Universidade Federal de Juiz de Fora

Departamento de Estatística

Curso de Estatística

Samuel Faria Cândido

Estimação de Densidades de Dados Viesados Via Bases de Cossenos

Juiz de Fora

2019

Samuel Faria Cândido

Estimação de Densidades de Dados Viesados Via Bases de Cossenos

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Professor Michel Helcias Montoril

Juiz de Fora

2019

Samuel Faria Cândido

Estimação de Densidades de Dados Viesados Via Bases de Cossenos

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em:

BANCA EXAMINADORA

Professor Marcel de Toledo Vieira - Presidente da Banca Examinadora
Universidade Federal de Juiz de Fora

Professora Camila Borelli Zeller
Universidade Federal de Juiz de Fora

Professor Lupércio França Bessegato
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Durante a graduação em Estatística, muitas pessoas me ajudaram, de alguma forma, a concluir essa etapa.

Quero agradecer em primeiro lugar a Deus, por ter me dado condições de concluir esse curso, aos meus pais, que sempre me apoiaram em todas as dificuldades pelas quais passei durante toda a minha vida. Sem eles, certamente não conseguiria terminar esta graduação.

Quero agradecer também a UFJF e a todos os professores, com quem tive contato na Universidade, em especial os professores do Departamento de Estatística que sempre foram muito prestativos e me ajudaram em tudo que eu precisei durante todos esses anos.

Quero agradecer também meu orientador Michel Helcias Montoril que me ajudou durante o curso principalmente nos dois anos em que foi meu orientador de iniciação cientifica

Finalmente, agradeço a todos os colegas principalmente o Filipe Fernandes e o João Gabriel dois grandes amigos com os quais cursei muitas das disciplinas da graduação.

A todos os citados acima, sou eternamente grato.

“Apenas busquem conhecimento”

Autor Desconhecido

RESUMO

Nesse trabalho, primeiramente consideramos o problema de estimação de densidades, nesse contexto, apresentamos um estimador não paramétrico via bases de cossenos, e aplicamos a metodologia proposta a um conjunto de dados relativo a velocidade de 82 galáxias medidas na região da Coroa Boreal. Posteriormente apresentamos o conceito de dados viesados e adaptamos esse estimador para este cenário. Além disso, avaliamos os resultados desse método em dados simulados os quais geramos com o auxilio de um algoritmo de aceitação-rejeição e finalmente aplicamos a metodologia em uma amostra, retirada de um conjunto de dados que corresponde a concentração de álcool no sangue de motoristas nos 50 estados americanos, no distrito de Columbia e em Porto Rico desde 1975.

Palavras-chave: Estimação de densidades, Dados viesados, Estimação não paramétrica, Bases de cossenos.

ABSTRACT

We consider first the problem of density estimation, we introduce a non-parametric estimator for this case, using a cosine basis. We apply the proposed methodology in a data set relative to the velocity of 82 galaxies measured in the region of Corona Borealis. Posteriorly, we presented the concept of biased data and we adapt the former estimator to this context. After that, we evaluate the results of this method using data simulated by an acceptance-rejection algorithm and later we apply this methodology in a sample taken from a dataset, containing information of blood alcohol concentration in drivers' of the 50 american states, the District of Columbia, and Puerto Rico since 1975.

Key-words: Density Estimation, Biased Data, Nonparametric Estimation, Cosine Basis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MISE	Erro quadrático integrado médio (do inglês, *Mean Integrated Squared Error*)
ISB	Viés Quadrático Integrado (do inglês, *Integrated Squared Bias*)
EQM	Erro quadrático médio
CAS	Concentração de álcool no sangue
iid	Independente e identicamente distribuída

SUMÁRIO

1 INTRODUÇÃO

Em [Ramirez e Vidakovic, 2010] os autores definem dados viesados como sendo dados provenientes de um processo amostral em que a probabilidade de uma determinada observação ser escolhida depende do seu valor. O objetivo desse trabalho é propor um estimador de séries de Fourier para dados com essas características. Nesse trabalho, o objetivo é estimar a densidade de uma variável aleatória X denotada por f_X através de observações indiretas de uma variável aleatória Y. A densidade viesada denotada por f_Y é escrita da seguinte forma

$$f_Y(y) := \frac{g(y)f_X(y)}{\mu}, \tag{1.1}$$

Nessa definição, f_X representa a verdadeira densidade que estamos interessados em estimar, $g(y)$ é uma função positiva denominada na literatura como função de viés e μ é definido como

$$\mu = \mathbb{E}[g(X)] < \infty. \tag{1.2}$$

Estimar densidades a partir de uma amostra viesada é uma necessidade recorrente em varias situações. Em [Efromovich, 1999] o autor propõe o seguinte exemplo de amostra viesada: Suponha, que estejamos interessados em estimar a distribuição da concentração de álcool no sangue de motoristas que transitam sobre uma determinada rodovia e que os dados estão disponíveis nos relatórios policiais sobre motoristas presos acusados de dirigir sob a influência de álcool. Como um motorista embriagado tem uma maior chance de chamar a atenção da polícia, esses dados são viesados.

Agora suponha que estejamos interessados em estimar o tamanho médio das turmas de uma determinada escola. Para isso, seleciona-se aleatoriamente alguns estudantes, pergunta-se a esses alunos quais os tamanhos das suas turmas e calcula-se a média dessas respostas. Em um primeiro momento parece não haver nada de errado nesse método, mas se perguntarmos à diretora da escola qual o tamanho médio das turmas ela provavelmente informará um valor menor do que o estimado com base nas respostas dos alunos. Isso acontece porque quando selecionamos os alunos, aqueles que pertencem a uma turma maior tem uma probabilidade maior de serem selecionados do que os que pertencem a turmas menores, fazendo com que a media amostral seja superestimada. Esse é um exemplo clássico do "paradoxo da inspeção" [Feller, 1971, Ross, 1993]. Mais exemplos práticos onde ocorrem amostras viesadas podem ser encontrados em [Cox, 1969, Laake et al., 1993, Cook e Martin, 1974].

Várias formas de estimação de densidades de dados viesados podem ser encontradas na literatura. Efromovich em [Efromovich, 2004] apresenta resultados de um estimador minimax para densidades de dados viesados e avalia como uma amostra viesada afeta

a precisão da estimativa. [De Uña Álvarez e Rodríguez Casal, 2007] desenvolveram um algoritmo EM para estimar densidades de dados viesados. [Vardi, 1982] deriva um estimador de máxima verossimilhança não paramétrico para f_X e [Jones, 1991] discute as propriedades do erro quadrático médio de um novo estimador de densidade via núcleo estimador para dados os viesados.

Este trabalho está organizado como segue. No capítulo 2 é introduzido o conceito de aproximação de funções por séries ortonormais. No Capítulo 3, é apresentado um estimador de séries de Fourier para densidades e aplicamos essa ideia em um conjunto de dados. No Capítulo 4, o estimador proposto no Capítulo 3 é adaptado para o contexto de dados viesados. No Capítulo 5 é aplicada a metodologia em amostras viesadas geradas a partir de um algoritmo de aceitação-rejeição. No Capítulo 6 a metodologia é aplicada em um conjunto dados reais e no Capítulo 7 são feitas as considerações finais sobre o trabalho.

2 Aproximação por séries ortonormais

Nessa seção, vamos falar um pouco sobre aproximação por séries e apresentar o sistema ortonormal de cossenos que será o método utilizado nos próximos capítulos deste trabalho

Vamos começar definindo o conceito de função. Uma função $f : A \to B$ é uma regra que associa para cada elemento $x \in A$ um elemento $y \in B$. Existem três métodos tradicionais usados para definir uma função: através de uma tabela, pela formula ou pelo seu gráfico.

Uma outra forma de descrever uma função é através de uma expansão em series. Suponha que uma função f tenha como domínio o intervalo $[0, 1]$, então

$$f(x) = \sum_{j=0}^{\infty} \theta_j \phi_j(x), \text{ em que } \theta_j = \int_0^1 \phi_j(x) f(x) dx. \tag{2.1}$$

Nessa definição, $\{\phi_0, \phi_1, \ldots\}$ são os elementos do sistema ortonormal e $\{\theta_0, \theta_1, ...\}$ são os coeficientes de Fourier. Um sistema de funções é dito ortonormal se

$$\int_0^1 \phi_j(x)\phi_k(x)dx = \begin{cases} 1, \text{se } k = j \\ 0, \text{caso contrário} \end{cases}.$$

Note que para descrever uma função pela expansão infinita definida em (2.1) precisaríamos conhecer um número infinito de coeficientes de Fourier, como isso não é possível, vamos aproximar f por uma série ortonormal truncada

$$f_J(x) = \sum_{j=0}^{J} \theta_j \phi_j(x). \tag{2.2}$$

Nessa definição θ_j são os coeficientes de Fourier que devem ser estimados e J é o ponto de corte (*Cutoff*) que pode ser predeterminado ou estimado, como mostraremos a seguir, uma vantagem dessa abordagem é que ela possibilita uma boa compreensão dos dados. Em termos estatísticos, a principal vantagem desse método é que permite a estimação da função com um número relativamente pequeno de coeficientes o que exige um custo computacional quase que irrisório se comparado a outros métodos de estimação de densidades. Vários sistemas ortonormais podem ser usados na análise de funções (veja, [Efromovich, 1999]), porém nesse trabalho vamos focar apenas no sistema ortonormal de cossenos pois ele é de fácil implementação e produz resultados satisfatórios. Os elementos do sistema de cossenos são definidos como:

$$\phi_j(x) := \begin{cases} 1, \text{se } j = 0 \\ \sqrt{2}\cos(\pi j x), \text{ se } j = 1, 2, 3 \ldots \end{cases}. \tag{2.3}$$

Na próxima seção, vamos utilizar esses conceitos de aproximação por séries ortonormais afim de obtermos um estimador para densidades.

3 Estimador Universal de Séries Ortogonais

Nessa seção supomos o modelo clássico de estimação de densidades em que temos n observações independentes e identicamente distribuídas (iid's) $X_1, ..., X_n$ de uma variável aleatória X. Vamos supor que X é distribuída de acordo com uma função densidade de probabilidade desconhecida $f(x)$ com suporte $[0, 1]$.

É importante salientar que essa metodologia é valida para qualquer suporte $[a, b] \in \mathbb{R}$ e o suporte $[0, 1]$ é escolhido escolhido por questão de conveniência. Caso os dados pertençam à um suporte $[a, b]$ que não seja o $[0, 1]$, os dados devem ser reescalados para o intervalo $[0, 1]$ da seguinte maneira: $W_l = (X_l - a)/(b - a)$. Ao fazermos isso temos que as observações reescaladas seguem uma distribuição $f_W(w)$ com $w \in [0, 1]$, então obtemos uma estimativa $\tilde{f}_W(w)$ para $f_W(w)$ e reescalamos novamente essa estimativa a fim de obtermos uma estimativa para $f_X(x)$, essa estimativa será dada por $\tilde{f}_X(x) := (b - a)^{-1}\tilde{f}_W((x - a)/(b - a))$ com $x \in [a, b]$.

Temos que, sob alguns pressupostos, a função $f(x)$, $x \in [0, 1]$ pode ser aproximada por uma soma parcial (série ortogonal truncada),

$$f_J(x) := \sum_{j=0}^{J} \theta_j \phi_j(x), 0 \leq x \leq 1, \text{ em que } \theta_j = \int_0^1 \phi_j(x) f(x) dx.$$

Nessa seção a base ortonormal $\{\phi_j\}$ será a base de cossenos (2.3), J é o ponto de corte (*cutoff*) e θ_j é o coeficiente de Fourier de correspondente ao elemento ϕ_j da base ortonormal usada.

Como o suporte de f é o intervalo $[0, 1]$, então $\theta_0 = \int_0^1 \phi_0(x) f(x) dx = \int_0^1 f(x) dx = 1$.

Portanto, para estimarmos a densidade desconhecida f, precisamos estimar os coeficientes de Fourier $\{\theta_j\}$ e o ponto de corte J. Como f é uma função densidade de probabilidade podemos escrever os coeficientes de Fourier em (2.1) como uma esperança, ou seja, temos que

$$\theta_j = \mathbb{E}\{\phi_j(X)\}.$$

Para estimar esses coeficientes usaremos a média amostral que é o estimador de momentos para a esperança. Ou seja,

$$\hat{\theta}_j = n^{-1} \sum_{l=1}^{n} \{\phi_j(X_l)\}. \tag{3.1}$$

O próximo passo é escolher o ponto de corte J. A escolha depende da medida de bondade de ajuste desejada para sua estimativa. Medidas de bondade de ajuste servem para medir a discrepância entre os valores observados e os valores esperados sob um modelo

de probabilidade. Nesse trabalho escolhemos mensurar a bondade de ajuste através do Erro Quadrático Integrado médio que aqui será abreviado por MISE (do inglês, *Mean Integrated Squared Error*), dado por

$$\text{MISE}\{\tilde{f}_J, f\} := \mathbb{E}\left\{\int_0^1 (\tilde{f}_J(x) - f(x))^2 dx\right\}.$$

Em que $\tilde{f}_J(x) = \sum_{j=0}^{J} \hat{\theta}_j \phi_j(x)$. Pela identidade de Parseval (veja apêndice A.2) temos que

$$\text{MISE}\{\tilde{f}_J, f\} = \sum_{j=0}^{J} \mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} + \sum_{j=J+1}^{\infty} \theta_j^2. \tag{3.2}$$

O J ótimo é o J que minimiza o MISE, ou seja, devemos definir o J que minimiza os dois termos da direita de (3.2), O primeiro termo é a variância de $\tilde{f}_J$. Esse termo é a soma das $J+1$ variâncias dos estimadores $\hat{\theta}_j$. Podemos escrever cada uma dessas variâncias como

$$\begin{aligned}
Var(\hat{\theta}_j) &= Var(n^{-1} \sum_{l=1}^{n} \{\phi_j(X_l)\}) \\
&= n^{-2} Var(\sum_{l=1}^{n} \{\phi_j(X_l)\}) \\
&= n^{-2} n Var(\phi_j(X)) \\
&= n^{-1} Var(\phi_j(X)) \\
&= n^{-1} [\mathbb{E}[(\phi_j(X))^2] - (\mathbb{E}[\phi_j(X)])^2] \\
&= n^{-1} \left[\mathbb{E}\left[(\phi_j(X))^2\right] - \left(\int_0^1 \phi_j(x) f_X(x) dx\right)^2\right] \\
&= n^{-1} \left[\mathbb{E}\left[(\phi_j(X))^2\right] - \theta_j^2\right].
\end{aligned}$$

Com isso, temos que

$$\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} = n^{-1} \left[\mathbb{E}\left[(\phi_j(X))^2\right] - \theta_j^2\right].$$

Sabendo disso, pela igualdade trigonométrica elementar temos que

$$\cos^2(\alpha) = [1 + \cos(2\alpha)]/2.$$

e, podemos escrever

$$\begin{aligned}
\mathbb{E}\left[(\phi_j(X))^2\right] &= \mathbb{E}\left[\phi_j^2(X)\right] \\
&= \mathbb{E}\left[[\sqrt{2}\cos(\pi jX)]^2\right] \\
&= \mathbb{E}\left[2\cos^2(\pi jX)\right] \\
&= \mathbb{E}\left[2\left[\frac{1}{2} + \frac{\cos(2\pi jX)}{2}\right]\right] \\
&= \mathbb{E}[1 + \cos(2\pi jX)] \\
&= 1 + \mathbb{E}[\cos(2\pi jX)] \\
&= 1 + \frac{1}{\sqrt{2}}\mathbb{E}[\sqrt{2}\cos(2\pi jX)] \\
&= 1 + \frac{1}{\sqrt{2}}\mathbb{E}[\phi_{2j}(X)] \\
&= 1 + \theta_{2j}2^{-1/2}.
\end{aligned}$$

Logo temos que

$$Var[\hat{\theta}] = \mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} = n^{-1}[1 + \theta_{2j}2^{-1/2} - \theta_j^2].$$

Como $\theta_0 = 1$ temos que para $j > 0$

$$Var[\hat{\theta}] = n^{-1}\theta_0 + [\theta_{2j}2^{-1/2} - \theta_j^2]n^{-1} = d_j n^{-1}.$$

Como θ_j decai quando j aumenta, podemos escolher $\hat{d} = \theta_0 = 1$ como uma estimativa para todos d_j, $j = 0, 1, 2, \ldots$. Logo, temos que n^{-1} será nosso estimador para $Var[\hat{\theta}]$.

O segundo termo em (3.2) é o Viés Quadrático Integrado que aqui será abreviado por ISB (do inglês, *Integrated Squared Bias*). Para mais detalhes, veja o apêndice A.1

Essa segunda soma em (3.2) é difícil de ser estimada diretamente porque contêm muitos termos. Porém, pela identidade de Parseval (Apêndice A), temos que

$$\int_0^1 f^2(x)dx = \sum_{j=0}^{\infty}\theta_j^2 = \sum_{j=0}^{J}\theta_j^2 + \sum_{j=J+1}^{\infty}\theta_j^2.$$

Logo,

$$\sum_{j=J+1}^{\infty}\theta_j^2 = \int_0^1 f^2(x)dx - \sum_{j=0}^{J}\theta_j^2,$$

onde o termo $\int_0^1 f^2(x)dx$ é constante. O problema de encontrar um ponto de corte que minimiza (3.2) é equivalente a encontrar o J que minimiza a soma $\sum_{j=0}^{J}(n^{-1} - \theta_j^2)$. Aqui utilizamos n^{-1} como estimativa para $Var(\hat{\theta}_j)$. Precisamos escolher um estimador para θ_j^2. Temos que $\hat{\theta}_j$ é um estimador não viciado para θ_j, porém $\hat{\theta}_j^2$ é um estimador viciado para θ_j^2 porque $\mathbb{E}(\hat{\theta}_j^2) = \theta_j^2 + Var(\hat{\theta}_j)$. Como n^{-1} é um estimador para $Var(\hat{\theta}_j)$, temos que $\hat{\theta}_j^2 - n^{-1}$ é um estimador não viciado para θ_j^2.

Combinando essas ideias, concluímos que encontrar o J que minimiza o MISE é equivalente a encontrar o J que minimiza a seguinte expressão:

$$\hat{J} = \text{argmin}_{0 \leq J \leq J_n} \sum_{j=0}^{J} (2n^{-1} - \hat{\theta}_j^2). \tag{3.3}$$

Para seleção dos candidatos a ponto de corte, precisamos determinar um valor para J_n. Uma alternativa proposta por [Efromovich, 1999] para J_n é $J_n = \lfloor c_{J_0} + c_{J_1} \ln(n) \rfloor$, em que a função $f(x) = \lfloor x \rfloor$ retorna o maior número inteiro menor ou igual a x, e com relação a c_{J_0} e c_{J_1} usaremos os valores propostos por Efromovich em [Efromovich, 1999] que são 4 e 0.5 respectivamente .

Em muitos casos é necessário suavizar os estimadores $\hat{\theta}_j$ multiplicando-os por pesos $w_j \in [0, 1]$ que também devem ser estimados.

Vamos usar como estimadores os pesos w_j que minimizam o EQM (erro quadrático médio) dos estimadores $\hat{\theta}_j$. Esses pesos são definidos como

$$\begin{aligned} \mathbb{E}\{(w_j\hat{\theta}_j - \theta_j)^2\} &= w_j^2\mathbb{E}(\hat{\theta}_j^2) - 2w_j\theta_j^2 + \theta_j^2 \\ &= \mathbb{E}(\hat{\theta}_j^2)[w_j - \theta_j^2/\mathbb{E}(\hat{\theta}_j^2)]^2 + \theta_j^2[\mathbb{E}(\hat{\theta}_j^2) - \theta_j^2]/\mathbb{E}(\hat{\theta}_j^2). \end{aligned}$$

Como $\mathbb{E}(\hat{\theta}_j^2) = \theta_j^2 + Var(\hat{\theta}_j^2)$, temos que $\mathbb{E}\{(w_j\hat{\theta}_j - \theta_j)^2\}$ pode ser escrito como

$$(\theta_j^2 + Var(\hat{\theta}_j^2))[w_j - \theta_j^2/(\theta_j^2 + Var(\hat{\theta}_j^2))]^2 + \theta^2 Var(\hat{\theta}_j^2)/(Var(\hat{\theta}_j^2) + \theta^2)$$

Baseado nisso, temos que os pesos que minimizam o EQM são

$$w_j = \frac{\theta_j^2}{\theta_j^2 + Var(\theta_j^2)}.$$

Devemos portanto estimar esses pesos. Temos que θ_j^2 é não negativo e que não precisamos suavizar $\hat{\theta}_0$ pois como o suporte de f é $[0, 1]$ temos que $\hat{\theta}_0 = 1$ fazendo com que o estimador de series integre 1. Logo, $\hat{w}_0 := 1$ e $\hat{w}_j = (1 - 1/n\hat{\theta}_j^2)_+$ para $j > 0$, onde $(x)_+ = \max(0, x)$.

Isso nos dá o seguinte estimador para a densidade de interesse [Efromovich, 1999]

$$\tilde{f}(x) = \sum_{j=0}^{\hat{J}} \hat{w}_j\hat{\theta}_j\phi_j(x).$$

Em [Efromovich, 1999] o autor propõe duas melhorias para esse estimador. A primeira é acrescentar termos de alta frequência a fim de melhoras as estimativas para densidades não homogêneas. O novo estimador será:

$$\tilde{f}(x) = \sum_{j=0}^{\hat{J}} \hat{w}_j \hat{\theta}_j \phi_j(x) + \sum_{j=\hat{J}+1}^{c_{JM} J_n} I_{\{\hat{\theta}_j^2 > c_T \hat{d} \ln(n)/n\}} \hat{\theta}_j \phi_j(x). \tag{3.4}$$

Nessa definição, para c_{JM} e c_T vamos utilizar novamente os valores sugeridos Efromovich em [Efromovich, 1999] que são 6 e 4, respectivamente. Esses termos de alta frequência são inclusos apenas nos casos em que o coeficiente de Fourier correspondente é grande, o que não prejudica a estimativa de funções suaves como a densidade normal.

A necessidade da segunda melhoria é que a estimativa de f pode assumir valores negativos. Para isso precisamos encontrar uma projeção de $\tilde{f} \in \mathbb{L}_2[0,1]$, onde $\mathbb{L}_2[0,1]$,é o espaço das funções quadrado integráveis no intervalo $[0,1]$, isso quer dizer que, se $f \in \mathbb{L}_2[0,1]$ temos que $\int_0^1 f^2(x)dx < \infty$. Essa projeção será

$$\hat{f}(x) = (\tilde{f}(x))_+.$$

A seguir vamos ilustrar essa metodologia a partir de um conjunto de dados reais

3.1 Exemplo: Velocidade das Galáxias

Vamos aplicar a metodologia acima mencionada a um conjunto de dados reais sobre astronomia. Esse conjunto foi apresentado em [Roeder, 1990] onde o autor propõe um estimador via núcleo estimador para densidades e utiliza esses dados para exemplificar seus resultados teóricos. Vamos fazer aqui uma breve descrição dos dados.

De acordo com a teoria do Big Bang, a matéria no universo se expandiu muito rapidamente e as forças gravitacionais causaram a formação de galáxias. Os astrônomos especulam que a atração gravitacional levou ao agrupamento de galáxias e pesquisas indicam a presença de conglomerados de galáxias cercados por grandes vazios. A velocidade relativa entre nossa galáxia e outras é estimada pelo desvio para o vermelho no espectro de luz de uma maneira similar a como o efeito Doppler mede as mudanças na velocidade através de mudanças no som. Pelo paradigma da expansão do universo as galáxias mais distantes da nossa devem estar se movendo com maior velocidade, porque velocidade e distâncias são proporcionais. Se as galáxias estão agrupadas, as velocidades devem ter uma distribuição multimodal, cada moda correspondendo a um cluster

Na região da Coroa Boreal foram medidas as velocidades de 82 galáxias com um erro relativo de medida menor que 0,5%. Os dados estão disponíveis no Apêndice C.

Aplicamos a metodologia descrita nesse capítulo a esse conjunto de dados, e a estimativa é apresentada na figura a seguir,

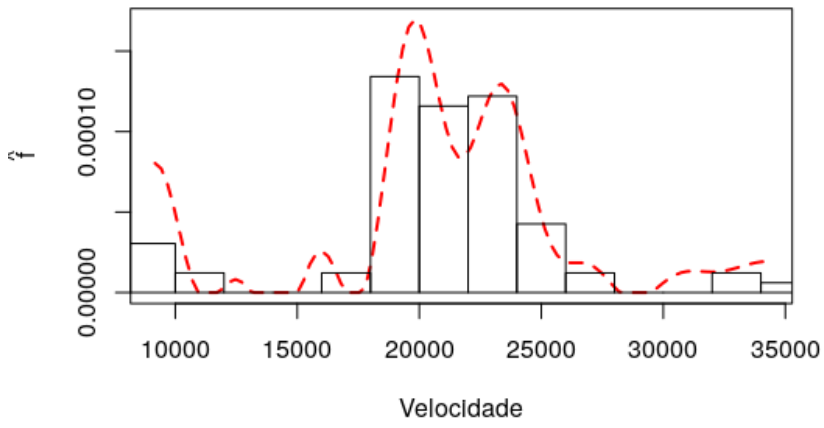

Figura 1 – Estimativa da densidade via séries de Fourier para os dados da Tabela 2, a velocidade esta em km/segundos.

Se cada moda representa um conglomerado de galáxias nessa estimativa observamos dois grupos grandes de galáxias com velocidade em torno de 20000km/s e 23000 km/s e outros grupos menores com velocidade em torno de 10000km/s e 30000km/s.

O $\hat{J}$ estimado para esse conjunto de dados foi de 15, logo estimamos 16 coeficientes e Fourier e 16 pesos pois nesse caso não foram necessários termos de alta, os valores dessas estimativas são apresentados na Tabela 1

No próximo capítulo, vamos adaptar esses conceitos para o caso de dados viesados.

No Apêndice E é apresentada a rotina em R [R Core Team, 2019] que utilizamos para computar essas estimativas e construir esse gráfico.

j	$\hat{\theta}_j$	$\hat{w}_j$
0	1.00000000	1.0000000
1	0.11864906	0.1337216
3	-0.89541321	0.9847897
4	-0.05672641	0.0000000
5	0.80895250	0.9813645
6	0.25370654	0.8105377
7	-0.13853049	0.3645295
8	-0.09495707	0.0000000
9	0.09318180	0.0000000
10	0.22267971	0.7540624
11	0.29751292	0.8622237
12	0.05268247	0.0000000
13	-0.30036786	0.8648303
14	-0.02176303	0.0000000
15	0.39350185	0.9212424
16	0.23123928	0.7719327

Tabela 1 – Estimativas dos coeficientes de Fourier e dos pesos dados galaxias

4 Estimação de densidades de dados viesados via bases de cossenos

Recapitulando, suponha que estamos interessados em estimar a densidade de uma variável aleatória X, denotada por f_X, através de observações indiretas de uma variável aleatória Y com suporte $[0,1]$ e densidade f_Y. Com base na equação (1.1), temos que essa densidade de interesse pode ser escrita como

$$f_X(x) = \frac{\mu f_Y(x)}{g(x)}. \tag{4.1}$$

Nessa definição, $g(x)$ corresponde a função de viés e μ é definido como

$$\mu = \int_0^1 g(x) f_X(x) dx = \mathbb{E}[g(X)].$$

O objetivo é estimar a densidade de interesse via séries trigonométricas, para isso vamos considerar a base de cossenos $\{\phi_0(x) = 1, \phi_j(x) = \sqrt{2}\cos(\pi j x), j = 1, 2, \ldots\}$ e estimar os coeficientes de Fourier $\theta_j = \int_0^1 \phi_j(x) f_X(x) dx$ da seguinte maneira

$$\begin{aligned}
\theta_j &= \int_0^1 \phi_j(x) f_X(x) dx \\
&= \int_0^1 \frac{\mu \phi_j(x) f_Y(x)}{g(x)} dx \\
&= \mu \int_0^1 \frac{\phi_j(x) f_Y(x)}{g(x)} dx \\
&= \mu \mathbb{E}\left[\frac{\phi_j(Y)}{g(Y)}\right].
\end{aligned}$$

Relacionando os momentos amostrais e populacionais, podemos considerar o estimador de momentos

$$\hat{\theta}_j = \mu n^{-1} \sum_{l=1}^{n} \frac{\phi_j(Y_l)}{g(Y_l)}. \tag{4.2}$$

Vamos verificar agora que (4.2) é um estimador não viciado para θ_j

$$\mathbb{E}\left[\hat{\theta}_j\right] = \mu n^{-1} \mathbb{E}\left[\sum_{l=1}^{n} \frac{\phi_j(Y_l)}{g(Y_l)}\right].$$

Supondo que as variáveis sejam independentes e identicamente distribuídas (iid), temos que

$$\begin{aligned}
\mu n^{-1}\mathbb{E}\left[\sum_{l=1}^{n}\frac{\phi_j(Y_l)}{g(Y_l)}\right] &= \mu n^{-1} n\mathbb{E}\left[\frac{\phi_j(Y)}{g(Y)}\right] \\
&= \mu\mathbb{E}\left[\frac{\phi_j(Y)}{g(Y)}\right] \\
&= \mu\int_0^1 \frac{\phi_j(y)f_Y(y)}{g(y)}dy \\
&= \mu\int_0^1 \frac{\phi_j(y)f_X(y)g(y)}{g(y)\mu}dy \\
&= \int_0^1 \phi_j(y)f_X(y)dy \\
&= \theta_j.
\end{aligned}$$

Em situações práticas, μ é desconhecido e deve ser estimado. Um estimador proposto em [Ramirez e Vidakovic, 2010] e [Efromovich, 1999] é dado por:

$$\hat{\mu} =: \frac{1}{n^{-1}\sum_{i=1}^{n} g^{-1}(Y_i)},$$

em que $g^{-k}(\cdot) = 1/g(\cdot)^k$. Nesses trabalhos os autores defendem esse estimador par μ pois para uma a amostra é iid $\frac{1}{\hat{\mu}}$ é um estimador não viesado para $\frac{1}{\mu}$, pois

$$\mathbb{E}\left(\frac{1}{\hat{\mu}}\right) = \mathbb{E}\left(g^{-1}(Y)\right) = \mu^{-1}\int_0^1 g(y)f_X(y)g^{-1}(y)dy = \frac{1}{\mu}.$$

Assim como no estimador 3.4 para o estimador de dados viesados também precisamos de um estimador para d, digamos $\hat{d}$, que nesse contexto, é dado por $n\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\}$. Como $\hat{\theta}_j$ é um estimador não viciado para θ_j, temos que

$$\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} = Var(\hat{\theta}_j).$$

Seja $r_j(Y_l) = \frac{\phi_j(Y_l)}{g(Y_l)}$, sabendo que a amostra é iid temos que

$$\begin{aligned}
Var(\hat{\theta}_j) &= Var(\mu n^{-1} \sum_{l=1}^{n} r_j(Y_l)) \\
&= \mu^2 n^{-2} Var(\sum_{l=1}^{n} r_j(Y_l)) \\
&= \mu^2 n^{-2} n Var(r_j(Y)) \\
&= \mu^2 n^{-1} Var(r_j(Y)) \\
&= \mu^2 n^{-1} [\mathbb{E}[r_j(Y)^2] - (\mathbb{E}[r_j(Y)])^2] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \left(\mathbb{E}\left[\frac{\phi_j(Y)}{g(Y)} \right] \right)^2 \right] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \left(\int_0^1 \frac{\phi_j(y) f_Y(y)}{g(y)} \right)^2 \right] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \left(\int_0^1 \frac{\phi_j(y) f_X(y) g(y)}{g(y)\mu} dy \right)^2 \right] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \left(\frac{1}{\mu} \int_0^1 \phi_j(y) f_X(y) dy \right)^2 \right] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \left(\frac{\theta_j}{\mu} \right)^2 \right] \\
&= \mu^2 n^{-1} \left[\mathbb{E}\left[\left(\frac{\phi_j(Y)}{g(Y)} \right)^2 \right] - \frac{\theta_j^2}{\mu^2} \right] \\
&= n^{-1} \left[\mathbb{E}\left[\frac{\mu^2 \phi_j(Y)^2}{g(Y)^2} \right] - \theta_j^2 \right] \\
&= n^{-1} \left[\mathbb{E}\left[\left(\frac{\mu \phi_j(Y)}{g(Y)} \right)^2 \right] - \theta_j^2 \right].
\end{aligned}$$

Logo, temos que

$$\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} = n^{-1} \left[\mathbb{E}\left[\left(\frac{\mu \phi_j(Y)}{g(Y)} \right)^2 \right] - \theta_j^2 \right].$$

Portanto,

$$n\mathbb{E}\{(\hat{\theta}_j - \theta_j)^2\} = \mathbb{E}\left[\left(\frac{\mu \phi_j(Y)}{g(Y)} \right)^2 \right] - \theta_j^2.$$

Pela igualdade trigonométrica elementar, temos que

$$\cos^2(\alpha) = [1 + \cos(2\alpha)]/2.$$

Sabendo disso, podemos escrever

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mu\phi_j(Y)}{g(Y)}\right)^2\right] &= \mathbb{E}\left[\frac{\mu^2\phi_j^2(Y)}{g^2(Y)}\right] \\
&= \mu^2\mathbb{E}\left[\frac{1}{g^2(Y)}\phi_j^2(Y)\right] \\
&= \mu^2\mathbb{E}\left[\frac{1}{g^2(Y)}[\sqrt{2}\cos(\pi jY)]^2\right] \\
&= \mu^2\mathbb{E}\left[\frac{2}{g^2(Y)}[\cos(\pi jY)]^2\right] \\
&= \mu^2\mathbb{E}\left[\frac{2}{g^2(Y)}\left[\frac{1}{2}+\frac{\cos(2\pi jY)}{2}\right]\right] \\
&= \mu^2\mathbb{E}[g^{-2}(Y)+\cos(2\pi jY)g^{-2}(Y)] \\
&= \mu^2\mathbb{E}[g^{-2}(Y)]+\mu^2\mathbb{E}[\cos(2\pi jY)g^{-2}(Y)] \\
&= \mu^2\mathbb{E}[g^{-2}(Y)]+\frac{\mu^2}{\sqrt{2}}\mathbb{E}[\sqrt{2}\cos(2\pi jY)g^{-2}(Y)] \\
&= \mu^2\mathbb{E}[g^{-2}(Y)]+\frac{\mu^2}{\sqrt{2}}\mathbb{E}[\phi_{2j}(Y)g^{-2}(Y)] \\
&= \mu^2\mathbb{E}[g^{-2}(Y)]+\frac{\mu^2}{\sqrt{2}}\int_0^1 f_Y(y)\phi_{2j}(y)g^{-2}(y)dy.
\end{aligned} \tag{4.3}$$

Podemos escrever o segundo termo da direita de (4.3) como

$$\frac{\mu^2}{\sqrt{2}}\int_0^1 f_Y(y)\phi_{2j}(y)g^{-2}(y)dy = \mu^2\int_0^1 f_Y(y)\cos(2\pi jy)g^{-2}(y)dy.$$

Pela regra da cadeia, temos que

$$\frac{d\ \cos(2j\pi y)}{dy} = -2j\pi\text{sen}(2j\pi y), \qquad \frac{d\ \text{sen}(2j\pi y)}{dy} = 2j\pi\cos(2j\pi y).$$

Utilizando esses resultados, temos que

$$\mu^2\int_0^1 f_Y(y)g^{-2}(y)\cos(2\pi jy)dy = \frac{\mu^2}{2j\pi}\int_0^1 f_Y(y)g^{-2}(y)d\ \text{sen}(2j\pi y). \tag{4.4}$$

Vamos calcular a integral obtida em (4.4) usando integração por partes. A fórmula da integral por partes é dada por

$$\int_0^1 udv = uv|_0^1 - \int_0^1 vdu.$$

Suponha que $f_Y(y)g^{-2}(y)$ seja diferenciável, denote $u = f_Y(y)g^{-2}(y)$ e $dv = d\left(\frac{\mu^2\text{sen}(2j\pi y)}{2j\pi}\right)$. Assim,

$$v = \frac{\mu^2\text{sen}(2j\pi y)}{2j\pi}, \qquad du = (f_Y(y)g^{-2}(y))'dy.$$

Assim,

$$uv|_0^1 = \frac{\mu^2}{2j\pi} f_Y(y)g^{-2}(y)\text{sen}(2j\pi y)|_0^1 = \frac{\mu^2}{2j\pi}[f_Y(0)g^{-2}(0)\text{sen}(0) - f_Y(1)g^{-2}(1)\text{sen}(2j\pi)].$$

Como $\text{sen}(\pi j) = 0$ para todo j inteiro, então

$$\frac{\mu^2}{2j\pi}\int_0^1 f_Y(y)g^{-2}(y)d\,\text{sen}(2j\pi y) = -\int_0^1 vdu = -\frac{\mu^2}{2j\pi}\int_0^1 \text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'dy.$$

Uma vez que o modulo é uma função convexa, temos que

$$|\int_0^1 \text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'dy| \leq \int_0^1 |\text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'|dy.$$

Como $0 \leq |\text{sen}(2j\pi y)| \leq 1$ para todo $y \in \mathbb{R}$ então temos que $|\int_0^1 \text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'dy| \leq \int_0^1 |(f_Y(y)g^{-2}(y))'|dy$. Com isso, concluímos que, se

$$f_Y(y)g^{-2}(y) \text{ é diferenciável} \quad \text{e} \quad \int_0^1 |(f_Y(y)g^{-2}(y))'|dy < \infty, \tag{4.5}$$

então $\int_0^1 \text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'dy$ será limitado. Como $\lim_{j\to\infty} \frac{-\mu^2}{2j\pi} = 0$, temos que

$$\lim_{j\to\infty} \frac{\mu^2}{2j\pi}\int_0^1 \text{sen}(2j\pi y)(f_Y(y)g^{-2}(y))'dy = 0.$$

Portanto, supondo que valem as condições estabelecidas em (4.5), para um j suficientemente grande, podemos reescrever o coeficiente de dificuldade para dados viesados como

$$\begin{aligned} d &:= \mu\int_0^1 f_X(y)g^{-1}(y)dy \\ &= \mu\int_0^1 \mu f_Y(y)g^{-2}(y)dy \\ &= \mu^2\int_0^1 f_Y(y)g^{-2}(y)dy \\ &= \mu^2\mathbb{E}\left[\frac{1}{g^2(Y)}\right] \\ &= \mu^2\mathbb{E}[g^{-2}(Y)]. \end{aligned}$$

Então, um estimador natural para o coeficiente de dificuldade será.

$$\hat{d} = \hat{\mu}^2 n^{-1}\sum_{l=1}^{n} g^{-2}(Y_l).$$

O estimador de f_X fica definido como

$$\hat{f}_X(y) = \sum_{j=0}^{\hat{J}} \hat{w}_j\hat{\theta}_j\phi_j(y) + \sum_{j=\hat{J}+1}^{c_{JM}J_n} I_{\{\hat{\theta}_j^2 > c_T\hat{d}\ln(n)/n\}}\hat{\theta}_j\phi_j(y).$$

No próximo capítulo, vamos fazer um breve estudo de simulação no R [R Core Team, 2019].

5 Simulação de dados viesados

Vamos ilustrar a metodologia com dois exemplos de amostras viesadas que foram geradas utilizando o algoritmo de aceitação-rejeição que e descrito no Apêndice B.3 e apresentado em código no Apêndice E. Esses exemplos foram propostos em [Ramirez e Vidakovic, 2010], onde os autores apresentam um estimador de ondaletas para dados viesados em amostras estratificadas.

Exemplo 1. Nesse exemplo foi considerada uma densidade $X \sim Beta(2, 2)$ e a partir dela geramos amostras viesadas considerando a função de viés $g(x) = 0.1 + 0.9x$ e $J = 7$ como proposto em [Ramirez e Vidakovic, 2010]. Vale ressaltar que essa função de viés é defendida por vários autores da área, em [Efromovich, 1999] por exemplo o autor advoga que essa função auxilia na obtenção de boas estimativas exceto em casos patológicos, nesse livro o autor apresenta alguns exemplos de caso problemáticos e diz que para esses casos podem ser feitos experimentos adicionais a fim se obter uma função de viés que produza resultados satisfatórios. Como fizemos várias simulações não apresentamos aqui as estimativas para ϕ_j e w_j como fizemos no exemplo das galáxias, porém essas estimativas podem ser obtidas facilmente no código disponível no Apêndice E. As estimativas das densidades para amostras de tamanho 100, 500, 1000 e 10000 são apresentadas nas figuras a seguir:

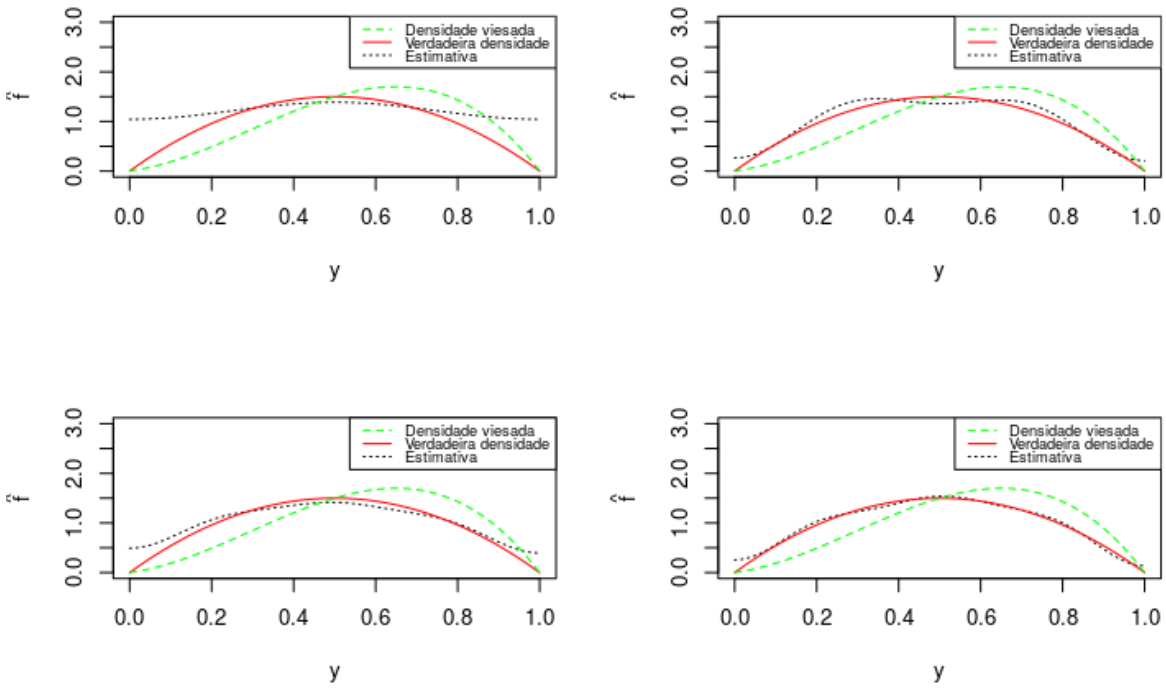

Figura 2 – Simulações para densidades beta com amostras de tamanho 100, 500, 1000 e 10000

Analisando essas simulações percebemos que a metodologia consegue corrigir o viés da amostra e nos dá uma estimava satisfatória para a verdadeira densidade, percebemos também, como era de se esperar, que quanto maior a amostra mais a densidade estimada se aproxima da verdadeira densidade.

Exemplo 2. Nesse exemplo consideramos uma densidade linear definida por partes [Stanley, 2004] da seguinte maneira:

$$f_X(y) = \begin{cases} 8y & y \in [0, \frac{1}{4}), \\ 4(1-2y) & y \in [\frac{1}{4}, \frac{1}{2}), \\ 8y-4 & y \in [\frac{1}{,2}, \frac{3}{4}), \\ 8-8y & y \in [\frac{3}{4}, 1]. \end{cases}$$

Novamente vamos utilizar a função de viés $g(x) = 0.1 + 0.9x$ proposta por [Efromovich, 1999]. As estimativas para amostras de tamanho 100, 500, 1000 e 10000 são apresentadas a seguir:

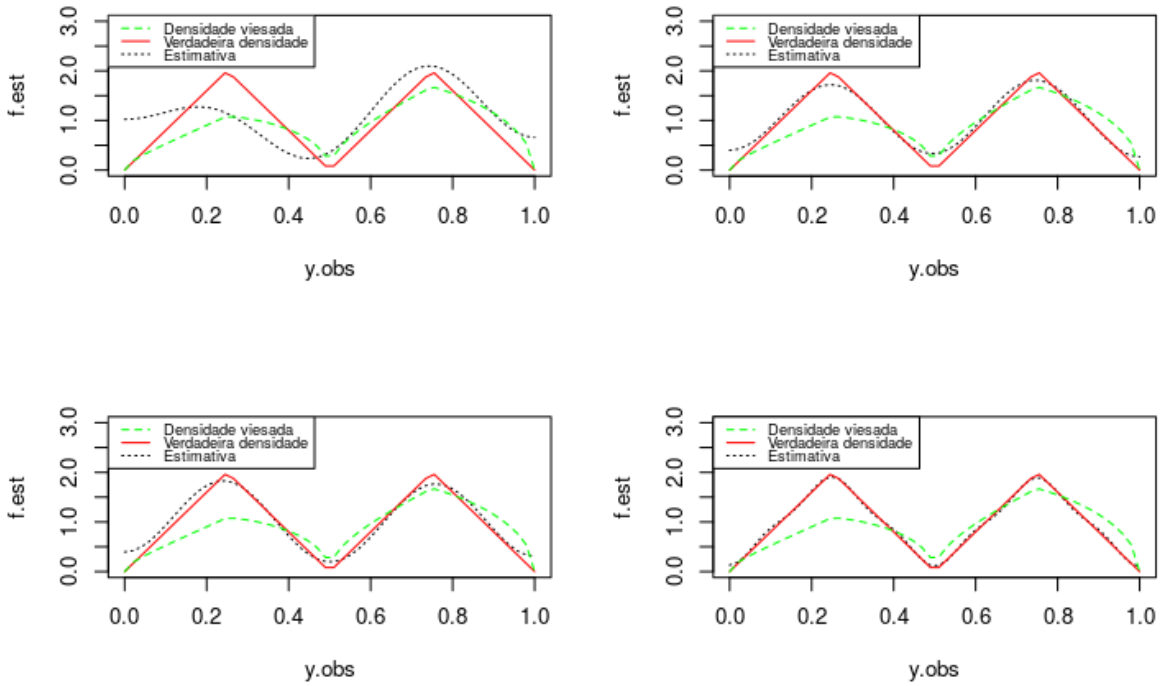

Figura 3 – Simulações para densidade linear definida por partes com amostras de tamanho 100, 500, 1000 e 10000

Analisando essas simulações percebemos novamente que a metodologia consegue corrigir bem o viés da amostra e nos dá uma estimava satisfatória para a verdadeira densidade.Porém como a função tem pontos onde ela não é diferenciável, são necessárias amostras maiores para que a estimativa convirja para a verdadeira densidade, diferente do primeiro exemplo em que para amostra de tamanho 100 a estimativa já é bem satisfatória.

No Capítulo a seguir vamos aplicar essa metodologia a um conjunto de dados reais.

6 Aplicação

Nesta seção, aplicamos a metodologia proposta a um conjunto de dados reais. Os dados são referentes à concentração de álcool no sangue de motoristas do sexo masculino envolvidos em acidentes fatais no ano de 1975, nos Estados Unidos. Essas observações são disponibilizadas pelo *"National Highway Traffic Safety Administration Department of Transportation"*. O conjunto de dados faz parte do Sistema de Relatórios de Análise de Fatalidades (FARS - do inglês, *Fatality Analysis Reporting System*), uma base de dados contendo todos os registros de acidentes fatais desde 1975 nos 50 estados americanos, Distrito de Columbia e Porto Rico. A variável de interesse será, portanto, a concentração de álcool no sangue (CAS), uma variável contínua expressa em grama/100ml de sangue. Para esse trabalho utilizamos uma amostra de tamanho 758 dessa base de dados, a amostra usada é apresentada no Apêndice D.

Nessa aplicação o J ótimo calculado (que minimza o MISE) foi de 24 e foram estimados 223 coeficientes e 223 pesos. O histograma dos dados é apresentado na Figura 4. Observe, por exemplo, que o histograma possui uma única moda por volta de 0,15 grama/100ml, indicando uma maior concentração de motoristas com CAS entre, aproximadamente, 0,1 grama/100ml e 0,2 grama/100ml. Isso nos induz a imaginar que houve mais motoristas com a CAS supracitada do que motoristas que não beberam (0 grama/100ml), o que é uma conclusão difícil de se acreditar que ocorra na prática. Isso decorre do fato da amostra ser viesada, pelos motivos já discutidos anteriormente.

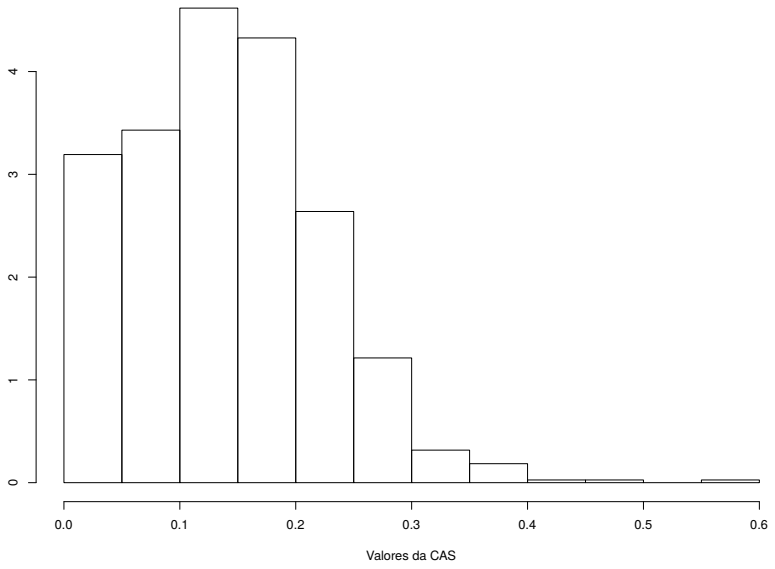

Figura 4 – Histograma da concentração de álcool no sangue

Para corrigir o viés da amostra, aplicamos a metodologia proposta aos dados,

novamente considerando a função de viés $g(x) = 0.1 + 0.9x$ proposta por Efromovich em [Efromovich, 1999]. A estimativa da densidade é apresentada na Figura 5.

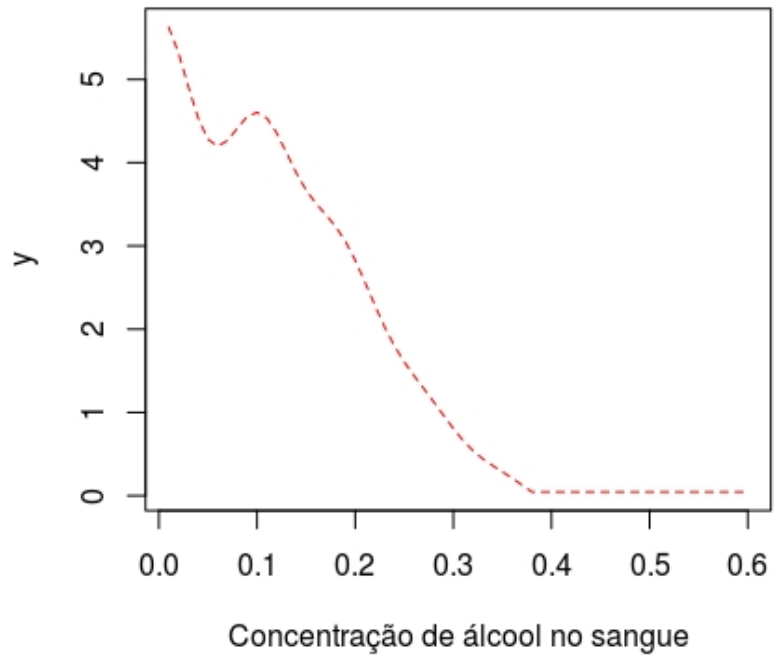

Figura 5 – Histograma da concentração de álcool no sangue

Ao analisar o resultado na Figura 5, observamos indicativos de que a distribuição da concentração de álcool no sangue dos motoristas aparenta ser bimodal, com um maior pico próximo de 0 grama/100ml e outro pico menor próximo de 0,1 grama/100ml . Se compararmos a estimativa com o histograma apresentado na Figura 4, vamos perceber que a amostra (como já era esperado) nos dá uma ideia errada sobre o comportamento da CAS. A metodologia proposta corrige, portanto, o viés que induzia à conclusão de que havia mais motoristas com CAS de 0,15 grama/100ml do que motoristas que não beberam (0 grama/100ml). Isso reflete a importância de se considerar o viés inerente aos dados no processo de estimação de sua densidade.

7 Conclusões e Comentários Adicionais

Nesse trabalho, apresentamos um estimador de densidades não paramétrico utilizando bases de cossenos. Para verificar o comportamento desse estimador fizemos uma aplicação a um conjunto de dados relativo à velocidade de 82 galáxias medidas na região da Coroa Boreal. Analisando essa estimativa percebemos que nessa região existem dois grandes grupos de galáxias que se movem com velocidade de aproximadamente 20000km/s e 23000 km/s e alguns grupos menores que se movem com outras velocidades.

Posteriormente adaptamos esse estimador para o contexto de dados viesados e avaliamos os resultados desse método. Primeiro, em um estudo de simulação no qual amostras foram geradas com o auxilio de um algoritmo de aceitação-rejeição, e posteriormente em uma amostra real, retirada de um conjunto de dados que corresponde a concentração de álcool no sangue de motoristas nos 50 estados americanos, no distrito de Columbia e em Porto Rico desde 1975. O método se mostrou eficiente na correção do viés tanto no estudo de simulação como no conjunto de dados reais.

Comparando o caso geral (para dados não viesados) e o caso em que a amostra é viesada, percebemos que o caso geral é um caso particular do modelo para dados viesados em que a função de viés é igual a 1 para todo espaço amostral. Isso explica porque a metodologia pode ser aplicada tanto para dados não viesados (como no exemplo das galáxias) quanto para dados viesados (como no exemplo da concentração de álcool no sangue).

Como sugestões para trabalhos futuros podem ser consideradas simulações de Monte Carlo, ou ainda a utilização de outras bases ortonormais como, base de polinômios, Haar ou Ondaletas. Além disso podem ser considerados outros métodos de estimação como núcleo estimador e algorítimo EM por exemplo.

REFERÊNCIAS

[Laake et al., 1993] BUCKLAND, S. T.; ANDERSON, D. R.; BURNHAM, K. P.; LAAKE, J. L. **Distance Sampling**. Dordrecht: Springer Netherlands, 1993.

[Cook e Martin, 1974] COOK, R. D.; MARTIN, F. B. A model for quadrat sampling with "visibility bias". **Journal of the American Statistical Association**, v. 69, n. 346, p. 345–349, 1974.

[Cox, 1969] COX, D. R. Some sampling problems in technology. **New Developments in Survey Sampling (N. L. Johnson and H. Smith, Jr., eds.)**, p. 506–527. Wiley, New York. 1969.

[De Uña Álvarez e Rodríguez Casal, 2007] DE UÑA ÁlVAREZ, J.; RODRÍGUEZ-CASAL, A. Nonparametric estimation from length-biased data under competing risks. **Computational Statistics and Data Analysis** v. 51, n. 5, p. 2653–2669, 2007.

[Efromovich, 1999] EFROMOVICH, Sam. **Nonparametric Curve Estimation**: Methods, Theory and Applications. New York: Springer, 1999.

[Efromovich, 2004] EFROMOVICH, Sam. Distribution estimation for biased data. **Journal of Statistical Planning and Inference**. v. 124, n. 1, p. 1–43, 2004.

[Feller, 1971] FELLER, W. **Introduction to Probability Theory and Its Applications.** Second edition. New York: Wiley, 1971.

[Jones, 1991] JONES, M. C. Kernel density estimation for length biased data. **Biometrika** v. 78, n. 3, p. 511, 1991.

[Ramirez e Vidakovic, 2010] Ramirez, P.; Vidakovic, B. Wavelet density estimation for stratified size-biased sample. **Journal of Statistical Planning and Inference.** v. 2, n.2, p. 419-432. 2010.

[R Core Team, 2019] R CORE TEAM. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[Roeder, 1990] ROEDER, Kathryn. Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. **Journal of the American Statistical Association**, v. 85, n. 411, p. 617–624, 1990.

[Ross, 1993] ROSS, S. M. **Stochastic processes**. New York: Wiley, 1983.

[Stanley, 2004] STANLEY, William D. **Technical Analysis And Applications With Matlab**. Cengage Learning. p. 143, 2004.

[Vardi, 1982] VARDI, Y. Nonparametric estimation in the presence of length bias. **The Annals of Statistics**, v. 10, n.2, p. 616–620, 1982.

APÊNDICE A – Identidade De Parseval

‘ Seja $\{\phi_j\}$ uma base ortonormal em $\mathbb{L}_2$ e seja $\theta_j = \langle f, \phi_j \rangle = \int_0^1 f(x)\phi_j(x)dx$ o j-ésimo coeficiente de Fourier de $f \in \mathbb{L}_2[0,1]$. Como já foi visto, podemos escrever f pela expansão em séries da seguinte forma

$$f(x) = \sum_{j=0}^{\infty} \theta_j \phi_j(x).$$

A norma $\mathbb{L}_2$ é dada por

$$||f||^2 = \int_0^1 f^2(x)dx.$$

Como $f \in \mathbb{L}_2$ então essa norma é finita. Usando a expansão em séries apresentada em (2.1),

$$\begin{aligned} \int_0^1 f^2(x)dx &= \int_0^1 \left[\sum_{j=0}^{\infty} \theta_j \phi_j(x)\right]^2 dx \\ &= \int_0^1 \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \theta_j\theta_k\phi_j(x)\phi_k(x)dx \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \theta_j\theta_k \int_0^1 \phi_j(x)\phi_k(x)dx. \end{aligned} \tag{A.1}$$

Como base é ortonormal, então

$$\int_0^1 \phi_j(x)\phi_k(x)dx = \begin{cases} 0, \text{se } k = j \\ 1, \text{caso contrário} \end{cases}$$

Sabendo disso, concluímos que

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \theta_j\theta_k \int_0^1 \phi_j(x)\phi_k(x)dx = \sum_{j=0}^{\infty} \theta_j^2.$$

Logo, a norma $\mathbb{L}_2$ de uma função $f \in \mathbb{L}_2$, pode ser escrita da seguinte forma

$$||f||^2 = \sum_{j=0}^{\infty} \theta_j^2. \tag{A.2}$$

Essa igualdade é chamada de identidade de Parseval.

A.1 Calculo do ISB via identidade de Parseval

O viés quadrático integrado (*integrated squared bias*), que vamos abreviar como ISB, é dado por

$$\text{ISB}_J(f) = \int_0^1 (f(x) - f_J(x))^2 dx, \tag{A.3}$$

em que $f(x)$ é a soma infinita definida em (2.1) e $f_J(x)$ é a soma truncada definida em (2.2).

Podemos reescrever a equação (A.3) da seguinte forma

$$\begin{aligned}\int_0^1 (f(x) - f_J(x))^2 dx &= \int_0^1 \left(\sum_{j=0}^{J} \theta_j \phi_j(x) - \sum_{j=0}^{\infty} \theta_j \phi_j(x) \right)^2 dx \\ &= \int_0^1 \left(\sum_{j=0}^{J} \theta_j \phi_j(x) - \sum_{j=0}^{J} \theta_j \phi_j(x) - \sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right)^2 dx \\ &= \int_0^1 \left(\sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right)^2 dx\end{aligned}$$

Pela identidade de Parseval,

$$\begin{aligned}\int_0^1 (f(x) - f_J(x))^2 dx &= \int_0^1 \left(\sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right)^2 dx \\ &= \sum_{j=J+1}^{\infty} \theta_j^2\end{aligned}$$

A.2 Calculo do MISE via identidade de Parseval

O erro quadrático médio integrado (*mean integrated squared error*), que vamos abreviar como MISE, é uma medida usada para avaliar a bondade de ajuste de um estimador $\tilde{f}_J(x)$ de uma densidade $f \in \mathbb{L}_2[0,1]$. O MISE é dado por

$$\text{MISE}\{\tilde{f}_J, f\} := \mathbb{E}\left\{ \int_0^1 (\tilde{f}_J(x) - f(x))^2 dx \right\}. \tag{A.4}$$

Seja $\tilde{f}_J(x) = \sum_{j=0}^{J} \hat{\theta}_j \phi_j(x)$ e $f(x) = \sum_{j=0}^{\infty} \theta_j \phi_j(x)$, o objetivo e escolher o J que minimiza o MISE. Assim, precisamos reescrever o MISE de uma forma que facilite o cálculo desse J. Note que podemos reescrever a equação (A.4) da seguinte forma

$$\begin{aligned}&\mathbb{E}\left\{ \int_0^1 (\tilde{f}_J(x) - f(x))^2 dx \right\} = \mathbb{E}\left\{ \int_0^1 \left(\sum_{j=0}^{J} \hat{\theta}_j \phi_j(x) - \sum_{j=0}^{\infty} \theta_j \phi_j(x) \right)^2 dx \right\} \\ &= \mathbb{E}\left\{ \int_0^1 \left(\sum_{j=0}^{J} \hat{\theta}_j \phi_j(x) - \sum_{j=0}^{J} \theta_j \phi_j(x) - \sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right)^2 dx \right\} \\ &= \mathbb{E}\left\{ \int_0^1 \left(\sum_{j=0}^{J} (\hat{\theta}_j - \theta_j) \phi_j(x) - \sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right)^2 dx \right\} \\ &= \mathbb{E}\left\{ \int_0^1 \left[\sum_{j=0}^{J} (\hat{\theta}_j - \theta_j) \phi_j(x) \right]^2 + \left[\sum_{j=J+1}^{\infty} \theta_j \phi_j(x) \right]^2 - \left[2 \sum_{j=0}^{J} \sum_{k=J+1}^{\infty} (\hat{\theta}_j - \theta_j) \theta_j \phi_j(x) \phi_k(x) \right] dx \right\}.\end{aligned}$$

Como a base é ortonormal então $\int_0^1 (\sum_{j=0}^{J} \sum_{k=J+1}^{\infty} (\hat{\theta}_j - \theta_j)\theta_j \phi_j(x)\phi_k(x))dx = 0$. Portanto,

$$\mathbb{E}\left\{\int_0^1 (\tilde{f}_J(x) - f(x))^2 dx\right\} = \mathbb{E}\left\{\int_0^1 \left[\sum_{j=0}^{J} (\hat{\theta}_j - \theta_j)\phi_j(x)\right]^2 dx\right\} + \mathbb{E}\left\{\int_0^1 \left[\sum_{j=J+1}^{\infty} \theta_j \phi_j(x)\right]^2 dx\right\}.$$

Pela identidade de Parseval, temos que

$$\begin{aligned}\mathbb{E}\left\{\int_0^1 (\tilde{f}_J(x) - f(x))^2 dx\right\} &= \mathbb{E}\left\{\sum_{j=0}^{J} (\hat{\theta}_j - \theta_j)^2\right\} + \mathbb{E}\left\{\sum_{j=J+1}^{\infty} \theta_j^2\right\} \\ &= \sum_{j=0}^{J} \mathbb{E}(\hat{\theta}_j - \theta_j)^2 + \sum_{j=J+1}^{\infty} \theta_j^2.\end{aligned}$$

APÊNDICE B – Geração de números aleatórios

B.1 Método da Transformação Integral da Probabilidade

Suponha que estamos interessados em gerar observações de uma variável aleatória W com função distribuição F_W e com função densidade de probabilidade f_W. Seja $U = F_W(W)$ e F_U a função distribuição acumulada de U. Temos que

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(F_W(W) \leq u) \\ &= P(W \leq F_W^{-1}(u)) \\ &= F_W(F_W^{-1}(u)) \\ &= u. \end{aligned}$$

Como $F_U(u) = u$ apenas no caso da distribuição Uniforme $(0, 1)$, então concluímos que $U = F_W(W)$ segue uma distribuição Uniforme $(0, 1)$. Podemos escrever a variável aleatória W como $W = F_W^{-1}(U)$, ou seja, pelo método da transformação integral da probabilidade, podemos gerar uma observação de uma variável aleatória W por meio da inversa da sua função de distribuição acumulada (F_W^{-1}) através do seguinte algoritmo.

B.1.1 Algoritmo do método da transformação integral da probabilidade

O algoritmo do método da transformação integral da probabilidade é apresentado nos passos abaixo.

Passo 1 Calcule a função F_W^{-1}

Passo 2 Gere uma observação U com distribuição Uniforme$(0, 1)$, independente de W

Passo 3 Aplique essa observação na função F_W^{-1}

Para gerar uma amostra de tamanho n repita esse procedimento n vezes.

Esse método funciona bem, porém nem sempre é possível obter uma expressão analítica para F_W^{-1}, nestes casos, é necessário a utilização de métodos de geração alternativos. Um dos mais utilizados nessa situação é o método de aceitação-rejeição.

B.2 Método de aceitação-rejeição caso geral

O método de aceitação-rejeição gera valores de uma variável aleatória de interesse a partir de candidatos de uma outra variável aleatória conhecida e então, rejeita um subconjunto dos candidatos gerados. O mecanismo de rejeição é construído de forma que os valores que serão aceitos sigam, de fato, a distribuição de interesse.

A ideia básica desse método é gerar uma observação de uma variável aleatória V, digamos v (que sabemos gerar) com função distribuição acumulada F_V e com função densidade de probabilidade f_V. Depois disso, vamos assumir que a razão $f_W(v)/f_V(v)$ seja delimitada por uma constante $c > 0$ tal que $\sup_w\{f_W(v)/f_V(v)\} \leq c$.

Após obter essa constante c, podemos gerar uma observação de uma variável aleatória W com base em uma variável aleatória V através do seguinte algoritmo.

B.2.1 Algoritmo do método de aceitação-rejeição caso geral

O algoritmo do método de aceitação-rejeição é apresentado nos passos abaixo.

P1 Gere uma observação de V;

P2 Gere uma observação U de uma Uniforme(0,1) independente de V;

P3 Se

$$U \leq \frac{f_W(V)}{cf_V(V)};$$

faça W = V ("Aceitar"), caso contrario, retorne ao passo P1 ("Rejeitar").

Esse procedimento é repetido até que se aceite a observação W gerada.

Para gerar uma amostra de tamanho n repita esse procedimento n vezes.

B.2.2 Prova algoritmo de aceitação-rejeição caso geral

Para provar que o algoritmo funciona, devemos mostrar que a distribuição condicional de V em $U \leq f_W(V)/cf_V(V)$ é a mesma distribuição de W. Para isso é necessário demonstrar que se A e B são variáveis aleatórias e h é uma função contínua, então vale os seguintes resultados

I) $f_{A|B}(a|b) = \frac{f_{A,B}(a,b)}{f_B(b)}$;

II) $P(A \leq h(B)) = \int_{-\infty}^{\infty} P\left(A \leq h(b)\right) f_B(b)db$;

Prova resultado I

$$\begin{aligned}
f_{A|B}(a|b) &= \lim_{\Delta_a,\Delta_b\to 0} \frac{P(a \leq A \leq a+\Delta_a | b \leq B \leq b+\Delta_b)}{\Delta_b} \\
&= \lim_{\Delta_a,\Delta_b\to 0} \frac{P(a \leq A \leq a+\Delta_a, b \leq B \leq b+\Delta_b)}{\Delta_a P(b \leq B \leq b+\Delta_b)} \\
&= \lim_{\Delta_a,\Delta_b\to 0} \frac{\frac{P(a\leq A\leq a+\Delta_a, b\leq B\leq b+\Delta_b)}{\Delta_a\Delta_b}}{\frac{P(b\leq B\leq b+\Delta_b)}{\Delta_b}} \\
&= \frac{f_{A,B}(a,b)}{f_B(b)}.
\end{aligned}$$

Prova resultado II

$$P(A \leq h(B)) = \int_{-\infty}^{\infty} \int_{-\infty}^{h(B)} f_{A,B}(a,b) da db.$$

Pelo resultado I temos que $f_{A,B}(a,b) = f_{A|B}(a|b) f_B(b)$. Com isso, temos

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{h(B)} f_{A,B}(a,b) da db &= \int_{-\infty}^{\infty} \int_{-\infty}^{h(B)} f_{A|B}(a|b) f_B(b) da db \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{h(B)} f_{A|B}(a|b) da \right] f_B(b) db \\
&= \int_{-\infty}^{\infty} P\left(A \leq h(B) | B = b\right) f_B(b) db. \\
&= \int_{-\infty}^{\infty} P\left(A \leq h(b)\right) f_B(b) db.
\end{aligned}$$

Em posse desses resultados é possível provar que a distribuição condicional de V dado $U \leq f_W(V)/c f_V(V)$ é de fato F_W. Pela fórmula de Bayes temos que essa probabilidade pode ser escrita como

$$P\left(V \leq w \mid U \leq \frac{f_W(V)}{c f_V(V)}\right) = \frac{P\left(V \leq w, U \leq \frac{f_W(V)}{c f_V(V)}\right)}{P\left(U \leq \frac{f_W(V)}{c f_V(V)}\right)}. \tag{B.1}$$

Primeiramente, vamos calcular a probabilidade de aceitação em cada tentativa, $P\left(U \leq \frac{f_W(V)}{c f_V(V)}\right)$. Pelo resultado II, temos que essa probabilidade pode ser escrita como

$$\begin{aligned}
P\left(U \leq \frac{f_W(V)}{c f_V(V)}\right) &= \int_{-\infty}^{\infty} P\left(U \leq \frac{f_W(t)}{c f_V(t)}\right) f_V(t) dt \\
&= \int_{-\infty}^{\infty} F_U\left(\frac{f_W(t)}{c f_V(t)}\right) f_V(t) dt \\
&= \int_{-\infty}^{\infty} \frac{f_W(t)}{c f_V(t)} f_V(t) dt \\
&= \int_{-\infty}^{\infty} \frac{f_W(t)}{c} dt \\
&= \frac{1}{c} \int_{-\infty}^{\infty} f_W(t) dt \\
&= \frac{1}{c}.
\end{aligned} \tag{B.2}$$

Como a probabilidade de aceitação em cada tentativa é igual a $1/c$, é interessante escolhermos uma variável aleatória V que maximize essa probabilidade, ou seja, devemos escolher uma variável aleatória V que faça com que a constante c se aproxime de um. Isso fará que o algoritmo seja mais eficiente para gerar a variável de interesse.

Voltando a equação (B.1) dada por

$$P\left(V \leq w \mid U \leq \frac{f_W(V)}{c f_V(V)}\right) = \frac{P\left(U \leq \frac{f_W(V)}{c f_V(V)}, V \leq w\right)}{P\left(U \leq \frac{f_W(V)}{c f_V(V)}\right)}.$$

Conjuntamente com a equação (B.2), temos que

$$\begin{aligned}
\frac{P\left(U \leq \frac{f_W(V)}{cf_V(V)}, V \leq w\right)}{P\left(U \leq \frac{f_W(V)}{cf_V(V)}\right)} &= \frac{\int_{-\infty}^{w} P\left(U \leq \frac{f_W(t)}{cf_V(t)}\right) f_V(t)dt}{\frac{1}{c}} \\
&= c\int_{-\infty}^{w} P\left(U \leq \frac{f_W(t)}{cf_V(t)}\right) f_V(t)dt \\
&= c\int_{-\infty}^{\infty} F_U\left(\frac{f_W(t)}{cf_V(t)}\right) f_V(t)dt \\
&= c\int_{-\infty}^{w} \frac{f_W(t)}{cf_V(t)} f_V(t)dt \\
&= \int_{-\infty}^{w} f_W(t)dt \\
&= F_W(w).
\end{aligned}$$

Ou seja, a observação gerada pelo método de aceitação-rejeição tem, de fato, a distribuição da variável de interesse W.

B.3 Método de aceitação-rejeição para dados viesados

Essa ideia do método de aceitação-rejeição pode ser adaptada para o caso de dados viesados. Nesse caso o objetivo é gerar uma observação de uma variável viesada Y com função densidade de probabilidade denotada por f_Y a partir de uma variável aleatória X com função densidade de probabilidade f_X. A densidade viesada pode ser escrita como

$$f_Y(y) := \frac{g(y)f_X(y)}{\mu}. \tag{B.3}$$

Com isso, temos que

$$c = \sup_y \left\{ \frac{f_Y(y)}{f_X(y)} \right\} = \sup_y \left\{ \frac{\frac{g(y)f_X(y)}{\mu}}{f_X(y)} \right\} = \sup_y \left\{ \frac{g(y)}{\mu} \right\}$$

Após obter essa constante c, podemos gerar uma observação de uma variável aleatória Y com base em uma variável aleatória X através do seguinte algoritmo

B.3.1 Algoritmo do método de aceitação-rejeição para dados viesados

O algoritmo do método de aceitação-rejeição para dados viesados é apresentado nos passos abaixo.

P1 Gere uma observação de X;

P2 Gere uma observação U de uma Uniforme(0,1) independente de X;

P3 Se

$$U \leq \frac{g(X)}{c\mu};$$

faça Y = X ("Aceitar"), caso contrario, retorne ao passo P1 ("Rejeitar").

Esse procedimento é repetido até que se aceite a observação X gerada.

Para gerar uma amostra de tamanho n repita esse procedimento n vezes.

B.3.2 Prova do algoritmo de aceitação-rejeição para dados viesados

Para provar que o algoritmo funciona de fato, temos que provar que, a distribuição condicional de X dado $U \leq g(X)/c\mu$ é de fato F_Y. Pela fórmula de Bayes temos que essa probabilidade pode ser escrita como

$$P\left(X \leq y \mid U \leq \frac{g(X)}{c\mu}\right) = \frac{P\left(X \leq y, U \leq \frac{g(X)}{c\mu}\right)}{P\left(U \leq \frac{g(X)}{c\mu}\right)}. \tag{B.4}$$

Primeiramente, vamos calcular a probabilidade de aceitação em cada tentativa, $P\left(U \leq \frac{g(X)}{c\mu}\right)$. Pelo resultado II, temos que essa probabilidade pode ser escrita como

$$\begin{aligned} P\left(U \leq \frac{g(X)}{c\mu}\right) &= \int_{-\infty}^{\infty} P\left(U \leq \frac{g(t)}{c\mu}\right) f_X(t)dt \\ &= \int_{-\infty}^{\infty} F_U\left(\frac{g(t)}{c\mu}\right) f_X(t)dt \\ &= \int_{-\infty}^{\infty} \frac{g(t)}{c\mu} f_X(t)dt. \end{aligned}$$

Pela equação (B.3), temos que

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{g(t)}{c\mu} f_X(t)dt &= \int_{-\infty}^{\infty} \frac{f_Y(t)}{c}dt \\ &= \frac{1}{c}\int_{-\infty}^{\infty} f_Y(t)dt \\ &= \frac{1}{c}. \end{aligned} \tag{B.5}$$

De forma análoga ao caso geral, para deixar o algoritmo mais eficiente é interessante escolhermos uma variável aleatória X que faça com que a constante c se aproxime de um.

Voltando a equação (B.4)

$$P\left(X \leq y \mid U \leq \frac{g(X)}{c\mu}\right) = \frac{P\left(U \leq \frac{g(X)}{c\mu}, X \leq y\right)}{P\left(U \leq \frac{g(X)}{c\mu}\right)}.$$

Utilizando o resultado II e a equação (B.5), temos que

$$\begin{aligned}
\frac{P\left(U \leq \frac{g(X)}{c\mu}, X \leq y\right)}{P\left(U \leq \frac{g(X)}{c\mu}\right)} &= \frac{\int_{-\infty}^{y} P\left(U \leq \frac{g(t)}{c\mu}\right) f_X(t)dt}{\frac{1}{c}} \\
&= c\int_{-\infty}^{y} P\left(U \leq \frac{g(t)}{c\mu}\right) f_X(t)dt \\
&= c\int_{-\infty}^{y} F_U\left(\frac{g(t)}{c\mu}\right) f_X(t)dt \\
&= c\int_{-\infty}^{y} \frac{g(t)}{c\mu} f_X(t)dt \\
&= \int_{-\infty}^{y} f_Y(t)dt \\
&= F_Y(y).
\end{aligned}$$

Ou seja, a observação gerada pelo método de aceitação-rejeição no caso de dados viesados tem, de fato, a distribuição da variável de interesse Y.

APÊNDICE C – Dados velocidade das galáxias

9172	9350	9483	19440	19473	19529	20221	20415	20629	22314	22374
22495	24289	24366	9558	9775	10227	19541	19547	19663	20795	20821
20846	22746	22747	22888	24771	32789	10406	16084	16170	19846	19856
19863	20875	20986	21137	22914	23206	23241	24990	34279	18419	18552
18600	19914	19918	19973	21492	21701	21814	23263	23484	23538	25633
26960	18927	19052	19070	19989	20166	20175	21921	21960	22185	23542
23666	23706	26995	32065	19330	19343	19349	20179	20196	20215	22209
22242	22249	23711	24129	24285						

Tabela 2 – Velocidade medida em km/s para 82 galáxias na região da Coroa Boreal

APÊNDICE D – Dados concentração de álcool no sangue

0.04	0.07	0.07	0.08	0.11	0.12	0.15	0.16	0.17	0.17	0.02	0.02	0.02
0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.09	0.10
0.10	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.14
0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.17	0.17
0.17	0.17	0.17	0.18	0.18	0.18	0.19	0.20	0.21	0.21	0.21	0.22	0.24
0.24	0.24	0.25	0.30	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.05	0.06	0.06	0.07
0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.10
0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.12	0.12	0.12	0.12	0.12
0.12	0.12	0.12	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14	0.14	0.14
0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16
0.16	0.16	0.17	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.20
0.20	0.21	0.23	0.24	0.24	0.26	0.26	0.27	0.60	0.01	0.01	0.01	0.03
0.03	0.03	0.04	0.05	0.06	0.06	0.07	0.08	0.08	0.08	0.09	0.09	0.09
0.09	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.12
0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.13	0.13	0.13	0.14
0.14	0.14	0.14	0.14	0.14	0.15	0.16	0.16	0.16	0.17	0.18	0.18	0.19
0.19	0.19	0.19	0.20	0.21	0.21	0.21	0.21	0.22	0.23	0.23	0.24	0.27
0.30	0.01	0.01	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.06
0.07	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.10	0.10	0.10	0.11	0.11
0.11	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.14	0.15	0.16	0.16	0.16
0.16	0.17	0.17	0.17	0.18	0.18	0.19	0.20	0.20	0.20	0.21	0.22	0.24
0.26	0.01	0.01	0.01	0.01	0.03	0.04	0.05	0.06	0.07	0.09	0.09	0.09
0.09	0.09	0.10	0.11	0.12	0.12	0.12	0.13	0.13	0.14	0.15	0.15	0.15
0.15	0.16	0.16	0.16	0.16	0.16	0.17	0.17	0.19	0.19	0.19	0.19	0.19
0.20	0.20	0.20	0.20	0.21	0.21	0.22	0.25	0.26	0.26	0.27	0.28	0.29
0.01	0.01	0.01	0.02	0.03	0.06	0.07	0.07	0.08	0.08	0.08	0.09	0.09
0.10	0.10	0.10	0.11	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.15	0.15
0.15	0.16	0.16	0.16	0.17	0.17	0.17	0.17	0.18	0.19	0.20	0.20	0.21
0.23	0.25	0.03	0.09	0.09	0.11	0.13	0.16	0.17	0.18	0.20	0.20	0.21
0.22	0.23	0.23	0.24	0.24	0.24	0.27	0.41	0.01	0.01	0.10	0.14	0.15
0.16	0.16	0.18	0.20	0.22	0.24	0.01	0.01	0.02	0.05	0.05	0.09	0.09
0.09	0.11	0.11	0.15	0.16	0.16	0.17	0.17	0.18	0.20	0.22	0.23	0.25
0.30	0.33	0.38	0.03	0.09	0.12	0.13	0.14	0.17	0.19	0.20	0.20	0.20
0.21	0.22	0.24	0.26	0.27	0.28	0.29	0.36	0.01	0.01	0.09	0.11	0.16
0.16	0.18	0.19	0.21	0.22	0.26	0.03	0.07	0.09	0.10	0.10	0.11	0.14
0.15	0.19	0.20	0.21	0.21	0.22	0.24	0.24	0.27	0.27	0.34	0.37	0.01

0.07	0.09	0.10	0.12	0.12	0.14	0.19	0.20	0.21	0.24	0.07	0.08	0.10
0.12	0.16	0.17	0.17	0.19	0.21	0.22	0.24	0.25	0.36	0.36	0.38	0.01
0.04	0.05	0.06	0.06	0.12	0.15	0.15	0.20	0.21	0.28	0.30	0.01	0.01
0.01	0.06	0.07	0.09	0.14	0.20	0.21	0.23	0.27	0.30	0.32	0.10	0.12
0.12	0.14	0.17	0.19	0.19	0.21	0.21	0.23	0.23	0.30	0.04	0.21	0.22
0.22	0.32	0.01	0.01	0.09	0.10	0.15	0.19	0.23	0.24	0.24	0.04	0.13
0.07	0.09	0.10	0.12	0.12	0.14	0.19	0.20	0.21	0.24	0.07	0.08	0.10
0.12	0.16	0.17	0.17	0.19	0.21	0.22	0.24	0.25	0.36	0.36	0.38	0.01
0.04	0.05	0.06	0.06	0.12	0.15	0.15	0.20	0.21	0.28	0.30	0.01	0.01
0.01	0.06	0.07	0.09	0.14	0.20	0.21	0.23	0.27	0.30	0.32	0.10	0.12
0.12	0.14	0.17	0.19	0.19	0.21	0.21	0.23	0.23	0.30	0.04	0.21	0.22
0.22	0.32	0.01	0.01	0.09	0.10	0.15	0.19	0.23	0.24	0.24	0.04	0.13
0.14	0.17	0.18	0.19	0.20	0.20	0.21	0.21	0.22	0.27	0.27	0.27	0.46
0.14	0.17	0.32	0.32	0.01	0.02	0.04	0.09	0.10	0.10	0.11	0.12	0.16
0.17	0.07	0.12	0.13	0.18	0.19	0.23	0.23	0.24	0.25	0.30	0.12	0.13
0.17	0.21	0.22	0.27	0.01	0.05	0.05	0.11	0.15	0.24	0.38	0.01	0.01
0.05	0.08	0.14	0.17	0.17	0.18	0.19	0.19	0.19	0.20	0.22	0.31	0.07
0.07	0.11	0.18	0.18	0.18	0.21	0.02	0.03	0.10	0.14	0.17	0.17	0.20
0.21	0.22	0.22	0.27	0.27	0.28	0.30	0.03	0.06	0.06	0.11	0.14	0.15
0.18	0.22	0.26	0.27	0.34	0.34	0.06	0.09	0.09	0.13	0.15	0.17	0.26
0.28	0.04	0.13	0.16	0.19	0.21	0.22	0.28	0.05	0.15	0.20	0.20	0.22
0.03	0.12	0.21	0.26	0.01	0.01	0.10	0.18	0.19	0.26	0.26	0.28	0.33
0.07	0.15	0.17	0.18	0.23	0.26	0.14	0.15	0.26	0.31	0.35	0.03	0.10
0.21	0.25	0.01	0.10	0.17	0.01	0.01	0.10	0.01	0.13	0.20	0.23	0.24
0.01	0.08	0.14	0.14	0.17	0.20	0.01	0.01	0.03	0.12	0.25	0.01	0.09
0.10	0.18	0.01	0.11	0.24	0.02	0.21	0.18	0.22	0.05	0.09	0.11	0.19
0.01	0.05	0.17	0.01	0.07	0.06	0.12	0.01	0.01	0.04	0.06	0.14	0.15
0.15	0.08	0.01	0.07									

Tabela 3 – Concentração de álcool no sangue dos motoristas em grama/100ml

APÊNDICE E – Rotinas no R

Rotina para o exemplo das galaxias

```
# Carregar os pacotes do R necessários para o script

library(MESS)

library(graphics)

# Ler a base de dados da Tabela 2

x <- data



# Funçoes usadas no calculo dos pesos e na obtençao da matriz dos valores de phi \\
# e do ponto de corte J


########################

w.opt2 <- function(k,x,d){

phimat <- BC(x,k)

theta.est <- colMeans(phimat)

w <- 1-d/(length(x)*theta.est^2)

w <- w*(w>0)

w[1] <- 1

w
}


##########################################
```

```
bc <- function(x,j) {

1*(j==0)+sqrt(2)*cos(pi*j*x)*(j!=0)

}

BC <- function(x,J){

sapply(0:J,bc,x=x)

}


################################################

JN <- function(cj0,cj1,x) {

Jn <- floor(cj0 + cj1*log(length(x)))

}

J.opt <- function(J,x,d) {

phimat=BC(x,max(J))

theta.est=colMeans(phimat)

which.min(cumsum(2*d/(length(x))-theta.est^2))-1

}

# Funçao principal usada para estimação

f2 <- function(x,d,cjm,c0,c1,ct){

# Reescalando os dados para o intervalo [0,1]

y <- (x-min(x))/(max(x)-min(x))
```

```
# Escolhendo o J ótimo

Jn <- JN(c0,c1,y)

Jgrid <- 0:Jn*cjm

J <- J.opt(Jgrid,y,d)

wts <- w.opt2(J,y,d)

# Gerando uma sequência do maximo ao minimo de y

y.obs <- seq(min(y),max(y),length.out = 100)

# Obtendo a matriz phi

phimat <- BC(y,Jn*cjm)

# Estimando theta

theta.est <- colMeans(phimat)

phi.obs <- BC(y.obs,Jn*cjm)


theta.reg <- c(theta.est[1:(J+1)]*wts,(theta.est[(J+2):(Jn*cjm+1)]^2>
(ct*d*log(length(x)))/length(x))*theta.est[(J+2):(Jn*cjm+1)])

# Estimando a densidade

f.est.y   <- drop(phi.obs%*%theta.reg)

f.est.y[f.est.y < 0] <- 0

f.est.x <- f.est.y/(max(x)- min(x))

x.obs <- seq(min(x),max(x),length.out = 100)

# Calculando a área para dividir e  garantir que a função integre 1
```

```
area = auc(x.obs,f.est.x,from = min(x.obs),to=max(x.obs))

# Plotando os gráficos
plot(x.obs, f.est.x/area, type = "l", lty = 2, col="red",xlab = "Velocidade",ylab =


return(list(J,wts,theta.est[1:(J+1)],theta.reg))

}

# Argumentos da função

c0 <- 4
c1 <- 5
cjm <- 6
ct <- 4
d <- 1

est <- f2(x, d, cjm, c0, c1, ct)


hist(x,probability = T,breaks = 10,add=T)
```

Rotina da simulação aceitação-rejeição

```
################################################################

# Funcão de vies

g <- function(y) 0.1 +0.9*y

# Funcao para gerar a amostra viesada (beta)

gbeta <- function(y, n, g, alfa, beta){
```

```
i <- function(y) dbeta(y, alfa, beta)

Multiply <- function(a, b){

force(a)

force(b)

function(x){a(x) * b(x)}

}

h <- Multiply(g, i)

mu <- integrate(h, 0, 1)$val

k <- g(optimise(g, c(0,1), maximum = T, tol = .Machine$double.eps)$maximum)/mu

idx <- 0

while(idx < n){

x <- rbeta(1, alfa, beta)

u <- runif(1, 0, 1)

r <- u*k

if(r <= g(x)/mu){

idx <- idx + 1

y[idx] <- x

}

}

y
```

```
}

# Funcao para gerar a amostra viesada da funcão estratificada

gest <- function(y, n, g, m){

a <- function(y) 8*y

b <- function(y) 4*(1-2*y)

c <- function(y) 8*y - 4

d <- function(y) 8 - 8*y

Multiply <- function(a, b)

{

force(a)

force(b)

function(x) {a(x) * b(x)}

}

h <- Multiply(g, a)

h1 <- Multiply(g, b)

h2 <- Multiply(g, c)

h3 <- Multiply(g, d)

mu <- integrate(h, 0, 1/4)$val + integrate(h1, 1/4, 1/2)$val + integrate(h2, 1/2, 3

k <- m/mu

f.ex3 <- function(x)
```

```
{
y <- rep(NA, length(x))
for(i in 1 : length(x))
{
if(x[i] <= 0.25)
{
y[i] <- sqrt(x[i]/4)
}
if(x[i] > 0.25 && x[i] <= 0.5)
{
y[i] <- (4 - sqrt(8 - 16*x[i]))/8
}

if(x[i] > 0.5 && x[i] <= 0.75)
{
y[i] <-(4 + sqrt(16*x[i] - 8))/8
}
if(x[i] > 0.75)
{
y[i] <- (8 - sqrt(16 - 16*x[i]))/8
```

```
}

}

y

}

idx <- 0
while(idx < n)

{

x <- f.ex3(runif(1, 0, 1))

u <- runif(1, 0, 1)

r <- u*k

if(r <= g(x)/mu)

{

idx <- idx + 1

y[idx] <- x

}

}

y

}


# Gerando a amostra
```

```
n <- 500

z <- rep(NA, n)

g <- function(y)0.1 + 0.9*y

alfa <- 2

beta <- 2

x <- gbeta(z, n, g, alfa, beta)

# Funções usadas na estimação

g1 <- function(x) 1/g(x)

########################

w.opt2<-function(k,y,d,mu.h){



phimat=BC(y,k)

theta.est=mu.h*colMeans(phimat/g(y))




w<-1-d/(length(x)*theta.est^2)
w<-w*(w>0)
w[1]<-1
w
}


##########################################
bc<-function(x,j){
1*(j==0)+sqrt(2)*cos(pi*j*x)*(j!=0)
```

```
}

BC<-function(x,J){
sapply(0:J,bc,x=x)
}


################################################

JN<-function(cj0,cj1,x){
Jn<-floor(cj0 + cj1*log(length(x)))

}

# Argumentos da função

c0=4

c1=5

cjm=6

ct=4

x.obs<-seq(min(x),max(x),length.out = 50)

y<-(x-min(x))/(max(x)-min(x))

y<-sort(y)

y<-y[-1]

y.obs<-(x.obs-min(x.obs))/(max(x.obs)-min(x.obs))

mu.h <- length(y)/sum(g1(y))

d <- mu.h^2* sum(g1(y)^2) /length(y)

Jn<-JN(c0,c1,x)
```

```
#

J <- 7

wts<-w.opt2(J,y,d,mu.h)

fx <- function(y) dbeta(y, 2, 2)

#Funcao

f2<-function(x,x.obs,y,y.obs,mu.h,d,cjm,Jn,J,wts,fx){


phimat<-BC(y,Jn*cjm)

theta.est<-mu.h*colMeans(phimat/g(y))



phi.obs<-BC(y.obs,Jn*cjm)

theta.reg<-c(theta.est[1:(J+1)]*wts,(theta.est[(J+2):(Jn*cjm+1)]^2>(ct*d*log(length

f.est<-drop(phi.obs%*%theta.reg/(max(x)-min(x)))



plot(y.obs, f.est, type = "l", ylim = c(0, 3), lty = 3,xlab="y",ylab=expression(hat

points(y.obs, fx(y.obs), type = "l", lty = 1, col = 2)


}

f2(x,x.obs,y,y.obs,mu.h,d,cjm,Jn,J,wts,fx)
```

```
mu <- integrate(g, 0, 1)$val

fy <- function(y) (dbeta(y,2,2)*g(y))/mu

points(y.obs,fy(y.obs),type = "l", col="green",lty=2)

legend(x="topright",c("Densidade viesada","Verdadeira densidade", "Estimativa"),
col=c("Green","red", "black"),lty=c(2,1,3),lwd=1,cex=.7)

#########################################################
############################################################
#######################################################

par(mfrow=c(2,2))

set.seed(1100)

# Gerar uma amostra viesada de tamanho 1000

n <- 10000

z <- rep(NA, n)

g <- function(y) 0.9*y + 0.1

m <- g(optimise(g, c(0,1), maximum = T, tol = .Machine$double.eps)$maximum)

x <- gest(z, n, g, m)

#funções usadas

g1 <- function(x) 1/g(x)

#######################
#######################

w.opt2<-function(k,y,d,mu.h){

phimat=BC(y,k)
```

```
theta.est=mu.h*colMeans(phimat/g(y))

w<-1-d/(length(x)*theta.est^2)

w<-w*(w>0)

w[1]<-1

w

}


##############################################

bc<-function(x,j){

1*(j==0)+sqrt(2)*cos(pi*j*x)*(j!=0)

}

BC<-function(x,J){

sapply(0:J,bc,x=x)

}


################################################

JN<-function(cj0,cj1,x){

Jn<-floor(cj0 + cj1*log(length(x)))

}

#ARGUMENTOS
```

```
c0=4
c1=5
cjm=6
ct=4

x.obs<-seq(min(x),max(x),length.out = 50)

y<-(x-min(x))/(max(x)-min(x))

y<-sort(y)

y<-y[-1]

y.obs<-(x.obs-min(x.obs))/(max(x.obs)-min(x.obs))

mu.h <- length(y)/sum(g1(y))

d <- mu.h^2* sum(g1(y)^2) /length(y)

Jn<-JN(c0,c1,x)

J<-7

wts<-w.opt2(J,y,d,mu.h)

fx <- function(x){

y <- rep(NA, length(x))

for(i in 1: length(x))

{

if(x[i] <= 0.25)

{

y[i] <- 8*x[i]
```

```
}

if(x[i] > 0.25 && x[i] <= 0.5)

{

y[i] <- 4*(1 - 2*x[i])

}

if(x[i] > 0.5 && x[i] <= 0.75)

{

y[i] <- 8*x[i] - 4

}

if(x[i] > 0.75)

{

y[i] <- 8 - 8*x[i]

}

}

y

}


#Funcao

f2<-function(x,x.obs,y,y.obs,mu.h,d,cjm,Jn,J,wts,fx){


phimat<-BC(y,Jn*cjm)
```

```
theta.est<-mu.h*colMeans(phimat/g(y))



phi.obs<-BC(y.obs,Jn*cjm)

theta.reg<-c(theta.est[1:(J+1)]*wts,(theta.est[(J+2):(Jn*cjm+1)]^2>(ct*d*log(length

f.est<-drop(phi.obs%*%theta.reg/(max(x)-min(x)))



plot(y.obs, f.est, type = "l", ylim = c(0, 3), lty = 3)

points(y.obs, fx(y.obs), type = "l", lty = 1, col = 2)


}

f2(x,x.obs,y,y.obs,mu.h,d,cjm,Jn,J,wts,fx)

mu <- integrate(g, 0, 1)$val
fy <- function(x){

y <- rep(NA, length(x))

for(i in 1: length(x))
{

if(x[i] <= 0.25)
{

y[i] <- sqrt((8*x[i]*g(x[i]))/mu)

}
if(x[i] > 0.25 && x[i] <= 0.5)
{
```

```
y[i] <- sqrt(4*g(x[i])*(1 - 2*x[i])/mu)

}

if(x[i] > 0.5 && x[i] <= 0.75)

{

y[i] <- sqrt(g(x[i])*(8*x[i] - 4)/mu)

}

if(x[i] > 0.75)

{

y[i] <- sqrt(g(x[i])*(8 - 8*x[i])/mu)

}

}

y

}
points(y.obs,fy(y.obs),lty=2,type="l",col="green")

legend("topleft",c("Densidade viesada","Verdadeira densidade", "Estimativa"),
col=c("Green","red", "black"),lty=c(2,1,3),lwd=1,cex=.7)
```

Rotina aplicação dados concentração de álcool no sangue

```
# Carregando pacote e ler os dados

library(MESS)

y <- data2

# Função de viés
```

```
g <- function(y) 0.1 + 0.9*y

# Função inversa

g1 <- function(x) 1/g(x)

# Estimação do mu

mu.h <- length(y)/sum(g1(y))

# Funções usadas na estimação

J.opt2 <- function(J,x,d,mu.h){


phimat=BC(x,max(J))

theta.est=mu.h*colMeans(phimat/g(x))



which.min(cumsum(2*d/(length(x))-theta.est^2))-1

}
w.opt2<-function(k,y,d,mu.h){



phimat=BC(y,k)

theta.est=mu.h*colMeans(phimat/g(y))




w<-1-d/(length(y)*theta.est^2)
w<-w*(w>0)
w[1]<-1
```

```
w
}


##############################################
bc<-function(x,j){
1*(j==0)+sqrt(2)*cos(pi*j*x)*(j!=0)
}

BC<-function(x,J){
sapply(0:J,bc,x=x)
}


################################################

JN<-function(cj0,cj1,x){
Jn<-floor(cj0 + cj1*log(length(x)))

}

#Argumentos da função

c0=4
c1=5
cjm=6
ct=4

# Gerando uma grade de pontos na amplitude dos dados

y.obs<-seq(min(y),max(y),length.out = 60)

# Estimando os parâmetros

mu.h <- length(y)/sum(g1(y))
d <- mu.h^2* sum(g1(y)^2) /length(y)
Jn<-JN(c0,c1,y)
Jgrid <- seq(0:Jn*cjm)
J <- J.opt2(Jgrid,y,d,mu.h)
```

```
wts<-w.opt2(J,y,d,mu.h)

#Função principal para estimar a densidade

f2<-function(y,y.obs,mu.h,d,cjm,Jn,J,wts){


phimat<-BC(y,Jn*cjm)

theta.est<-mu.h*colMeans(phimat/g(y))



phi.obs<-BC(y.obs,Jn*cjm)

theta.reg<-c(theta.est[1:(J+1)]*wts,(theta.est[(J+2):(Jn*cjm+1)]^2>(ct*d*log(length

f.est<-drop(phi.obs%*%theta.reg)
f.est[f.est < 0] <- min(f.est[f.est>0])


c = auc(y.obs,f.est,from = min(y.obs),to=max(y.obs))


plot(y.obs, f.est/c, type = "l", lty = 2, col = 2,ylab = "y",xlab = "Concentração d



}

#Chamando a função principal
f.est <- f2(y,y.obs,mu.h,d,cjm,Jn,J,wts)
```