

5.

Métodos de Estimação de Densidade

As previsões probabilísticas para geração de energia eólica são um desenvolvimento relativamente recente. Este tipo de previsão possui dois tipos principais de abordagens: uma abordagem relacionada com o *erro de previsão*, isto significa que fornece previsões probabilísticas dos erros de um modelo de previsão determinístico existente. A outra abordagem é chamada *abordagem direta*, a qual fornece diretamente as previsões probabilísticas da variável considerada.

A abordagem do *erro de previsão* acrescenta uma estimativa da incerteza para os sistemas de previsão determinísticos existentes. Abordagens anteriores usavam critérios de avaliação globais (como o desvio padrão dos erros de previsão calculados ao longo de várias corridas), como avaliação da incerteza da previsão. No entanto, este fornece valores constantes para um determinado período de tempo.

Por outra parte, diversas abordagens diretas de previsão probabilística têm sido propostas como foi expresso na Seção 3.1. Por exemplo, a Regressão Quantílica local é usada em Bremnes (2004) para calcular quantis específicos da produção de energia. Uma comparação de três abordagens quantílicas, chamadas Regressão Quantílica Local, modelagem Gaussiana local e o estimador de *Nadaya-Watson* foi efetuada por Bremnes (2006). Similarmente, em Juban et al. (2007b), realiza-se uma comparação dos métodos de Estimação da Densidade Condicional por *kernel*, Regressão Quantílica Forests e Regressão Quantílica B-Spline. No entanto, o foco desta tese se centra nas previsões que tem como saída uma função de densidade de probabilidade.

5.1.

Estimação de Densidade de Probabilidade

Seja X uma variável aleatória com função de densidade de probabilidade f , isto significa que para quaisquer valores a e b , a probabilidade de que X esteja entre a e b está dada por:

$$P(a < X < b) = \int_a^b f(x)dx \quad \text{para qualquer } a < b, \quad (5.1)$$

onde a função f possui as seguintes propriedades:

- Toma só valores positivos
- É integrável
- A integral de $f(x)$ desde $-\infty$ a $+\infty$ é igual a 1.

Quando se tem uma amostra de dados, a qual se supõe vem de uma função de densidade desconhecida, o cálculo da função de densidade estimada será de interesse só para essa amostra de dados observados. Diversas abordagens para a estimação desta densidade podem ser utilizadas, e encontram-se classificadas em estimações paramétricas e não paramétricas. A estimação paramétrica requer uma especificação a priori de uma família de distribuições da qual se supõe a amostra de dados foi obtida. Porém, na estimação não paramétrica este pressuposto não é necessário e técnicas como o histograma ou a estimação por *kernel* podem ser aplicadas.

O histograma apresentou-se como o único estimador de densidade não paramétrico até a década de 1950, quando importantes e simultâneos avanços foram feitos tanto para a estimação da densidade quanto para a estimação da densidade espectral. Em um artigo pouco conhecido, Fix & Hodges (1951) introduziram o algoritmo básico da estimação da densidade não paramétrica. Neste trabalho, abordou-se o problema da discriminação estatística quando a forma paramétrica da densidade de amostragem é não conhecida. Durante a década seguinte, vários algoritmos gerais e modos teóricos alternativos de análise foram introduzidos por Rosenblatt (1956), Parzen (1962), e Cencov (1962). Seguiu-se uma segunda onda de documentos importantes e principalmente teóricos por Watson & Leadbetter (1963), Loftsgaarden & Quesenberry (1965), Schwartz (1967), Epanechnikov (1969), Tarter & Kronmal (1970), e Wahba (1971). A generalização multivariada natural foi introduzida por Cacoullos (1966). Finalmente, na década de 1970, vieram os primeiros trabalhos com foco na aplicação prática destes métodos: Scott et al. (1978) e Silverman (1978).

5.2.

Estimador Inocente

O conceito inicial do estimador da densidade de probabilidade não paramétrica veio do histograma, como foi dito anteriormente. Esta abordagem considera uma partição de dados em m distintos intervalos da mesma largura. Neste caso $(z_k; z_{k+1})$ conforma um

dos intervalos para a amostra de dados de interesse para todo $m < n$ e $z_{k+1} - z_k = h$. O estimador da função de densidade através do histograma está dado por:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n I[z_k < x_i \leq z_{k+1}] \quad \forall k \text{ tal que } z_k < x \leq z_{k+1}, \quad (5.2)$$

onde $I(\cdot)$ é uma função indicadora, igual a 1 se seu argumento é válido ou 0 caso contrario. A estimação da função de densidade via histograma possui as seguintes características:

1. A representação gráfica dos dados é não suavizada
2. A representação depende da escolha do ponto de origem
3. A representação também depende do número de intervalos usados

tornando-se em desvantagens para o calculo da densidade. Estas dificuldades podem ser contornadas através do uso do estimador *kernel*. Para evitar esta situação, ou seja, para evitar ter que fazer este tipo de escolhas arbitrárias (itens 2 e 3), o estimador inocente considera intervalos que podem-se sobrepor ao invés de estar separados uns dos outros. O principio aqui é estimar a função de densidade em um ponto contando o número de observações que estão "próximas" a este ponto de avaliação. Para uma amostra de n observações, x_1, \dots, x_n o estimador inocente da densidade é:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n I\left[x - \frac{h}{2} < x_i < x + \frac{h}{2}\right], \quad (5.3)$$

sendo h a largura do intervalo e $I(\cdot)$ é uma função indicadora igual a 1 se seus argumentos são válidos e 0 caso contrario. Nesta equação, a estimativa da densidade no ponto x é dada pela proporção de observações que estão dentro de uma distância de $h/2$ ou menos desde o ponto x . A densidade global é obtida fazendo deslizar esta janela de largura h ao longo de todos os pontos de avaliação.

Este estimador pode ser comparado com um histograma com intervalos de largura h , isto é, um histograma no qual todos os pontos estão no centro do intervalo. Se o ponto de avaliação x da função de densidade corresponde ao centro de um dos intervalos do histograma, a avaliação dada pelo estimador inocente (5.3) neste ponto é exatamente a mesma que a dada pelo histograma (5.2). Porém, para todos os pontos que não correspondem ao centro de um intervalo, o estimador inocente produz uma representação gráfica diferente, revelando muito mais detalhe sobre a curvatura da função de densidade,

já que a estimação é realizada em todos os pontos, e não por intervalos separados, o que faz que o problema da escolha do ponto de origem desapareça.

A escolha da largura do intervalo permite escolher o grau de suavização: valores grandes de h produzem funções mais suavizadas, em quanto que valores muito menores de h produzem curvas que não são em absoluto suavizadas. O valor de h tem de ser suficientemente grande para garantir que há um número suficiente de observações por intervalo, mas suficientemente pequeno para evitar o excesso de suavização da função resultante. Critérios de escolha ótima ou inspeção visual podem ser usados para determinar um valor satisfatório de h . Este valor dependerá do número de observações que se tenham disponíveis, e tipicamente será menor quantos mais dados se tenham.

Desde o ponto de vista matemático o estimador inocente não é um estimador desejável de funções de densidade, pois não é diferenciável em todos os pontos, o que significa a não continuidade da função nos pontos $x \pm h/2$ e a nulidade da derivada em todos os pontos onde a função é contínua. Para obter um estimador que seja diferenciável em todos os pontos, recorre-se à estimação por *kernel*. Para fazê-lo, é útil reescrever o estimador inocente como segue:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n w\left(\frac{x - x_i}{h}\right), \quad (5.4)$$

sendo

$$w(x) = \begin{cases} 1 & \text{se } |x| < 1/2 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.5)$$

A função $w(\cdot)$ pode ser representada por um retângulo. Como tal este estimador é a soma de retângulos, de largura h e altura $(nh)^{-1}$, colocado ao redor de cada uma das observações; e onde a função $w(\cdot)$ atribui um peso de 1 a todas as observações que pertencem ao intervalo centrado em x , e peso de 0 às observações restantes.

5.3.

Estimador *Kernel* Univariado

Este estimador é uma generalização do estimador inocente, o qual permite superar o problema de diferenciabilidade em todos os pontos, que como pode se observar na Equação (5.5), este problema vem de $w(\cdot)$. O princípio da estimação por *kernel* é simples: ao invés de dar o mesmo peso a todas as observações no intervalo, o peso

atribuído é maior quanto mais perto a observação estiver de x , isto é, a transição de 1 para 0 nos pesos não é realizada de forma abrupta, mas sim de forma gradual. É provável que o método seja o mais utilizado para a estimação de densidade, e é obtido substituindo a função de ponderação $w(\cdot)$ por uma função *kernel* não negativa $K(\cdot)$:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (5.6)$$

sendo h é o parâmetro de suavização (sabe-se que a escolha de h afeta extremamente a forma da densidade estimada), $X_i, i = 1, 2, \dots, N$ é a amostra de dados e K é uma função *kernel* corretamente escolhida. Exemplos destas funções são os *kernels* tais como o uniforme, gaussiano, epanechnikov, quadrático, entre outros. A fim de que $\hat{f}(x)$ conserve as propriedades de uma função de densidade, a função *kernel* tem de satisfazer a seguinte condição:

$$\int_{-\infty}^{\infty} K(u) du = 1. \quad (5.7)$$

O *kernel* $K(\cdot)$ é usualmente tomado como uma função densidade de probabilidade simétrica. Isto implica que o mesmo peso é atribuído a todas as observações que estão situadas à mesma distância de x . Pelas propriedades da função $K(\cdot)$, tem-se que \hat{f} é uma função densidade, haja vista que sua integral sobre os reais é igual à unidade. Analogamente, \hat{f} herda as propriedades de continuidade e diferenciabilidade de $K(\cdot)$.

Na estimação por *kernel* dois parâmetros têm de ser determinados: o *kernel* $K(\cdot)$ e o parâmetro de suavização h . Embora a escolha do *kernel* afeta a estimativa da densidade, a literatura sugere que esse efeito é bastante pequeno, com resultados empíricos muito semelhantes para diferentes escolhas de $K(\cdot)$. Segundo Scott (1992), a escolha da função *kernel* tem um papel menor na qualidade final da estimativa e sugere evitar o uso que classicamente tem-se do *kernel gaussiano* para reduzir a sobrecarga computacional. O *kernel gaussiano* corresponde à distribuição normal padrão, representado por:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \quad (5.8)$$

O *kernel Epanechnikov*(1969) é um polinômio de segundo grau, ajustado para satisfazer as propriedades de uma função de densidade. Tem a vantagem de ser analiticamente simples, tendo também propriedades de otimalidade e está dado por:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{se } |u| \leq 1 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.9)$$

O *kernel triangular* nem sempre é diferenciável, no entanto, pode ser escrito como:

$$K(u) = \begin{cases} 1 - |u| & \text{se } |u| \leq 1 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.10)$$

O *kernel uniforme* está dado por:

$$K(u) = \begin{cases} 1/2 & \text{se } |u| \leq 1 \\ 0 & \text{caso contrario,} \end{cases} \quad (5.11)$$

observe-se que este *kernel* corresponde ao estimador inocente, tal como definido na Equação (5.3), com uma função de ponderação normalizada.

O *kernel quadrático* ou *biweight* está definido por:

$$K(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2 & \text{se } |u| \leq 1 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.12)$$

O *kernel cúbico* ou *triweight* se define como:

$$K(u) = \begin{cases} \frac{35}{32}(1 - u^2)^3 & \text{se } |u| \leq 1 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.13)$$

Estes *kernels* estão representados na Figura 5.1. Com exceção do *kernel uniforme* todos os *kernels* são simétricos em torno de zero e diminuem gradualmente à medida que se afastam de zero. A diferença entre eles é que, por exemplo, o *kernel Gaussiano* sempre assume valores positivos, mesmo que estes sejam extremamente pequenos fora do intervalo $[-3, 3]$, enquanto que o *kernel Epanechnikov* produz valores de zero fora do

intervalo $[-1, 1]$. Na prática, os *kernels* *Gaussiano* e *Epanechnikov* são usados com maior frequência.

Em geral, a função *kernel* está normalizada; isto é, a variância é definida igual a 1. Isto implica que o parâmetro h desempenha o papel da variância na estimativa da densidade na Equação (5.5). Este parâmetro corresponde à escolha da largura do intervalo, e determina o grau de suavização da estimativa. Este procedimento é chamado de escolha do parâmetro de suavização ou seleção do tamanho da janela.

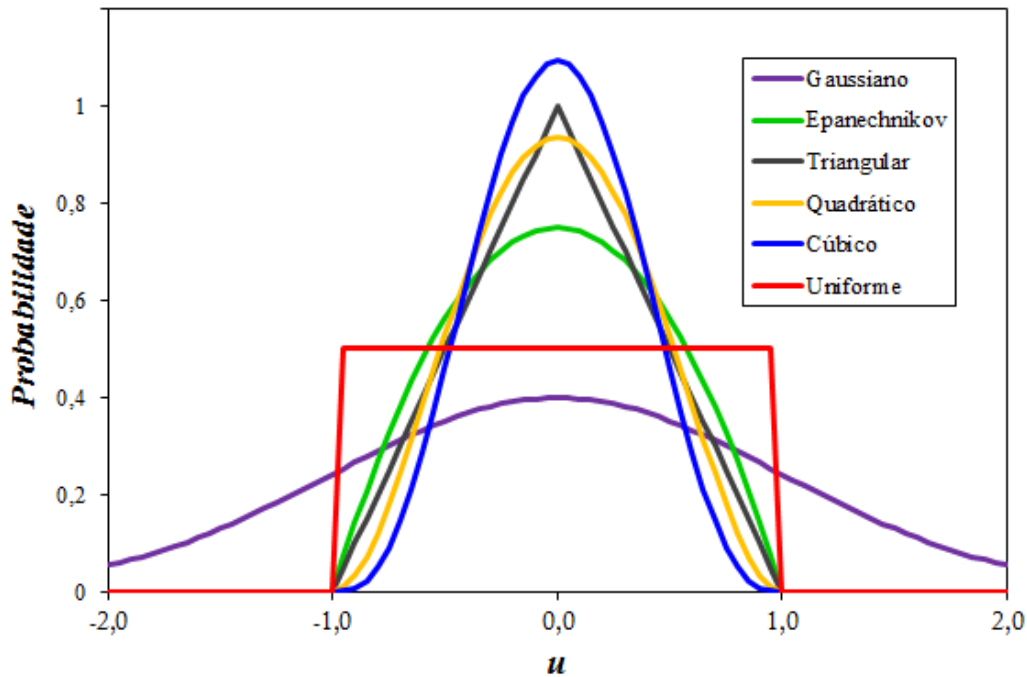


Figura 5.1 – Funções *Kernel*. (Fonte: Elaboração do Autor)

Por outra parte, as medidas de discrepância entre o estimador de densidade \hat{f} e a verdadeira densidade são de suma importância na construção das funções *kernel* e na escolha do parâmetro de suavização h . Quando se considera a estimação em um único ponto, uma medida natural da discrepância entre o estimador e a densidade real é o Erro Quadrático Médio (MSE), o qual está dado pela distância média entre o valor estimado e o valor real:

$$MSE(\hat{f}(x)) = E([\hat{f}(x) - f(x)]^2), \quad (5.14)$$

esta Equação (5.14) pode ser descomposta em dois termos:

$$MSE(\hat{f}(x)) = (E[\hat{f}(x)] - f(x))^2 + Var[\hat{f}(x)]. \quad (5.15)$$

A primeira expressão (quadrática) corresponde ao viés do estimador, a segunda expressão a sua variância e $\hat{f}(x)$ é a densidade estimada no ponto x . Na prática, o interesse radica na precisão da estimação em todos os pontos e não somente em um ponto específico. Por conseguinte, uma medida de precisão global de \hat{f} é o Erro Quadrático Médio Integrado (MISE), dado por:

$$MISE(\hat{f}(x)) = E \left(\int [\hat{f}(x) - f(x)]^2 dx \right). \quad (5.16)$$

Analogamente ao MSE, o MISE pode ser decomposto em dois termos:

$$MISE(\hat{f}(x)) = \int (E[\hat{f}(x)] - f(x))^2 dx + \int Var[\hat{f}(x)] dx, \quad (5.17)$$

isto é, a integral do quadrado do viés somada à integral da variância do estimador. O primeiro método de estimação do parâmetro h foi proposto por Silverman (1986), e se baseia na minimização do Erro Quadrático Médio Integrado (MISE), para encontrar h que permita obter a melhor estimação de $\hat{f}(x)$. De acordo com Silverman (1986), o viés pode ser representado por $\frac{1}{2}h^2f''(x)k_2$ e a variância por $\frac{1}{nh} \int K(t)^2 dt$, desta forma, a Equação (5.17) é reescrita como:

$$MISE(\hat{f}(x)) = \frac{1}{4}h^4k_2^2 \int f''(x)^2 dx + \frac{1}{nh} \int K(t)^2 dt, \quad (5.18)$$

sendo h a largura do *kernel*, k^2 uma constante resultante pertencente ao segundo termo no processo de expansão da Série de Taylor, $f''(x)$ a derivada segunda da densidade real, n o tamanho da amostra dos dados e $K(\cdot)$ a função *kernel* utilizada. Silverman (1986) estabelece que o valor de h ótimo é aquele que minimiza o MISE. O uso do critério MISE requer da avaliação de uma derivada que depende da função de densidade verdadeira, a qual é desconhecida. Uma forma natural de desenvolver um método de seleção automática, que é útil na prática, é usar uma distribuição de referência. Desta forma, a distribuição normal com variância de σ^2 é empregada como a distribuição de referência, e como função *kernel* o *kernel Gaussiano*. Assim, o valor de h ótimo é dado por:

$$h_{opt} = 1,059 \sigma n^{-\frac{1}{5}}. \quad (5.19)$$

Uma forma rápida de calcular o parâmetro σ é substituí-lo pelo desvio padrão dos dados da amostra, que é um estimador de σ . Uma desvantagem do desvio padrão dos dados é que é muito sensível a outliers presentes na amostra. Esta situação se apresenta quando a amostra de dados é pequena ou quando a densidade verdadeira é caracterizada por caudas pesadas. Silverman (1986) sugere usar um estimador alternativo ou uma dispersão que seja mais robusta a outliers; propondo para a determinação do parâmetro de suavização h o intervalo interquartil, definido como a distância entre o terceiro e o primeiro quartil ($q_3 - q_1$), dado por:

$$h_{opt} = 1,059 \left(\frac{q_3 - q_1}{1,349} \right) n^{-\frac{1}{5}}, \quad (5.20)$$

onde o valor de 1,349 corresponde ao intervalo interquartil de uma distribuição normal padrão $N(0,1)$. Estas duas propostas de largura de h funcionam bem na prática quando a densidade real está perto da distribuição de referência, i.e., à distribuição normal. No entanto, este não é o caso quando a densidade verdadeira está distante da distribuição normal. Silverman (1986) analisa o desempenho de diversas opções de larguras automáticas quando a densidade verdadeira não está perto de ser normal e propõe utilizar:

$$\hat{h}_{opt} = 0,9 \min \left(\hat{\sigma}; \frac{\hat{q}_3 - \hat{q}_1}{1,349} \right) n^{-\frac{1}{5}}, \quad (5.21)$$

onde σ é o desvio padrão dos dados, q_3 e q_1 são, respectivamente, o primeiro e o terceiro quartil calculados a partir dos dados. Esta outra abordagem resume-se à utilização do mínimo das duas opções anteriores, substituindo o valor de 1,059 por 0,9 para reduzir o risco de sobre-suavização da função, funcionando bem na maioria dos casos. Contudo, quando a densidade real é multi-modal e muito assimétrica, a formulação da Equação (5.21) tende a sobre-suavizar a distribuição.

Pode-se pontuar que na estimação de densidade a escolha do parâmetro de suavização h constitui um problema crucial, pois ele tem uma grande influência na qualidade da distribuição estimada. Além dos enfoques apresentados anteriormente, existem vários métodos que indicam caminhos para a escolha do parâmetro de suavização h . Em algumas situações, por exemplo, examinam-se gráficos ponderados por diferentes parâmetros de suavização, e escolhe-se aquele que mais se assemelha com ideias que se

tinha *a priori*. Outros métodos de escolha do parâmetro de suavização são: validação cruzada por mínimos quadrados, validação cruzada por verossimilhança, teste gráfico, referência a f como uma distribuição padrão.

5.4.

Estimador *Kernel* para Dados Bivariados

Até agora só foi considerada a estimação da densidade para uma única variável, no entanto, em muitos casos, é bastante útil estimar a densidade conjunta de uma série de variáveis. Este método é uma extensão para o caso multivariado trabalhado na Seção anterior. A extensão para duas dimensões da estimação de densidade pelo método de *kernel* é simples e informativa, porém as dificuldades na representação dos dados para além de duas dimensões torna improvável o uso deste método de estimação para fins exploratórios. De outro modo, se a intenção não é ver a forma da densidade estimada, mas sim usar a estimação de densidade como um passo intermediário para outra técnica, esta técnica se torna interessante para estimar densidades de altas ordens.

Considerando o caso bidimensional, a densidade conjunta de duas variáveis x e y , para as quais se têm n observações, pode ser estimada com uma função *kernel* bivariada:

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^n K\left(\frac{x - X_i}{h_x}; \frac{y - Y_i}{h_y}\right), \quad (5.22)$$

o qual é equivalente ao produto de dois *kernels* univariados no caso de gaussiano, com $K(\cdot)$ definida como uma função densidade de probabilidade bivariada, radialmente simétrica e unimodal.

Por sua vez, o *kernel* gaussiano bivariado baseia-se na distribuição normal padrão bivariada:

$$K(u_1, u_2) = (2\pi)^{-1} \exp\left(-\frac{u_1^2 + u_2^2}{2}\right). \quad (5.23)$$

Esta função é igual ao produto de duas distribuições normais padrão. Como tal, a utilização do *kernel* Gaussiano multivariado ou do produto de dois *kernels* gaussianos univariados produz os mesmos resultados.

O *kernel* Epanechnikov bivariado é definido por:

$$K(u_1, u_2) = \begin{cases} 2\pi^{-1}(1 - u_1^2 - u_2^2) & \text{se } u_1^2 + u_2^2 < 1 \\ 0 & \text{caso contrario.} \end{cases} \quad (5.24)$$

Esta expressão não é igual ao produto de duas distribuições univariadas Epanechnikov. Conseqüentemente, a utilização do *kernel* Epanechnikov multivariado ou o produto de dois *kernels* Epanechnikov univariados não produz os mesmos resultados. É apenas o *kernel* Epanechnikov bivariado quem atribui um peso aos pontos que diminuem, de um modo simétrico e circular, com a distância a partir do ponto de referência.

A escolha do parâmetro de alisamento é mais complexa no caso multivariado. Isto acontece, em particular, porque a avaliação visual dos diferentes valores dos parâmetros torna-se consideravelmente mais difícil em mais de duas dimensões, mas também a precisão da estimativa. Acima de duas dimensões, a escolha dos parâmetros de suavização é, portanto, determinada por regras de seleção automática. Estas podem ser desenvolvidas como extensões naturais daqueles desenvolvidas para o caso univariado. Por conseguinte, os métodos analisados anteriormente propostos por Silverman (1986) estão baseados em estimativas da densidade através de *kernel*, e sua extensão para o caso multivariado é direta com o uso de *kernels* multivariados. A sua utilização, no entanto, torna-se mais complexa, pois requer o cálculo das derivadas de distribuições multivariadas e uma grande quantidade de cálculos numéricos. Porém é possível contornar esta situação, apelando para uma distribuição de referência, como veremos a continuação.

Tomando as distribuições de referência como sendo Normais multivariadas $N(\mu, \Sigma)$, onde a matriz de covariância é a diagonal $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$, obtém-se uma regra empírica que é igual de simples ao caso univariado. A minimização de uma aproximação para o MISE produz uma largura ótima do parâmetro de suavização $H = \text{diag}(h_1, \dots, h_d)$ dada por:

$$h_j = \left(\frac{4}{d+2} \right)^{1/(d+4)} \cdot \sigma_j \cdot n^{-1/(d+4)}, \quad (5.25)$$

sendo $j = 1, \dots, d$, e σ_j os desvios padrão das amostras. Quando $d = 1$, obtém-se a expressão da Equação (5.19). Scott (1992) notou que os primeiros elementos desta expressão sempre tomam valores entre 0.924 e 1.059, pelo que propôs uma extensão da regra de Silverman (1986) para calcular H dada por:

$$\hat{h}_j = \hat{\sigma}_j \cdot n^{-1/(d+4)}. \quad (5.26)$$

A mesma análise não pode ser efetuada para proporcionar uma regra no caso de uma matriz de covariância cujos elementos fora da diagonal da distribuição de referência sejam diferentes de zero. No entanto, a Equação (5.25) sugere que H é proporcional a $\Sigma^{1/2}$. Pode-se, portanto, generalizar a regra de Scott (1992), considerando:

$$\hat{H} = \hat{\Sigma}^{1/2} \cdot n^{-1/(d+4)}. \quad (5.27)$$

A aplicação desta regra empírica é análoga à:

1. Transformar os dados de forma tal a que a sua matriz de covariância seja a matriz identidade;
2. Estimar a densidade com um único parâmetro de alisamento $h = n^{-1/(d+4)}$;
3. Transformar a densidade estimada para trazê-la de volta à escala original.

Esta transformação prévia dos dados é simples e muito útil para a estimação prática de uma densidade multivariada. Tal transformação permite usar um único parâmetro de alisamento na estimação da densidade, tendo em conta as diferentes concentrações dos dados em toda a dimensão e qualquer correlação entre as variáveis.

Sabe-se que para o mesmo conjunto de dados, a mudança do parâmetro de suavização altera significativamente a forma da densidade estimada. De modo análogo ao caso univariado, existem mecanismos na mesma linha de desenvolvimento para a escolha do parâmetro de suavização usando validação cruzada.

5.5.

Estimação da Densidade Condicional por *Kernel*

A estimação densidade condicional foi uma importante derivação da estimação da função de densidade, a qual foi introduzida por Rosenblatt (1969). Uma correção do viés foi proposta por Hyndman et al. (1996). Em Fan et al. (1996), propôs-se um estimador direto baseado na estimação polinomial local. Regras de seleção do tamanho da janela têm sido propostas por Bashtannyk & Hyndman (2001), Fan & Yim (2004) e Hall et al. (2004). O problema relacionado com estimação da distribuição condicional é examinado em Hall et al. (1999). Outros trabalhos têm utilizado estimativas da densidade condicional como uma entrada para outros problemas, incluindo Robinson (1991), Tjøstheim (1994),

Polonik & Yao (2000), Hyndman & Yao (2002), Juban et al. (2007a, 2007b), Bessa et al. (2012b).

Formalmente, deseja-se estimar a densidade condicional da variável aleatória Y dada $X = x$, para isto suponha que a variável explicativa X , é uma variável aleatória univariada e que ambas variáveis possuem uma distribuição de densidade conjunta $f(x, y)$. Tem-se uma amostra aleatória bivariada que será denotada por $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ e cujas observações estão dadas por $\{(x_1, y_1), \dots, (x_n, y_n)\}$. Assuma-se também que estas observações bivariadas são independentes. Desta forma, a média condicional é denotada por $r(x) = E[Y|X = x]$ pode-se escrever $Y_j|X_j = x_j = r(x_j) + \varepsilon_j$, onde $E[\varepsilon_j] = 0$, $V[\varepsilon_j] = \sigma_\varepsilon^2$ e os ε_j são independentes ($cov[\varepsilon_i, \varepsilon_j] = 0$ para $i \neq j$), mas não necessariamente identicamente distribuídos.

Ao expressar a função de densidade condicional em termos da densidade conjunta $f(x, y)$, pode-se derivar seu estimador. Para isto, seja $f(x)$ a densidade marginal de X . Assim a densidade condicional de $Y|(X = x)$ está dada por:

$$f(y|x) = \frac{f(x, y)}{f(x)}. \quad (5.28)$$

Será assumido que $f(y|x)$ e $f(x)$ são funções tais que as suas segundas derivadas são contínuas e são quadraticamente integráveis, e que $r(x)$ tem sua segunda derivada contínua.

Deseja-se estimar o numerador e o denominador da Equação (5.28) separadamente utilizando estimadores de *kernel* ao invés de estimativas paramétricas das funções de densidade. Em primeiro lugar, a densidade conjunta $f(x, y)$ pode ser estimada usando uma função *kernel* do produto de dois *kernels* univariados, fornecendo o mesmo resultado do estimador *kernel* bivariado da Equação (5.22) para o caso gaussiano, como exemplificado em Scott (1992). Segundo Wand & Jones (1995), outro tipo de *kernel* multivariado (bivariado) ao invés de produto de *kernels* pode ser usado para definir $\hat{f}(x, y)$. No entanto, recomenda-se utilizar o produto de *kernels* devido a que é mais simples de ser trabalhado, e a que conduz a estimadores de densidade condicional com várias propriedades assintóticas boas. O estimador para $\hat{f}(x, y)$ é dado por:

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{j=1}^n K\left(\frac{x - X_j}{h_x}\right) K\left(\frac{y - Y_j}{h_y}\right) = \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j) K_{h_y}(y - Y_j), \quad (5.29)$$

que substituí a densidade bivariada $f(x, y)$ desconhecida, no numerador da Equação (5.28). Esta formulação contém dois parâmetros de suavização, h_x e h_y . Estes são parâmetros fixos de escala que controlam a quantidade de suavização para a estimativa da densidade conjunta na direção de x e y respectivamente. $K_h(\cdot) = K(\cdot/h)$ é uma função de densidade de probabilidade simétrica, conhecida como função *kernel*, a qual cumpre com os seguintes pressupostos gerais (Silverman, 1986):

$$\begin{aligned} K(u) > 0, \quad \int_{\mathbb{R}} K(u) du = 1, \\ \int_{\mathbb{R}} uK(u) du = 0 \quad \text{e} \quad \sigma_K^2 = \int_{\mathbb{R}} u^2 K(u) du < \infty \\ \neq 0 \end{aligned} \tag{5.30}$$

Para obter a estimativa do denominador da Equação (5.28) correspondente à função densidade marginal $f(x)$, lembremos que:

$$f(x) = \int f(x, y) dy. \tag{5.31}$$

Desta forma, empregando o estimador de densidade por *kernel* tem-se:

$$\begin{aligned} \int \hat{f}(x, y) dy &= \int \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j) K_{h_y}(y - Y_j) dy \\ &= \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j) \int K_{h_y}(y - Y_j) dy. \end{aligned} \tag{5.32}$$

Pelo pressuposto da função *kernel* em que $\int K(u) du = 1$, obtém-se:

$$\hat{f}(x) = \int \hat{f}(x, y) dy = \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j). \tag{5.33}$$

Por conseguinte, ao substituir as Equações (5.29) e (5.33) em (5.28) é obtido o estimador seguinte:

$$\hat{f}(y|x) = \frac{\frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j) K_{h_y}(y - Y_j)}{\frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j)}. \quad (5.34)$$

Reescrevendo os termos:

$$\hat{f}(y|x) = \sum_{j=1}^n \mathcal{W}_j(x) K_{h_y}(y - Y_j), \quad (5.35)$$

sendo

$$\mathcal{W}_j(x) = \frac{K_{h_x}(x - X_j)}{\sum_{i=1}^n K_{h_x}(x - X_i)}. \quad (5.36)$$

A média do estimador de densidade condicional $\hat{f}(y|x)$ fornece um estimador da média condicional de $r(x)$:

$$\hat{r}(x) := \int y \hat{f}(y|x) dy = \sum_{j=1}^n \mathcal{W}_j(x) Y_j, \quad (5.37)$$

sendo esta expressão é idêntica à função de regressão por *kernel* de Nadaraya(1964) e Watson(1964), e que é na verdade, a forma como o estimador *Nadaya-Watson* é frequentemente derivado (por exemplo, Scott 1992). Por definição, a média condicional de $Y|X = x$ está dada por:

$$r(x) = E(y|x) = \int y f(y|x) dy = \int \frac{y f(x, y)}{f(x)} dy. \quad (5.38)$$

Estimando o numerador da Equação (5.38), tem-se:

$$\begin{aligned} \int y \hat{f}(x, y) dy &= \frac{1}{n} \int y \sum_{j=1}^n K_{h_x}(x - X_j) K_{h_y}(y - Y_j) \\ &= \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j) \int y K_{h_y}(y - Y_j) dy. \end{aligned} \quad (5.39)$$

Pelos pressupostos da função *kernel* em que $\int uK(u)du = 0$ e $\int K(u)du = 1$, obtém-se:

$$\int yK_{h_y}(y - Y_j)dy = Y_j. \quad (5.40)$$

Portanto, a estimativa do numerador está dada por:

$$\int y\hat{f}(x, y)dy = \frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j)Y_j. \quad (5.41)$$

Em cálculos anteriores o estimador do denominador da Equação (5.38) foi determinado e está dado pela Equação (5.33). Substituindo as Equações (5.33) e (5.41) em (5.38), obtém-se o estimador da média condicional $r(x)$:

$$\begin{aligned} \hat{r}(x) = E(y|x) &= \int y\hat{f}(y|x)dy = \int \frac{y\hat{f}(x, y)}{\hat{f}(x)} dy \\ &= \frac{\frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j)Y_j}{\frac{1}{n} \sum_{j=1}^n K_{h_x}(x - X_j)}. \end{aligned} \quad (5.42)$$

Reescrevendo os termos:

$$\hat{r}(x) = E(y|x) = \int y\hat{f}(y|x)dy = \sum_{j=1}^n \mathcal{W}_j(x)Y_j, \quad (5.43)$$

expressão que corresponde ao estimador *Nadaya-Watson* da Equação (5.37) da média condicional $r(x)$, e com $\mathcal{W}_j(x)$ dado pela Equação (5.36)

Outros estimadores da densidade condicional como o estimador em dois-estágios proposto por (Hyndman et al., 1996), o estimador da media Loess proposto por Cleveland (1979) e o estimador Parzen-Rosenblatt (Parzen, 1962; Rosenblatt, 1969) podem ser usados para o cálculo de dita densidade; este último estimador é utilizado no trabalho de Jeon & Taylor (2012) e indica-se que não existem muitos exemplos aplicados a séries temporais. Nesta tese será usado o estimador de *Nadaraya-Watson* quem possui boas propriedades estatísticas. Desta forma, o seguinte Capítulo apresenta uma estratégia de especificação que leve em consideração a previsão da velocidade do vento através de SSA, e a estimação da densidade condicional da geração eólica para sua previsão probabilística utilizando o este estimador *kernel*.