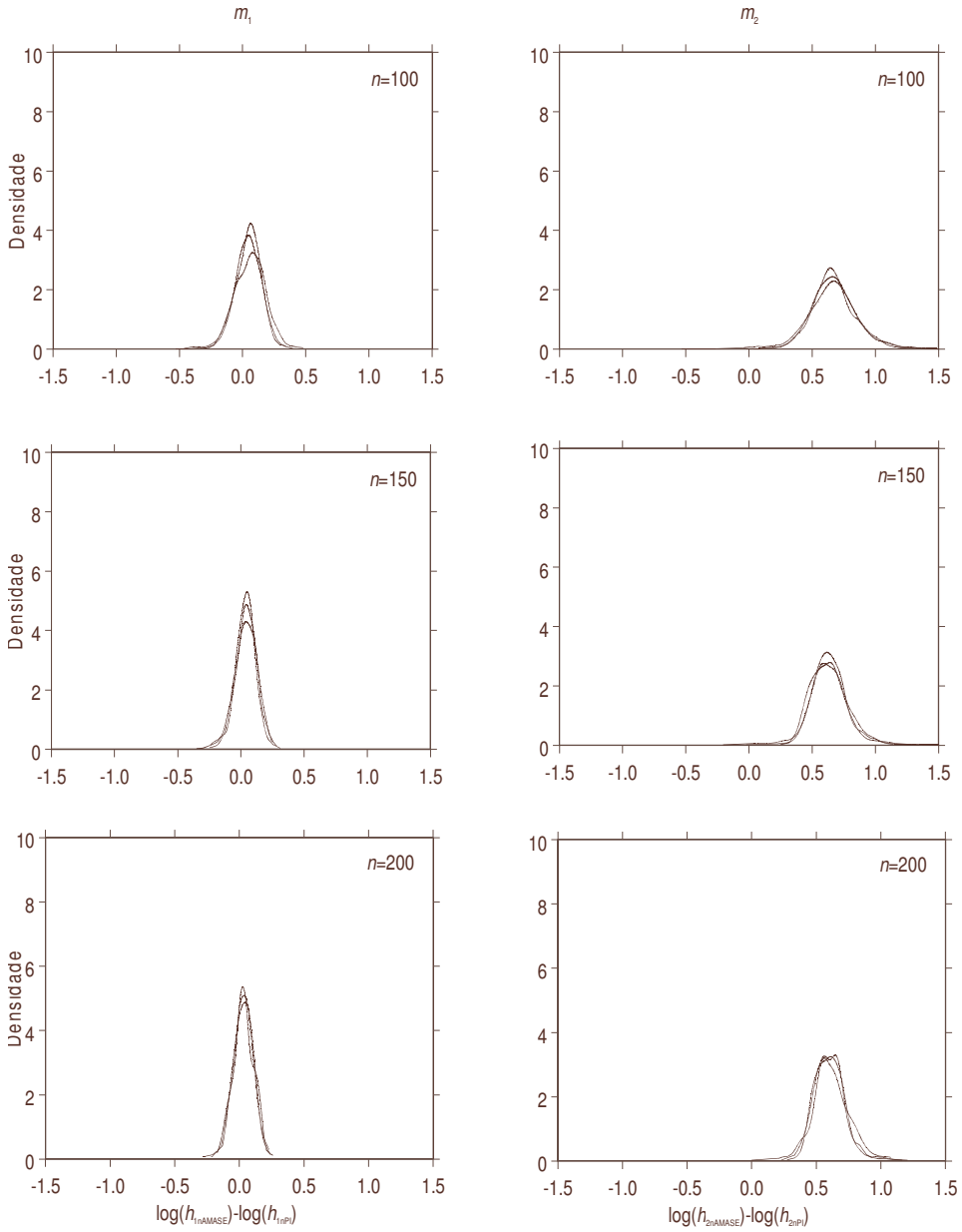

Figura 1: Densidades dos Estimadores *PI* para os três níveis de correlação.

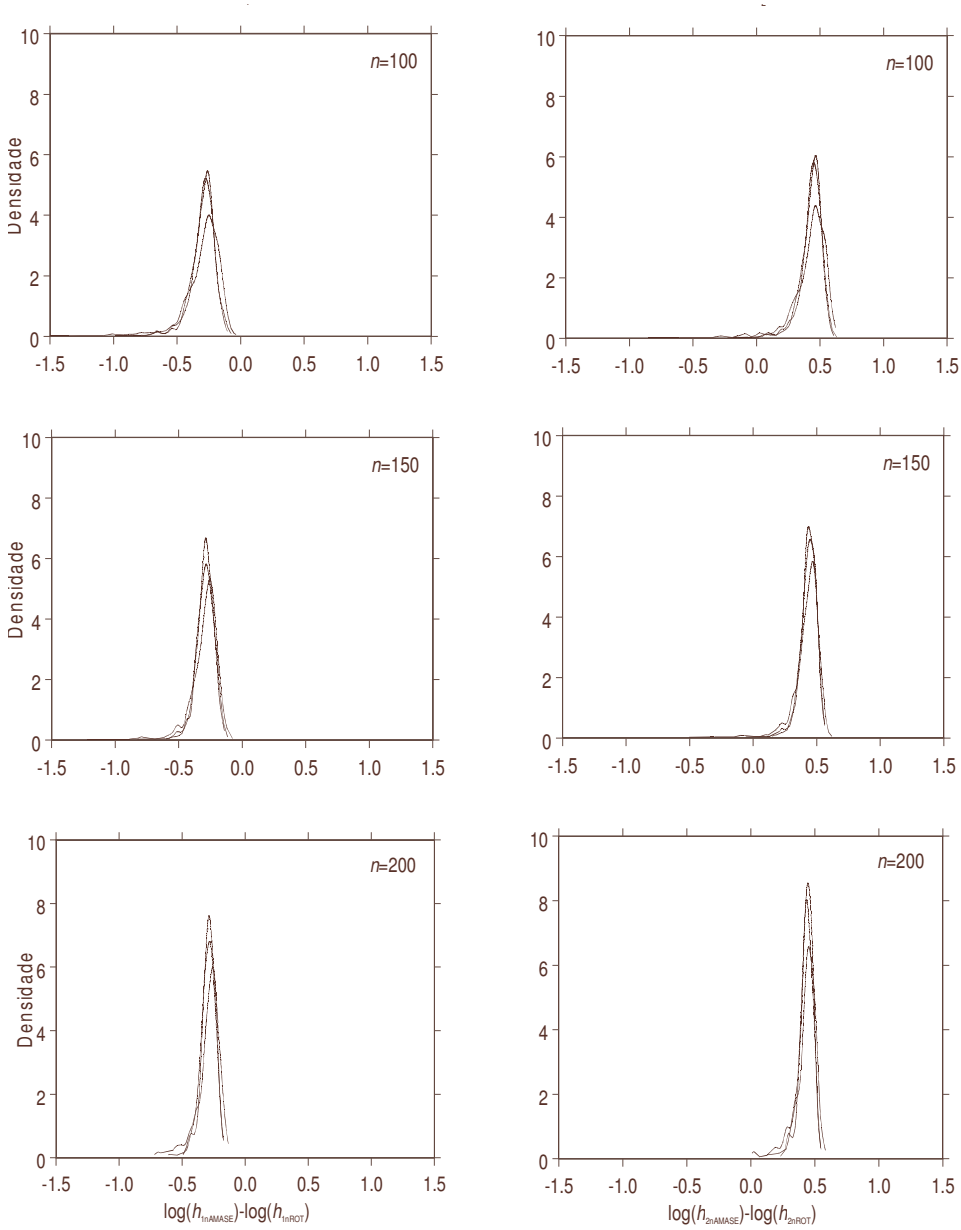

Figura 2: Densidades dos Estimadores ROT para os três níveis de correlação.

7 Conclusões

A literatura atual propõe dois métodos para a estimação de uma regressão não paramétrica aditiva: estimadores-B e MI. A comparação feita via o estudo de Monte Carlo sugere que o estimador-B tem uma performance superior ao estimador-MI. Embora o estudo de simulação apresentado aqui seja um pouco pequeno, isto é confirmado em um estudo mais extensivo, ver Martins-Filho(2001).[25]

O estimador de Linton & Nielsen é baseado em uma ótima idéia, porém envolve o produto das janelas. No caso bivariado, se as estimativas das duas janelas forem subsuavizadas ou sobresuavizadas o efeito será magnificado. Além disso, como comentado na página 16 a estimação via Integração Marginal demanda mais computacionalmente[26], o que é inconveniente para usuários. De fato, o estimador-MI apresenta problemas associados a regressões multivariadas irrestritas o que é indesejável.

O objetivo principal da dissertação era comparar os dois métodos atuais para a estimação de uma regressão não paramétrica aditiva. Os principais pontos encontrados estão resumidos abaixo.

1. A quebra da suposição de independência entre os regressores não afeta a estimação feita via o algoritmo de *Backfitting*. O mesmo não ocorre quando se utiliza Integração Marginal.
2. Uma diferença interessante entre o estimador-B e o estimador-MI acontece quando os regressores $\mathbf{X}$ e $\mathbf{Z}$ são independentes. Neste caso, espera-se que o viés assintótico de estimadores de um modelo aditivo quando se deseja estimar um dos componentes aditivos não dependa do comportamento do outro componente. O estimador-B tem esta propriedade, enquanto que o estimador-MI não a possui. Em virtude disto, em geral, o estimador-MI apresentará um viés superior em relação ao estimador-B.

[25] Além das janelas utilizadas nesta dissertação, usam-se validação cruzada e a janela h_{DPI} (Opsomer & Ruppert 1998). Esta última é baseada em um método *plug-in* que aparentemente apresenta um desempenho superior aos estimadores utilizados aqui.

[26] A diferença nos estudos de simulação feitos foi considerável. Para amostras de tamanho 100 uma replicação usando Integração Marginal demorou em média 33,7s. Usando *Backfitting* com janela *PI* durou em média 2,1s e usando a janela *ROT* 1,2s. O equipamento utilizado foi um Pentium III 500Mhz. Os programas foram feitos em Gauss versão 3.5 e serão disponibilizados pelo autor sob pedido.

3. A variância assintótica dos estimadores-B é sempre menor do que a dos estimadores-MI, a não ser que os regressores sejam independentes.

4. O estimador-MI, em geral, necessita computar um número maior de operações para estimar os componentes aditivos do que o estimador-B, isto é, a demanda computacional do estimador-MI é superior a do estimador-B.

5. O estimador-MI, no caso bivariado, envolve o produto de 2 janelas. Se as estimativas das janelas forem subsuavizadas ou sobresuavizadas o efeito será magnificado. Isto age de maneira semelhante a maldição da dimensionalidade.

Referências

[1] Breiman, L. and Friedman, J.(1985), "Estimating Optimal Transformations for Multiple Regression and Correlation", *Journal of the American Statistical Association*, **80,** 580-619.

[2] Brown, L.D., Cai, T., Low, M.G. and Zhang, C.(2002), "Asymptotic Equivalence Theory for Nonparametric Regression with Random Design", *Annals of Statistics*, in press.

[3] Buja, A., Hastie, T. and Tibshirani, R.(1989), "Linear Smoothers and Additive Models(with discussion)", *Annals of Statistics,* **17,** 453-555.

[4] Cacoullos,T. (1996), "Estimation of a Multivariate Density", *Annals of the Institute of Statistical Mathematics,* **18,** 178-189.

[5] Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association,* **74,** 829-836.

[6] Draper, N.R., and Smith, H. (1998), *Applied Regression Analysis*, 3rd Edition, New York: John Wiley & Sons.

[7] Epanechnikov, V.A.(1969), "Nonparametric Estimates of a Multivariate Probability Density", *Theory of Probability and Applications*, **14,** 153-158.

[8] Fan, J. (1992), "Design Adaptive Nonparametric Regression", *Journal of the American Statistical Association,* **87**, 998-1004.

[9] Fan, J., Gasser, T., Gijbels, I., Brockmann, M., Engel, J. (1993), "Local Polynomial Fitting: a Standard for Nonparametric Regression", *Departament of Statistics, UNC.*

[10] Fix, E. and Hodges, J. L. (1951), "Discriminatory analysis. nonparametric discrimination: Consistency properties", Tech. Rep. 4 Project number 2149 -004, USAF School of Aviation Medicine, Randolph Field, Tx.

[11] Friedman, J.H. and Stuetzle, W.(1981), "Projection Pursuit Regression", *Journal of the American Statistical Association,* **76**, 817-823.

[12] Friedman, J.H. and Tukey, J.W.(1974), "A Projection Pursuit Algorithm for Explaratory Data Analysis", IEEE Transactions on Computers, **C-23**, 881-890.

[13] Gozalo, P. L. and Linton, O. B. (2001), "Testing Additivity in Generalized Nonparametric Regression Models with Estimated Parameters", *Journal of Econometrics*, **104,** 1-48.

[14] Greene, W.H. (1999), *Econometric Analysis*, 4th Edition, New York: Prentice-Hall.

[15] Härdle, W. (1990), *Applied Nonparametric Regression*, New York: Cambridge University Press.

[16] Härdle, W., Hall, P. and Marron, J. S.(1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?", *Journal of the American Statistical Association*, **83**, 86-95.

[17] Hastie, T. J. and Tibshirani, R. J.(1986), "Generalized Additive Models", *Statistical Science*, **1**, 297-318.

[18] Hastie, T. J., and R. J., Tbishirani (1990), *Generalized Additive Models*. London: Chapman and Hall.

[19] Ibragimov, I. A. and Hasminskii, R. Z.(1983), "Estimation of Distribution Density", *J. Sov. Math.*, **25**, 40-57.

[20] Johnston, J. and Dinardo, J. (1996), *Econometric Methods*, 4th Edition, Irvine: McGraw-Hill.

[21] Jones, M. C., Marron, J. S. and Sheater, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association,* **91,** 401-407.

[22] 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,* **8**, 2, 279-297.

[23] Kuan, C.M. and White, H.(1994), "Artificial Neural Networks: An Econometric Perspective", *Econometric Reviews,* **13**, 1-91.

[24] Linton, O. and Nielsen, J.P.(1995), "A Kernel Method of Estimating Structured Nonparametric Regression based on Marginal Integration", *Biometrika*, **82**, 1, 93-100.

[25] Linton, O. and Härdle, W.(1996), "Estimation of additive regression models with known links.ructured Nonparametric Regression based on Marginal Integration", *Biometrika*, **83**, 3, 529-540.

[26] 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.

[27] Martins-Filho, C.B. (2001), "Additive Nonparametric Regression Estimator via *Backfitting* and Marginal Integration: Finite Sample Performance", *Working paper*, OSU, Departament of Economics.

[28] Martins-Filho, C.B. and Bin, O. (1999), "Estimation of Hedonic Price Functions via Additive Nonparametric Regression, *Working paper,* OSU, Department of Economics.

[29] Nadaraya, E.A.(1964), "On estimating regression", *Theor. Prob. Appl.* **9**, 141-142.

[30] Opsomer, J.D. and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polinomial Regression", *Annals of Statistics.*, **25**, 1, 186-211.

[31] Opsomer, J.D. and Ruppert, D.(1998),."A Fully Automated Bandwidth Selection Method for Fitting Additive Models", *Journal of the American Statistical Association,* **93**, 442, 605-619.

[32] Pagan, A. and Ullah, A. (1999), *Nonparametric Econometrics,* Cambridge University Press.

[33] Rosenblatt, M.(1956), "Remarks on some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics,* **27,** 832-837.

[34] Ruppert, D., Sheather, S. J. and Wand, M. P.(1995), "An Effective Bandwidth Selector for Local Least Squares Regression", *Journal of the American Statistical Association*, **25**, 186-211.

[35] Ruppert, D., and Wand, M. P.(1994), "Multivariate Locally Weighted Least Squares Regression", *Annals of Statistics*, **22**, 1346-1370.

[36] Sheater, S. J. and Jones, M. C. (1991), "A Reliable Data-Based Bandwitdh Selection Method for Kernel Density Estimation", *Journal of the Royal. Statistical Society*, Series B, **53**, 3, 683–690.

[37] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[38] Simonoff, J.S. (1996), "*Smoothing Methods in Statistics",* New York: Springer-Verlag.

[39] Stone, C.J. (1977), "Consistent Nonparametric Regression. *Annals of Statistics",* **5**, 595-645.

[40] Stone, C. J. (1980), "Optimal rates of convergence for nonparametric estimators", *Annals of Statistics,* **8**, 1348-1360.

[41] Stone, C.J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression", *Annals of Statistics,* **10**, 1040-1053.

[42] Stone, C.J. (1985), "Additive Regression and Other Nonparametric Models", *Annals of Statistics,* **13**, 689-705.

[43] Stone, C.J. (1986), "The Dimension Reduction Principle for Generalized Additive Models", *Annals of Statistics,* **14**, 590-606.

[44] Watson, G. S. (1964), "Smooth Regression Analysis", *Sankhya*, Series A, **26**, 359-372.