Estimação de funções do redshift de galáxias com base em dados fotométricos

Gretta Rossi Ferreira

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

GRETTA ROSSI FERREIRA

ESTIMAÇÃO DE FUNÇÕES DO REDSHIFT DE GALÁXIAS COM BASE EM DADOS FOTOMÉTROCOS

Dissertação apresentada ao Departamento de Estatística – Des/UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre ou Doutor em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP.

Orientador: Prof. Dr. Rafael Izbicki

São Carlos

Novembro de 2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

GRETTA ROSSI FERREIRA

GALAXIES REDSHIFT FUNCTION ESTIMATION USING PHOTOMETRIC DATA

Dissertation submitted to the Departamento de Estatística – Des/UFSCar and to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in the partial fulfillment for the Master degree in Statistics – Interinstitucional Program of Pos-Graduation in Satatistics UFSCar-USP.

Advisor: Prof. Dr. Rafael Izbicki

São Carlos

November 2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de dissertação de mestrado da candidata Gretta Rossi Ferreira realizada em 18/09/2017:

Prof. Dr. Rafael Izbicki
UFSCar

Prof. Dr. Paulo Henrique Ferreira da Silva
UFBA

Profa. Dra. Teresa Cristina Martins Dias
UFSCar

Certifico que a sessão de defesa foi realizada com a participação à distância do membro Prof. Dr. Paulo Henrique Ferreira da Silva e, depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da comissão examinadora redigido no relatório de defesa do(a) aluno(a) Gretta Rossi Ferreira.

Prof. Dr. Rafael Izbicki
Presidente da Comissão Examinadora
UFSCar

RESUMO

FERREIRA, G. R. **Estimação de funções do redshift de galáxias com base em dados fotométricos**. 2017. 52 p. Dissertação (Mestrado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Em uma quantidade substancial de problemas de astronomia, tem-se interesse na estimação do valor assumido, para diversas funções g, de alguma quantidade desconhecida $z \in \Re$ com base em covariáveis $\mathbf{x} \in \Re^d$. Isto é feito utilizando-se uma amostra $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$. As duas abordagens usualmente utilizadas para resolver este problema consistem em (1) estimar a regressão de Z em $\mathbf{x}$, e plugar esta na função g ou (2) estimar a densidade condicional $f(z|\mathbf{x})$ e plugá-la em $\int g(z)f(z|\mathbf{x})dz$. Infelizmente, poucos estudos apresentam comparações quantitativas destas duas abordagens. Além disso, poucos métodos de estimação de densidade condicional tiveram seus desempenhos comparados nestes problemas. Em vista disso, o objetivo deste trabalho é apresentar diversas comparações de técnicas de estimação de funções de uma quantidade desconhecida. Em particular, damos destaque para métodos não paramétricos. Além dos estimadores (1) e (2), propomos também uma nova abordagem que consistem em estimar diretamente a função de regressão de $g(Z)$ em $\mathbf{x}$. Essas abordagens foram testadas em diferentes funções nos conjuntos de dados DEEP2 e Sheldon 2012. Para quase todas as funções testadas, o estimador (1) obteve os piores resultados, exceto quando utilizamos florestas aleatórias. Em diversos casos, a nova abordagem proposta apresentou melhores resultados, assim como o estimador (2). Em particular, verificamos que métodos via florestas aleatórias, em geral, levaram a bons resultados.

Palavras-chave: Astroestatística, Densidade Condicional, Inferência não Paramétrica, Regressão.

ABSTRACT

FERREIRA, G. R. **Galaxies redshift function estimation using photometric data**. 2017. 52 p. Dissertação (Mestrado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

In a substantial amount of astronomy problems, we are interested in estimating values assumed of some unknown quantity $z \in \Re$, for many function g, based on covariates $\mathbf{x} \in \Re^d$. This is made using a sample $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$. Two approaches that are usually used to solve this problem consist in (1) estimating a regression function of Z in $\mathbf{x}$ and plugging it into the g or (2) estimating a conditional density $f(z|\mathbf{x})$ and plugging it into $\int g(z) f(z|\mathbf{x}) dz$. Unfortunately, few studies exhibit quantitative comparisons between these two approaches. Besides that, few conditional density estimation methods had their performance compared in these problems. In view of this, the objective of this work is to show several comparisons of techniques used to estimate functions of unknown quantity. In particular we highlight nonparametric methods. In addition to estimators (1) and (2), we also propose a new approach that consists in directly estimating the regression function from $g(Z)$ on x. These approaches were tested in different functions in the DEEP2 and Sheldon 2012 datasets. For almost all the functions tested, the estimator (1) obtained the worst results, except when we use the random forests methods. In several cases, the proposed new approach presented better results, as well as the estimator (2). In particular, we verified that random forests methods generally present to good results.

Keywords: Astrostatistics, Conditional Density, Nonparametric Inference, Regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Cada vez mais, diversas áreas da ciência necessitam de estimadores não-paramétricos eficientes para uma densidade condicional $f(z|\mathbf{x})$, em que $\mathbf{x}$ é um vetor com dimensionalidade alta e z é uma variável real. Em particular, astronomia é um campo em que a estimação desta quantidade desempenha um papel cada vez mais importante.

Um exemplo de extrema importância é a estimação do redshift de galáxias com base em dados fotométricos. Redshift nada mais é do que a distância da galáxia8 até a terra. Esse problema utiliza $f(z|\mathbf{x})$ e possibilita a estimação de parâmetros cosmológicos com grande precisão, em particular uma precisão maior que aquela dada por métodos de regressão (CUNHA *et al.*, 2009; SHELDON *et al.*, 2012; KIND; BRUNNER, 2013; RAU *et al.*, 2015; IZBICKI; LEE; FREEMAN, 2017; IZBICKI; LEE, 2016). O ganho de precisão é tão grande que Ball e Brunner (2010) mencionam o uso de densidades condicionais como uma das futuras tendências deste campo.

Em uma quantidade substancial de problemas de astronomia, estimadores de $f(z|\mathbf{x})$ podem ser usados em situações em que se deseja estimar uma função g de z em novas observações. Mais especificamente, utiliza-se uma amostra $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$ para se descrever a relação entre $\mathbf{x}$ e z. Com base na relação estimada, pode-se então estimar $g(z_{n+1}), \ldots, g(z_{n+m})$ em m novas observações com covariáveis conhecidas $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}$, veja a Tabela 1.

Neste trabalho, focamos nesta categoria de problemas, que descrevem a relação entre $\mathbf{x}$ e z em novas observações. Um exemplo no qual tal abordagem é utilizada é no problema da estimação do efeito de lentes gravitacionais, que é apresentado mais detalhadamente na Seção 5.1 do Capítulo 5.

Tabela 1 – Covariável $\mathbf{x}$, variável resposta z e a função $g(z)$. O objetivo deste trabalho é desenvolver métodos para prever $g(Z_{n+1}),\ldots,g(Z_{n+m})$ com base em $\mathbf{X}_{n+1},\ldots,\mathbf{X}_{n+m}$.

$\mathbf{x}$	z	$g(z)$
$\mathbf{X}_1$	Z_1	$g(Z_1)$
$\mathbf{X}_2$	Z_2	$g(Z_2)$
$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_n$	Z_n	$g(Z_n)$
$\mathbf{X}_{n+1}$	?	?
$\vdots$	$\vdots$	$\vdots$
$\mathbf{X}_{n+m}$	?	?

As seguintes abordagens para estimar o valor de $g(z)$ em novas observações são comparadas:

1. **Estimador *plugin*** $g(\widehat{\mathbb{E}}[Z|\mathbf{x}])$. Inicialmente cria-se $\widehat{\mathbb{E}}[Z|\mathbf{x}]$, um estimador da função de regressão de Z em $\mathbf{x}$, e pluga-se ele na função g. Esse é um estimador tradicional e bastante utilizado na literatura.

2. **Estimador via densidade condicional** $\int_z g(z)\widehat{f}(z|\mathbf{x})dz$. Este estimador é motivado pelo fato de que $\mathbb{E}[g(Z)|\mathbf{x}] = \int_z g(z)f(z|\mathbf{x})dz$. Isto é, a esperança condicional pode ser escrita como uma integral da função $g(z)$, multiplicada pela densidade condicional. Dessa forma, estima-se a função substituindo a mesma na densidade condicional. Tal estimador foi inicialmente proposto por (MANDELBAUM *et al.*, 2008).

3. **Estimador *direto*** $\widehat{\mathbb{E}}[g(Z)|\mathbf{x}]$. Estima-se a função de regressão da variável aleatória $g(Z)$ em $\mathbf{x}$. Esse método é diferente do método de estimação *plugin* $g(\widehat{\mathbb{E}}[Z|\mathbf{x}])$, já que neste é feita a regressão direta da variável transformada $g(Z)$ em $\mathbf{x}$. Destacamos que esta abordagem é inédita na literatura.

Apesar de diversos trabalhos indicarem que a abordagem baseada em densidade condicional leva a resultados melhores que o estimador ingenuo $g(\hat{z})$, em que $\hat{z}$ é um estimador da regressão de z em $\mathbf{x}$, poucos são os estudos que comparam o efeito de diferentes estimadores da densidade condicional nas estimativas obtidas. Dessa forma, o objetivo deste trabalho é apresentar e investigar o uso de diversas técnicas de estimação de uma função de uma quantidade desconhecida. Em particular, daremos destaque para métodos não paramétricos, com ênfase em estimadores de densidades condicionais em problemas ligados à astronomia.

Neste trabalho, foram utilizados tanto exemplos simulados quanto exemplos reais. Em particular, alguns dos exemplos apresentados por Rau *et al.* (2015) foram investigados sob a ótica das novas ferramentas desenvolvidas.

Este trabalho está organizado da seguinte maneira: o Capítulo 2 apresenta uma breve explicação sobre estimação não paramétrica de regressão, mostrando alguns métodos de estimação, que serão utilizadas ao longo deste trabalho. No Capítulo 3 é feita uma apresentação sobre estimação de densidades condicionais e alguns métodos utilizados neste problema. O Capítulo 4 apresenta um exemplo com dados simulados, no qual podemos motivar e observar o funcionamento dos métodos propostos. Em seguida, no Capítulo 5 aplicamos tais técnicas de estimação a dados reais de astronomia. Finalmente, o Capítulo 6 apresenta as principais conclusões e considerações finais a respeito deste trabalho.

CAPÍTULO

2

ESTIMAÇÃO NÃO PARAMÉTRICA DE REGRESSÃO

Neste capítulo, o objetivo principal é expor o que é e como funcionam os métodos de regressão utilizados neste trabalho. Descrevemos também o que é função de risco e como podemos estimá-la.

2.1 Objetivos da Regressão

O principal objetivo da regressão é descrever a relação entre uma variável resposta Z e um vetor de covariáveis $\mathbf{x}$. Uma forma de descrever essa relação é utilizando a função de regressão. Matematicamente, a função de regressão é definida pela seguinte relação:

$$r(\mathbf{x}) := \mathbb{E}[Z|\mathbf{x} \tag{2.1}$$

Assim, quando temos uma amostra i.i.d. $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$, uma forma de estimar a relação entre $\mathbf{X}$ e Z é estimando $r(\mathbf{x})$. A notação utilizada para a variável resposta é dada pelo vetor $Z_1, \cdots, Z_n$. Para as d covariáveis $\mathbf{X}$, pode ser encontrada na Tabela 2.

Tabela 2 – Notação utilizada nas variáveis de um problema de regressão.

X_{11}	$\cdots$	X_{1d}	$(=\mathbf{X}_1)$
$\vdots$	$\ddots$	$\vdots$	
X_{n1}	$\ldots$	X_{nd}	$(=\mathbf{X}_n)$

Assim, $X_{i,j}$ é o valor da i-ésima observação na j-ésima covariável. Além disso, $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})$.

2.2 Função de Risco

Como o foco neste trabalho é predizer novas observações, precisamos medir a performance dessa função de predição. Quando o objetivo é prever Z, esta será medida através do risco quadrático $R(h) = \mathbb{E}[(Z - h(\mathbf{X}))^2]$, em que $(\mathbf{X}, Z)$ é uma observação nova, que não foi utilizada na estimação de h.

Em problemas reais, é comum ajustar várias regressões $h(\mathbf{x})$ e encontrar qual destas possui o maior poder preditivo. Isso pode ser feito através da estimação do risco para cada uma destas regressões. A que obtiver o menor risco é, então, a que possui melhor poder preditivo, dentre estas.

Modelos que possuem muitos parâmetros podem levar ao que chamamos de superajuste (*overfitting*), descrevendo bem os dados observados mas tendo um poder preditivo bem baixo em novas observações. Modelos com um número de parâmetros muito baixo pode ser simplista demais e não descrever bem os dados, o que chamamos de sub-ajuste (*underfitting*). Precisamos, então, encontrar uma função h que não sofra nem superajuste, nem sub-ajuste e que tenha um bom poder preditivo.

2.3 Validação Cruzada

Desejamos encontrar a melhor função de predição h dentro do conjunto de candidatos h's. Para isso, precisamos estimar o risco preditivo. Um possível estimador para o risco é o Erro Quadrático Médio do conjunto de treinamento, dado por:

$$EQM(h) := \frac{1}{n}\sum_{i=1}^{n}(Z_i - h(\mathbf{X}_i))^2 \tag{2.2}$$

Tal estimador é muito otimista para o risco, uma vez que este leva ao superajuste dos dados (isso porque h foi escolhida para justar bem aos dados observados $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$). Uma possível solução é dividir o conjunto de dados em duas partes, treinamento e validação. Podemos usar por exemplo, 70% para o conjunto treinamento e 30% para o conjunto validação:

$$\underbrace{(\mathbf{X}_1, Z_1), (\mathbf{X}_2, Z_2), \ldots, (\mathbf{X}_n, Z_n)}_{\text{Conjunto Treinamento}}, \underbrace{(\mathbf{X}'_1, Z'_1), \ldots, (\mathbf{X}'_{n'}, Z'_{n'})}_{\text{Conjunto Validação}}$$

O conjunto treinamento é utilizado para estimar a função de predição g, enquanto o conjunto validação estima apenas o risco $R(h)$ por:

$$R(g) \approx \frac{1}{n' - n}\sum_{i=n+1}^{n'}(Z'_i - h(\mathbf{X}'_i))^2. \tag{2.3}$$

Através da equação 2.3, podemos perceber que o Erro Quadrático Médio é avaliado no conjunto validação. Tal estimador de $R(h)$ é consistente pela lei dos grandes números (IZBICKI; SANTOS, 2017), pois o Conjunto Validação não foi utilizado para estimar a h.

O Erro Quadrático Médio é muito otimista para o conjunto treinamento, porque escolhemos cada h que minimize tal erro. Da mesma forma, se h é escolhida de forma a minimizar a estimativa do risco na validação, também haverá superajuste no conjunto de validação. Para evitar essa situação, podemos dividir em três parte o conjunto de dados: Treinamento, Validação e Teste. Os Conjuntos Treinamento e Validação são usados para estimar a função de predição h e para estimar o risco $R(h)$, respectivamente. O Conjunto Teste é então usado para estimar o erro do melhor estimador de regressão encontrado a partir do Conjunto de Validação.

2.4 Métodos de Regressão

Os métodos de regressão utilizados neste trabalho são: Método dos k Vizinhos mais Próximos (BENEDETTI, 1977), Florestas Aleatórias (JAMES *et al.*, 2013), SpAM (RAVIKUMAR *et al.*, 2009) e Séries Espectrais (LEE; IZBICKI, 2016). A seguir, temos uma descrição mais detalhada para esses métodos.

2.4.1 Método dos k Vizinhos mais Próximos

O método dos *k-nearest neighbours* (KNN) é denominado em português por k vizinhos mais próximos (BENEDETTI, 1977). Esse nome se dá pois tal método tem como ideia principal estimar $r(\mathbf{x})$ utilizando os Z's dos k vizinhos mais próximos de $\mathbf{x}$, representado por $\mathscr{N}_x$. Como k é o número de vizinhos mais próximos a $\mathbf{x}$, ele precisa ser inteiro e positivo. Formalmente, definimos a estimativa de $r(\mathbf{x})$ por:

$$\hat{r}(\mathbf{x}) = \frac{1}{K} \sum_{i \in \mathscr{N}_\mathbf{x}} z_i \tag{2.4}$$

Para medir a proximidade dos vizinhos mais próximos a $\mathbf{x}$, pode ser utilizada a distância Euclidiana, ou a distância de Mahalanobis, entre outros métodos existentes na literatura. O número de vizinhos k pode ser escolhido via validação cruzada.

2.4.2 Árvores e Florestas

2.4.2.1 Árvore de Regressão

Árvores são um método de predição útil e simples para interpretação de resultados (JAMES *et al.*, 2013). A ideia desse método é dividir o espaço das covariáveis em J regiões geradas a partir de divisões binárias das covariáveis, como é mostrado na Figura 1, de modo que não haja interseção entre elas: $R_1, R_2, \ldots, R_j$. Para prever uma nova observação com covariáveis

x, precisamos observar a qual região tal **x** pertence. Seja $R_{\mathbf{x}}$ a região a que **x** pertence. A predição para Z é dada por:

$$\hat{r}(x) = \frac{1}{k} \sum_{i \in R^x} z_i \tag{2.5}$$

A partição de cada covariável recebe o nome de nó, e cada região $R_1, R_2, \ldots, R_j$, recebe o nome de folha.

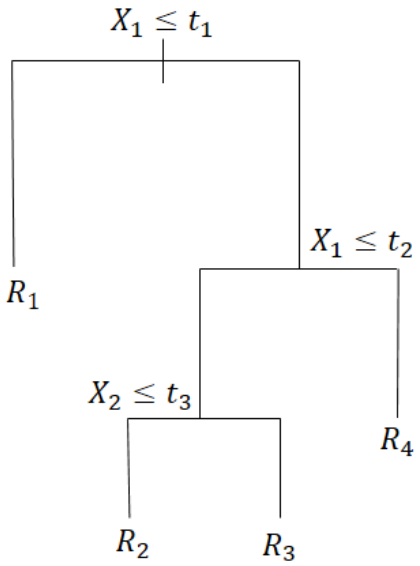


Figura 1 – Exemplo de árvore de regressão.

Para predizer uma nova observação, começamos pelo topo. Se a primeira condição for satisfeita, seguimos para a esquerda e caso contrário, seguimos para a direita. No caso da Figura 1, quando queremos classificar uma nova observação, começamos com a variável X_1 e se tal observação tiver um valor de covariável menor que t_1 será classificada na região R_1, caso contrário encontramos uma nova condição, em que se X_1 for maior que t_2, a observação será classificada em R_4 e se for menor ou igual, encontramos uma nova condição e assim sucessivamente até que esse dado atinja uma folha.

Para criar uma árvore, utilizamos duas etapas: (1) criar uma árvore grande e complexa e (2) podar tal árvore, para evitar o super ajuste.

A etapa 1 é feita de forma recursiva, inicialmente particionando o espaço das covariáveis em duas diferentes regiões. No passo seguinte, particionamos os espaços em regiões menores e assim sucessivamente, até encontrarmos subdivisões suficientes que correspodem à árvore criada. Formalmente, tal árvore divide o espaço das covariáveis em j regiões distintas, $R_1, R_2, \ldots, R_j$. Para determinar tais regiões, utiliza-se divisões binárias de cada uma das covariáveis. Assim, para executar tal divisão, primeiro selecionamos uma covariável $\mathbf{X}$ e um corte t de tal forma que

o espaço das covariáveis seja dividido em $\{X|X < t\}$ e $\{X|X \geq t\}$. A variável e seu respectivo corte são escolhidos de forma a minimizar o erro quadrático médio da partição criada:

$$P(T) = \sum_{R} \sum_{k \in R} (z_k - \hat{z}_R)^2 \tag{2.6}$$

em que $\hat{z}_R$ é a média da resposta para as observações do conjunto treinamento na região R (JAMES *et al.*, 2013).

A etapa 2 consiste em tornar menos complexa a árvore de interesse, e é o que chamamos de poda. Nessa parte do processo, cada nó é retirado da árvore e o EQM no conjunto de validação é calculado. Escolhe-se então o tamanho da árvore que possui menor EQM no conjunto validação. Essa etapa evita que a árvore se ajuste demais aos dados.

2.4.2.2 Florestas Aleatórias

Cada vez que sorteamos uma amostra de treinamento de forma aleatória, as árvores resultantes podem divergir muito umas das outras. Ou seja, tais árvores possuem grande variância. Uma forma de solucionar o problema é utilizar a ideia do bagging, que reduz a variância de métodos estatísticos de predição e assim aumenta o poder de preditivo (JAMES *et al.*, 2013). Esse método é a base para o método de Florestas Aleatórias.

A ideia principal do bagging é retirar várias amostras de treinamento da população para podermos, assim, construir diferentes modelos de predição para cada uma das amostras. Como não temos várias amostras de treinamentos disponíveis, retiramos B amostras do conjunto original de treinamento, a partir do método de bootstrap. Para cada uma das B amostras, são construídas árvores de regressão e calculadas as respectivas médias das predições obtidas em cada uma delas. Seja $g^b(\mathbf{x})$ a função de predição obtida segundo a b-ésima árvore. A função de predição pelo *bagging* (e também florestas aleatórias) é dada pela seguinte equação:

$$g(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} g^b(\mathbf{x}) \tag{2.7}$$

No bagging as árvores não são podadas para diminuir o viés de cada função de predição.

O problema do bagging é que essas B árvores são correlacionadas e g^b's tendem a ser muito próximos uns dos outros, uma vez que são construídas com as mesmas covariáveis. Para diminuir essa correlação entre as árvores é utilizado o método de florestas aleatórias, que segue a mesma ideia do bagging uma vez que as árvores de regressão são construídas a partir da amostra bootstrap do conjunto de treinamento. Mas, ao invés de se utilizar todas as p covariáveis para a construção de uma árvore, apenas m covariáveis (escolhidas de forma aleatória) são utilizadas para a construção de cada nó em cada uma dessas árvores. É costume tomarmos $m \approx \sqrt{d}$ (JAMES *et al.*, 2013).

2.4.3 SpAM - Modelos Aditivos Esparsos

Os Modelos Aditivos Esparsos (RAVIKUMAR *et al.*, 2009) são métodos de regressão que permitem descrever uma relação não linear entre a variável resposta Z e cada covariável $\mathbf{x}$. Neste caso, assume-se que a regressão pode ser decomposta em uma soma de funções suaves de cada covariável. Além disso, esse método dá peso zero para algumas dessas funções.

Mais especificamente, supondo d o número de covariáveis, a regressão $r(\mathbf{x})$ pode ser escrita da seguinte forma:

$$r(\mathbf{x}) = \beta_1 g_1(x_1) + \beta_2 g_2(x_2) + \ldots + \beta_d g_d(x_d) \tag{2.8}$$

Para ajustar tal modelo, busca-se por coeficientes tais que $\sum_{i=1}^{n} |\beta_i| \leq L$, de modo que garante-se esparsidade dos coeficientes. O valor de L pode ser escolhido via validação cruzada.

2.4.4 Séries Espectrais

O método de Séries Espectrais (LEE; IZBICKI, 2016) permite expandir a função de regressão em uma combinação linear de bases espectrais (ϕ_i) de cada uma das covariáveis $\mathbf{x}$. As bases espectrais consistem em uma extensão do método de séries ortogonais, sendo assim mais adequadas para dados com dimensionalidade alta. O método de séries ortogonais consiste em expandir uma função de regressão em termos de uma base ortogonal. Para isso, precisamos escolher uma base ortonormal ϕ_i para o conjunto de funções

$$L^2(\Re^d)) := \{f : [0,1] \to \Re : \int_0^1 f^2(\mathbf{x}) d\mathbf{x} < \infty\} \tag{2.9}$$

O estimador de séries espectrais utiliza como base um conjunto de funções construído baseado nos dados. Assim, no lugar de se usar uma base de Fourier, por exemplo, utilizamos uma base $(\phi_i)_{i\in\mathbb{N}}$ criada com base em $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Dessa forma, o estimador pode ser escrito da seguinte maneira:

$$r(\mathbf{x}) = \sum_{i=1}^{\infty} \beta_i \phi_i(\mathbf{x}) \tag{2.10}$$

Neste caso, não podemos considerar todos os termos da soma, uma vez que isso pode causar o superajuste. Sendo assim, precisamos escolher um ponto de corte, de forma que o modelo de regressão seja bom para os dados. Isso pode ser feito via validação cruzada. Os coeficientes β_i podem ser estimados com base nos dados. Para mais detalhes veja (LEE; IZBICKI, 2014).

CAPÍTULO

3

ESTIMAÇÃO DE DENSIDADES CONDICIONAIS

Neste capítulo são descritas as técnicas utilizadas para a estimação de densidade condicional. A estimação via Kernel é a técnica mais utilizada pelos astrônomos (IZBICKI; LEE; FREEMAN, 2017). A estimação de densidade condicional FlexCode é uma técnica proposta por Izbicki e Lee (2017).

3.1 Função de perda

Na seção 2.2 do capítulo 2, mostramos que para medir a performance de um estimador de regressão, utilizamos a função de risco. Para densidades condicionais, avaliamos a qualidade de um estimador a partir de uma segunda função de risco.

Para um dado estimador $\hat{f}(z|\mathbf{x})$, podemos medir a discrepância entre $\hat{f}(z|\mathbf{x})$ e $f(z|\mathbf{x})$ pelo risco da equação 3.1, que é ponderado por $f(\mathbf{x})$ (IZBICKI; LEE; FREEMAN, 2017).

$$R(\hat{f},f) = \int\int [(\hat{f}(z|\mathbf{x}) - f(z|\mathbf{x}))^2] f(\mathbf{x}) d\mathbf{x} dz \tag{3.1}$$

Para estimar $R(\hat{f},f)$, note que desenvolvendo a fórmula, temos:

$$\int\int [\hat{f}^2(z|\mathbf{x}) - 2\hat{f}(z|\mathbf{x}) f(z|\mathbf{x}) + f^2(z|\mathbf{x})] f(\mathbf{x}) d\mathbf{x} dz =$$

$$= \underbrace{\int\int \hat{f}^2(z|\mathbf{x}) f(x) d\mathbf{x} dz}_{(1)} - \underbrace{\int\int 2\hat{f}(z|\mathbf{x}) f(z,\mathbf{x}) d\mathbf{x} dz}_{(2)} + \underbrace{\int\int f^2(z|\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} dz}_{(3)}$$

Como a terceira parte (3) não depende do estimador, então será substituída por uma constante K. Dessa forma, temos:

$$\underbrace{\int\int \hat{f}^2(z|\mathbf{x})f(\mathbf{x})dxdz}_{(1)} - \underbrace{\int\int 2\hat{f}(z|\mathbf{x})f(z,\mathbf{x})d\mathbf{x}dz}_{(2)} + K$$

Seja $(Z'_1,\mathbf{X}'_1),\ldots,(Z'_n,\mathbf{X}'_n)$ o conjunto de validação. Desenvolvendo as fórmulas acima, temos para a parte (1):

$$\begin{aligned}&\int\int \hat{f}^2(z|\mathbf{x})f(\mathbf{x})dxdz = \int E[\hat{f}^2(z|\mathbf{x})]dz\\ &\approx \int \frac{1}{n}\sum_{i=1}^{n}\hat{f}^2(z|\mathbf{x}'_i)dz \approx \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}^2(z|\mathbf{x}'_i)dz\end{aligned}$$

Para a parte (2) da fórmula que também depende do estimador ficará da seguinte maneira:

$$\int\int 2\hat{f}(z|x)f(z,\mathbf{x})dxdz = 2E[\hat{f}(Z|\mathbf{X})] \approx \frac{2}{n}\sum_{i=1}^{n}\hat{f}(z'_i|\mathbf{x}'_i)$$

Assim, $R(\hat{f},f)$ pode ser aproximado (a menos de uma constante) por:

$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}^2(z|\mathbf{x}'_i)dz + \frac{2}{n}\sum_{i=1}^{n}\hat{f}(z'_i|\mathbf{x}'_i) \tag{3.2}$$

3.2 Estimação de Densidade Condicional via Kernel

Inicialmente descrevemos como estimar uma densidade (não condicional) f com base em uma amostra i.i.d. $Z_1,\ldots,Z_n \sim f$, utilizando o método baseado em Kernel. A partir disso, mostraremos como essa ideia pode ser utilizada para a estimação de densidades condicionais, que é o principal interesse deste trabalho.

A palavra Kernel se refere a qualquer função suave K tal que $K(z) \geq 0$ e

$$\int K(z)dz = 1 \tag{3.3}$$

Aqui utilizamos o Kernel Gaussiano, dado por:

$$K(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} \tag{3.4}$$

Dada uma amostra i.i.d. $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$ e um número positivo h, que é chamado de banda (bandwith), o estimador da densidade Kernel de $f(z)$, aplicado no ponto z, é dado pela seguinte expressão:

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{z-Z_i}{h}\right) \tag{3.5}$$

Dessa forma, se tivéssemos várias observações $\mathbf{X} = \mathbf{x}$, poderíamos utilizar os z's correspondentes da equação acima para estimar $f(z|\mathbf{X} = \mathbf{x})$. Mais especificamente, supondo que $Z'_1, \ldots, Z'_l$ são tais que $\mathbf{x}'_i = \mathbf{x}\ \forall i = 1, \ldots, l$, poderíamos estimar $f(z|\mathbf{x})$ via

$$\hat{f}(z|\mathbf{x}) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{h}K\left(\frac{z-Z'_i}{h}\right) \tag{3.6}$$

Porém, em geral não temos várias observações com os valores exatos de $\mathbf{x}$. Logo, para cada $\mathbf{x}$ de interesse, selecionamos k amostras de forma que estas tenham valores próximos de $\mathbf{x}$, e então, a estimativa de $f(z|\mathbf{x})$ é dada por:

$$\hat{f}(z|\mathbf{x}) = \frac{1}{k}\sum_{i\in\mathscr{N}_x}\frac{1}{h}K\left(\frac{z-Z_i}{h}\right) \tag{3.7}$$

Neste caso, a soma não percorre todos os valores da amostra, mas sim os índices dos k vizinhos mais próximos a $\mathbf{x}$. O valor da banda h e o número de vizinhos k são tuning parameter e dessa forma são escolhidos por validação cruzada, minimizando a estimativa da função de perda, dada pela equação 3.2.

3.3 FlexCode

Seja $(\mathbf{X}_1, Z_1), \ldots, (\mathbf{X}_n, Z_n)$, em que $\mathbf{x}$ são as covariáveis; $\mathbf{x} \in \Re^d$, e a variável resposta $Z \in [0,1]$. Em linhas gerais, para estimar $f(z|\mathbf{x})$, o FlexCode propõe que se expanda tal função em uma base ortonormal $(\phi_i)_{i\in\mathbb{N}}$ para funções em $\Re$. Cada coeficiente dessa expansão pode ser estimado diretamente através de uma regressão. Existem muitas bases ortonormais que podem ser escolhidas para capturar qualquer forma da função de interesse. Para este trabalho, será utilizada a base de Fourier para modelar $f(z|\mathbf{x})$, que é descrita nas equações 3.8, 3.9 e 3.10:

$$\phi_1(z) = 1 \tag{3.8}$$

$$\phi_{2i+1}(z) = \sqrt{2}\sin(2\pi i z), i \in \mathbb{N} \tag{3.9}$$

$$\phi_{2i}(z) = \sqrt{2}\cos(2\pi i z), i \in \mathbb{N} \tag{3.10}$$

Mais especificamente, fixada a base $(\phi_i)_{i\in\mathbb{N}} \in \mathfrak{R}$, a expansão é dada por

$$f(z|\mathbf{x}) = \sum_{i\in\mathbb{N}} \beta_i(\mathbf{x})\phi_i(z), \tag{3.11}$$

em que

$$\beta_i(\mathbf{x}) = \int_{\mathfrak{R}} \phi_i(z) f(z|\mathbf{x}) dz = \mathbb{E}[\phi_i(Z)|\mathbf{x}] \tag{3.12}$$

ou seja, cada coeficiente dessa expansão pode ser escrito como uma esperança condicional e, dessa forma, ser estimado via função de regressão. Assim, para um i fixo, podemos estimar $\beta_i(\mathbf{x})$ através da regressão de $\phi_i(z)$ em $\mathbf{x}$, utilizando a amostra $(\mathbf{X}_1, \phi_i(Z_1)), \ldots, (\mathbf{X}_n, \phi_i(Z_n))$. Para a estimativa desses coeficientes, qualquer método de regressão pode ser utilizado.

O estimador FlexCode da função $f(z|\mathbf{x})$ é definido como:

$$\hat{f}(z|\mathbf{x}) = \sum_{i=1}^{I} \hat{\beta}_i(\mathbf{x})\phi_i(z), \tag{3.13}$$

em que $\hat{\beta}_i(\mathbf{x}) = \hat{\mathbb{E}}[\phi_i(Z)|\mathbf{x}]$ é obtido via estimação por regressão. O corte I da expansão de série é um tuning parameter, que controla o balanço entre viés e variância na densidade estimada final. Para a escolha de I, utilizamos validação cruzada, novamente minimizando a estimativa do risco dada pela equação 3.2.

CAPÍTULO

4

ESTIMANDO $g(z)$

Este capítulo tem por objetivo ilustrar, através de dados simulados, como funcionam as três técnicas de estimação de uma função de quantidade desconhecida: regressão *plug-in* $g(\hat{E}[Z|\mathbf{x}])$, regressão direta $\hat{E}[g(Z)|\mathbf{x}]$ e densidade condicional $\int g(z)\hat{f}(z|\mathbf{x})dz$. Também será apresentada uma descrição detalhada de cada método. Como critério de comparação, utilizaremos o risco estimado para regressão (Seção 2.2) e para densidade condicional (Seção 3.1) desses métodos.

A qualidade destes estimadores foi avaliada via o seu risco estimado. Mais precisamente, para avaliar o desempenho de um estimador $h(\mathbf{X})$ de $g(Z)$, utilizaremos o risco quadrático, dado pela expressão 4.1:

$$R(h) = \mathbb{E}[(Y - h(\mathbf{X}))^2], \tag{4.1}$$

em que $Y = g(Z)$. A intenção é que $R(h)$ seja baixo, pois assim teremos predições boas em observações novas.

4.1 Métodos

1. O estimador *plugin* $g(\widehat{\mathbb{E}}[Z|\mathbf{x}])$, i.e., inicialmente cria-se $\widehat{\mathbb{E}}[Z|\mathbf{x}]$, um estimador da função de regressão de Z em $\mathbf{x}$, e pluga-se ele na função g. Esse é um estimador tradicional, e é o mais utilizado na literatura. Tal método estima valores de Z com base nas covariáveis $\mathbf{x}$ através da estimação da regressão $\widehat{\mathbb{E}}[Z|\mathbf{x}]$. Para essa primeira abordagem diversos métodos de regressão foram propostos (Capítulo 2).

2. Estimador via densidade condicional, $\int_z g(z)\widehat{f}(z|\mathbf{x})dz$, que é motivado pelo fato de que $\mathbb{E}[g(Z)|\mathbf{x}] = \int_z g(z)f(z|\mathbf{x})dz$. Isto é, a esperança condicional pode ser escrita como uma integral da função $g(z)$, multiplicada pela densidade condicional. Dessa forma, estima-se a função substituindo a mesma na densidade condicional e, assim, podemos ter uma

estimativa da esperança dessa função. Tal estimador foi inicialmente proposto por (MANDELBAUM *et al.*, 2008). Para esse método, alguns estimadores de f foram comparados (Capítulo 3).

3. Estimador *direto* $\widehat{\mathbb{E}}[g(Z)|\mathbf{x}]$, i.e., estima-se a função de regressão de $g(Z)$ em $\mathbf{x}$. Esse método é diferente do método de estimação *plugin* $g(\widehat{\mathbb{E}}[Z|\mathbf{x}])$, já que neste é feita a regressão direta de g em $\mathbf{x}$. Esse método estima valores de g com base nas covariáveis $\mathbf{x}$, através da regressão $\widehat{\mathbb{E}}[g(Z)|\mathbf{x}]$. Destacamos que esta abordagem é inédita na literatura. Nesta terceira abordagem, também foram utilizados diversos métodos de regressão descritos no Capítulo 2.

4.1.1 Conjunto de dados

Os dados foram gerados utilizando o Software R. O conjunto apresenta uma variável resposta Z, uma covariável X e n=1000 amostras. A variável X foi gerada a partir de uma distribuição Uniforme no intervalo de -1 a 1, e a variável Z foi gerada a partir de uma distribuição Normal da mistura $\frac{1}{2}N(x,0,25^2)+\frac{1}{2}N(-x,0,25^2)$. Como o objetivo do trabalho está não só no valor de Z mas também em funções do Z, vamos utilizar $g(Z)=Z^2$ para ilustrar os métodos investigados. A Figura 2 apresenta graficamente X versus Z e X versus $g(Z)=Z^2$, respectivamente.

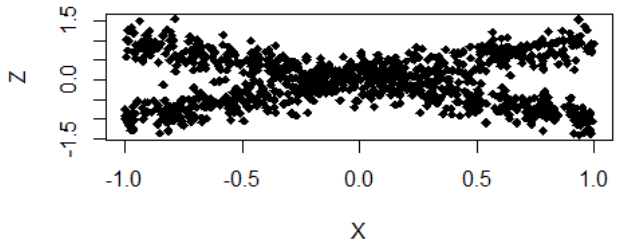


Figura 2 – Gráfico de dispersão x versus z.

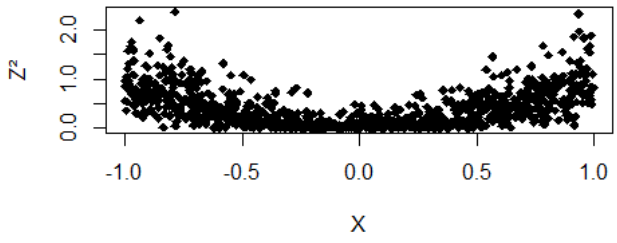


Figura 3 – Gráfico de dispersão x versus z^2.

O conjunto de dados foi separado em três partes: treinamento (60%), validação (20%) e teste (20%).

4.1.2 Regressão Plug-in $g(\hat{E}[Z|x])$

O primeiro método investigado consiste em estimar z com base em $\mathbf{x}$, através da estimação da regressão $\mathbb{E}[Z|\mathbf{x}]$. A título de ilustração, utilizamos o método de regressão K-Nearest Neighbors (KNN), dado pela equação 2.4 do Capítulo 2. Neste caso, o melhor número de vizinhos encontrado foi de k=84, ou seja, esse é o número de vizinhos de X utilizado para estimar $\mathbb{E}[Z|\mathbf{x}]$. Para medir o quanto essas obervações estão próximas de X, foi utilizada a distância euclidiana.

A partir disso, foi calculada a curva estimada da função de regressão de $E[Z|x]$, apresentada em vermelho na Figura 4. Como é possível perceber, essa curva estimada fica próxima da reta $z = 0$ (a verdadeira regressão neste problema). Dessa forma, para todos os valores de x, o valor predito de z (e portanto z^2) será sempre o mesmo e próximo de zero.

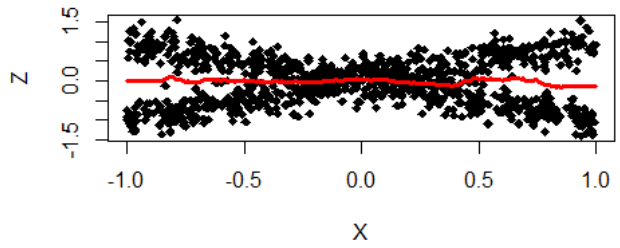


Figura 4 – Gráfico de dispersão x versus Z com a curva estimada (em vermelho) da função regressão $E[Z|x]$, $\hat{E}[Z|x]$.

Para o gráfico de x versus $g(Z) = Z^2$, na Figura 5, permite notar que $g(\hat{E}[Z|x])$ não é uma boa aproximação para $g(Z) = Z^2$. Isso ocorre pois a distribuição dos dados em estudo é bimodal.

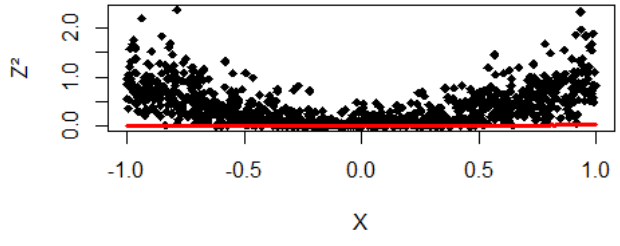


Figura 5 – Gráfico de dispersão X versus Z^2 com a curva estimada $g(\hat{E}[Z|x])$.

Para essa função de regressão Plug-in $g(\hat{E}[Z|x])$, o risco estimado foi:

$$\hat{R}(g(\hat{E}[Z|x])) = 0,317$$

4.1.3 Regressão direta $\hat{E}[g(Z)|x]$

Nesta seção, foi testada a abordagem de estimar diretamente $g(z)$ em x através da função de regressão $E[g(Z)|x]$. O estimador da função de regressão de $g(z)$ e x é dado pela equação 2.4 da Seção 2, substituindo z por $g(z) = z^2$.

O critério para a escolha no melhor número de vizinhos foi o mesmo utilizado na Seção 4.1.2, mas aqui o melhor número de vizinhos utilizado para prever $g(Z) = Z^2$ em X foi de k=23.

A Figura 6 evidencia que esse método tem maior poder preditivo quando comparado com o estimador $g(\hat{E}[Z|x])$, já que a curva dos valores estimados prevê z^2 com base em **x** de modo muito mais preciso que a Figura 5.

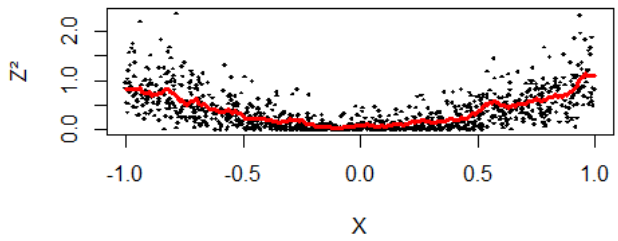


Figura 6 – Gráfico de dispersão X versus Z^2 com a curva estimada da função regressão $E(g[Z|x])$.

O risco estimado deste segundo estimador confirma que a função de regressão direta $\hat{E}[g(Z)|x])$ é melhor que a função de regressão Plug-in.

$$\hat{R}(\hat{E}[g(Z)|x]) = 0,077$$

4.1.4 Densidade Comdicional $\int g(z)\hat{f}(z|x)dz$

Dadas as covariáveis x, é possível calcular a densidade condicional e plugá-la na integral. Uma vez que encontrado o melhor número de vizinhos para estimar a densidade, e a melhor banda, estimamos, a título de ilustração, as densidades para x assumindo valores -0.5, 0 e 0.5 ($f(z|x = 0.5)$, $f(z|x = 0)$) e $f(z|x = -0.5)$, como é possível observar nas Figuras 7, 8 e 9.

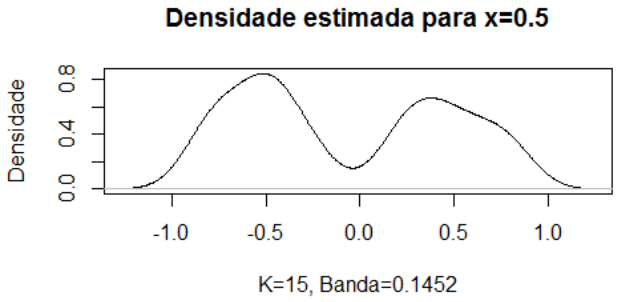


Figura 7 – Densidade estimada $f(z|x=0.5)$.

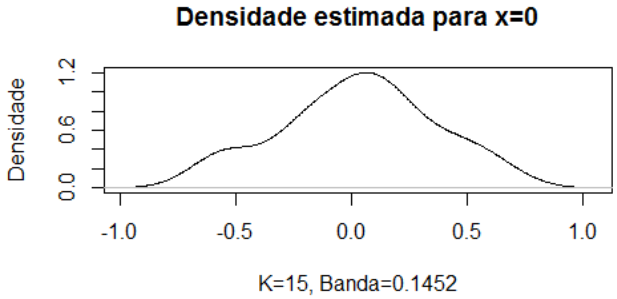


Figura 8 – Densidade estimada $f(z|x=0)$.

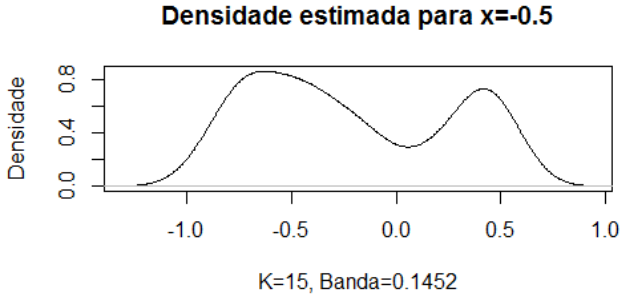


Figura 9 – Densidade estimada $f(z|x=-0.5)$.

Aqui, foi calculado o risco da densidade Plug-in $\int g(z)\hat{f}(z|x)dz$:

$$\hat{R}\left(\int g(z)\hat{f}(z|x)dz\right) = 0,079$$

4.1.5 Comparação dos três métodos utilizados

Para ilustrar os resultados dos riscos para as três abordagens acima, a Figura 10 apresenta o risco estimado das três funções de predição, $g(\hat{E}[Z|x])$, $\hat{E}[g(Z)|x]$ e $\int g(z)\hat{f}(z|x)dz$, com um intervalo de 95% de confiança.

Como foi observado, o risco da função de regressão Plug-in $g(\hat{E}[Z|x])$ foi maior que o risco quando comparado com as abordagens de Regressão direta $\hat{E}[g(Z)|x]$ e de Densidade Plug-in $\int g(z)\hat{f}(z|x)dz$. Essas duas últimas abordagens apresentaram uma performance muito parecida, uma vez que o risco de ambas foram bem próximos.

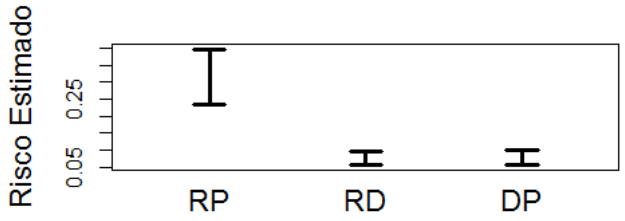


Figura 10 – Risco estimado para os três estimadores, RP-regressão Plug-in, RD-regressão direta e DP-densidade Plug-in.

CAPÍTULO

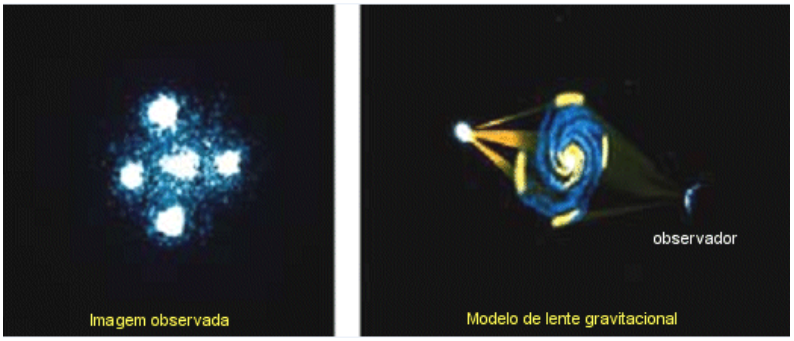
5

EXEMPLO COM DADOS REAIS

Nesse capítulo é apresentada uma explicação mais detalhada do problema cosmológico enunciada na introdução e também é feita uma análise de dados reais dos conjuntos DEEP2 e Spectroscopic utilizando os métodos investigados, assim como várias funções de interesse $g(z)$.

5.1 Lentes Gravitacionais

A abordagem de estimar $g(z)$ para m novas observações, com covariáveis conhecidas $\mathbf{x}$, pode ser utilizada no problema da estimação do efeito de lentes gravitacionais (MANDELBAUM *et al.*, 2008), que nada mais são do que uma espécie de amplificador de imagens, formadas por campos gravitacionais que, situados entre o objeto que se observa e a terra, funcionam como uma lente. Assim, essas massas gravitacionais podem apresentar uma imagem distorcida da galáxia original. A Figura 11 apresenta como se comporta uma Lente Gravitacional.



*Fonte: Clube de Astronomia, http://cacarlsagan.blogspot.com.br/, 2016.

Figura 11 – À esquerda da figura, podemos observar a imagem de diversas galáxias. À direita da figura percebemos que se trata de apenas uma galáxia, em que a massa gravitacional, funcionando como uma lente, está distorcendo a imagem quando esta chega no observador, parecendo, assim, ser mais de uma galáxia.

A função que caracteriza o sistema de lentes é denotada por $\Sigma(z_l, z_s)$ e chamada de

densidade crítica da superfície de um par de galáxias – lente e de fundo – com redshifts[1] z_l e z_s, respectivamente. Assim, para cada z_l fixo, em astronomia, é importante estimar $\Sigma(z_l, z_s)$ para um conjunto de galáxias de fundo com redshifts desconhecidos. Uma vez que $g_l(z) = \Sigma(z_l, z)$ são estimadas, utilizam-nas para estimar parâmetros cosmológicos importantes para caracterizar a evolução do universo. Dessa forma, é importante ter estimativas precisas desta quantidade. Para estimar $g_l(z)$, utiliza-se *covariáveis fotométricas* $\mathbf{x}$ que nos trazem informação sobre z.

Neste trabalho, compararemos a eficácia de diversas técnicas para estimar essa quantidade. Note que para lente l há uma função g_l diferente.

5.2 Funções utilizadas

A título de ilustração os métodos foram comparados utilizando algumas funções de z. Tais funções foram: $g_1(z) = z^2$, $g_2(z) = \sin(2\pi z)$, $g_3(z) = 5z + 1$, $g_4(z) = (z-1)^7$, $g_5(z) = I_{(0.2;0.5)}(x)$. Também utilizamos a função do Sigma crítico com diferentes valores de z_l: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, $g_6f(z) = \Sigma(0.19, z_s)$, $g_6g(z) = \Sigma(0.22, z_s)$, $g_6h(z) = \Sigma(0.24, z_s)$, $g_6i(z) = \Sigma(0.27, z_s)$ e $g_6j(z) = \Sigma(0.30, z_s)$.

A Tabela 3 apresenta os métodos de regressão e de densidade condicional utilizados, tais como suas respectivas siglas. Os métodos de regressão plug-in se referem à abordagem (1) do Capítulo 1; nestes utilizamos os métodos de regressão KNN e Florestas (RP_Knn e RP_Fl). O mesmo foi feito com a regressão direta (abordagem (2) também do Capítulo 1) (RD_Knn e RD_Fl). A terceira abordagem é referente à densidade condicional, em que os métodos utilizados foram KNN, Florestas, SpAM e Series para estimar os coeficientes $\beta_i(\mathbf{x})$ no FlexCode (DC_Knn, DC_Fl, DC_SpAM e DC_Ser). Temos também a densidade condicional tradicional (DC_Tradi), dada pelo método baseado em Kernel.

[1] Redshift é essencialmente uma medida da distância entre a galáxia e a Terra.

Tabela 3 – Métodos de estimação e seus respectivos símbolos.

Métodos	Símblolos
Regressão Plug-in KNN	RP_Knn
Regressão Plug-in Florestas	RP_Fl
Regressão Plug-in SpAM	RP_SpAM
Regressão Plug-in Series	RP_Ser
Regressão Direta KNN	RD_Knn
Regressão Direta Florestas	RD_Fl
Regressão Direta SpAM	RD_SpAM
Regressão Direta Series	RD_Ser
Densidade Condicional KNN	DC_Knn
Densidade Condicional Florestas	DC_Fl
Densidade Condicional SpAM	DC_SpAM
Densidade Condicional Series	DC_Ser
Densidade Condicional Tradicional KNN	DC_Tradi

5.3 Viés e Variância

Além do risco estimado para avaliar a performance de cada método, as medidas de variância e viés (MANDELBAUM *et al.*, 2008) também foram utilizadas para avaliar a qualidade das funções associadas ao Σ critico. Tais medidas de performance são muito utilizadas em problemas de lentes pelos astrônomos. Seja $(\mathbf{X}''_1, Z''_1), \ldots, (\mathbf{X}''_n, Z''_n)$ o conjunto de teste, temos que:

$$V = \frac{\left(\sum_{i=1}^{n''} \sqrt{(g^2(z''_i)h^2(\mathbf{x}''_i))}\right)^2}{\sum_{i=1}^{n''} g^2(z''_i) \sum_{i=1}^{n''} h^2(\mathbf{x}''_i)} \tag{5.1}$$

$$B = \left| \frac{\sum_{i=1}^{n''} g^2(z''_i)h^2(\mathbf{x}''_i)}{\sum_{i=1}^{n''} h^2(\mathbf{x}''_i)h^2(\mathbf{x}''_i)} - 1 \right| \tag{5.2}$$

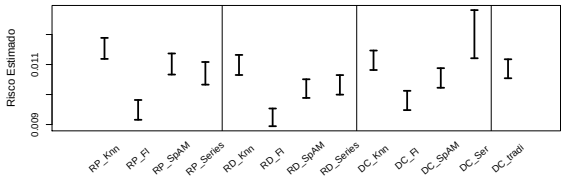
em que $h(\mathbf{x})$ é a função de predição criada e avaliada em $\mathbf{x}$ e $g(z)$ é a verdadeira função para o problema avaliada em g. Vale lembrar que são medidas elaboradas pelos astrônomos e não são o viés e a variância que usualmente utilizamos em problemas de estatística. O ideal é que o viés seja próximo de zero e a variância assuma valores próximos que 1.

5.4 Dados DEEP2 EGS Region

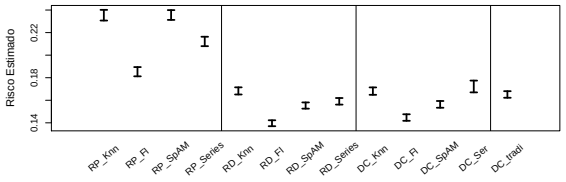
Os dados utilizados nesse trabalho são fotométricos de DEEP2 EGS Region (WEINER *et al.*, 2005), em que temos cinco covariáveis representando as cores das galáxias e a variável resposta corresponde ao redshift das mesmas. Esse conjunto tem um total de 1383 galáxias. As amostras foram divididas em 60% para treinamento, 20% para validação e 20% para teste.

5.4.1 Resultados

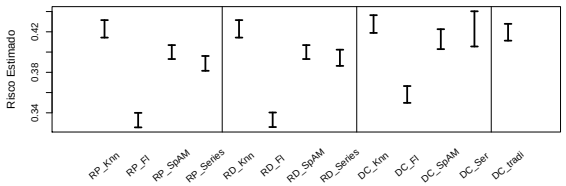
As Figuras 12, 13 e 14 apresentam o risco estimado para cada método em cada uma das funções de interesse, bem como seus respectivos intervalos de confiança de 95% de confiança.



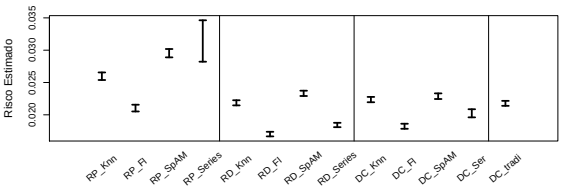
(a) Risco da função $g_1(z) = z^2$



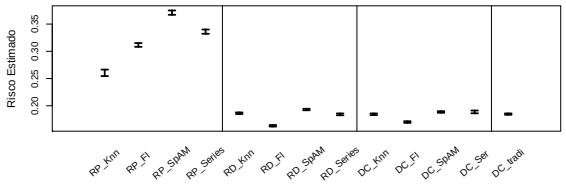
(b) Risco da função $g_2(z) = \sin(2\pi z)$



(c) Risco da função $g_3(z) = 5z+1$

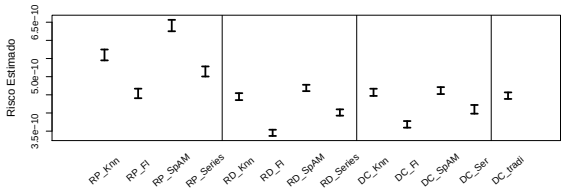


(d) Risco da função $g_4(z) = (z-1)^7$

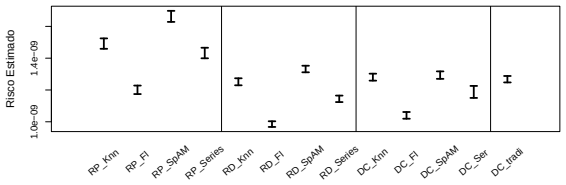


(e) Risco da função $g_5(z) = I_{(0.2;0.5)}(x)$

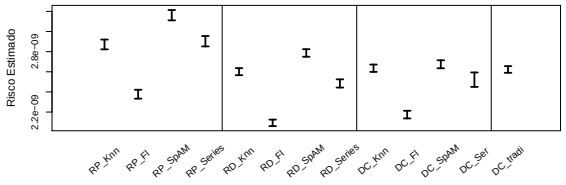
Figura 12 – Risco estimado para os nove estimadores nas funções: $g_1(z) = z^2$, $g_2(z) = \sin(2\pi z)$, $g_3(z) = 5z+1$, $g_4(z) = (z-1)^7$, $g_5(z) = I_{(0.2;0.5)}(x)$, respectivamente.



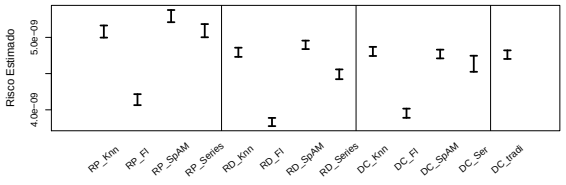
(a) Risco da função $g_6a(z) = \Sigma(0.05, z_s)$



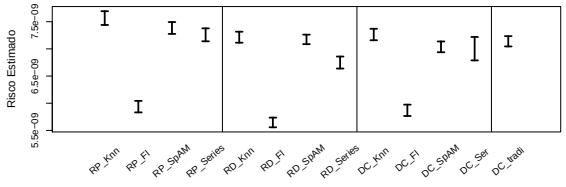
(b) Risco da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) [Risco da função $g_6c(z) = \Sigma(0.11, z_s)$

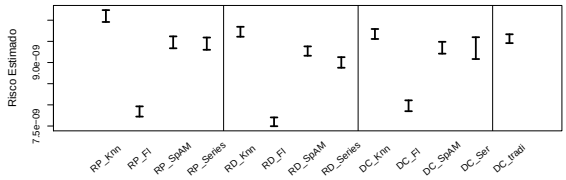


(d) Risco da função $g_6d(z) = \Sigma(0.13, z_s)$

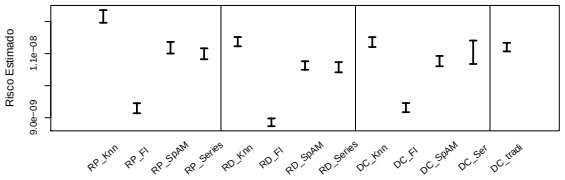


(e) Risco da função $g_6e(z) = \Sigma(0.16, z_s)$

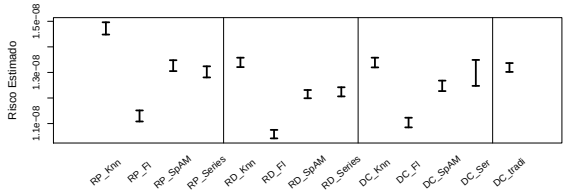
Figura 13 – Risco estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



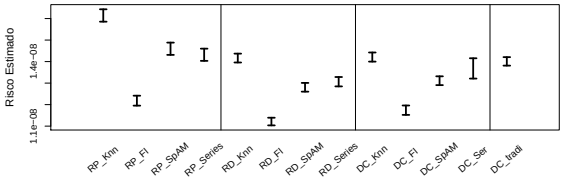
(a) Risco da função $g_6 f(z) = \Sigma(0.19, z_s)$



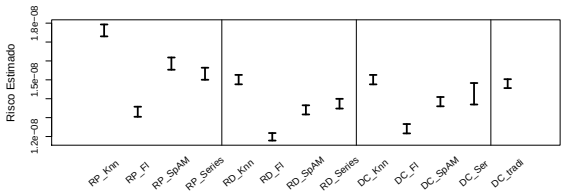
(b) Risco da função $g_6 g(z) = \Sigma(0.22, z_s)$



(c) Risco da função $g_6 h(z) = \Sigma(0.24, z_s)$



(d) Risco da função $g_6 i(z) = \Sigma(0.27, z_s)$



(e) Risco da função $g_6 j(z) = \Sigma(0.30, z_s)$

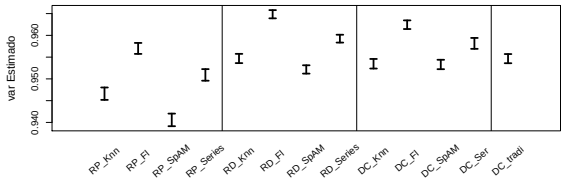
Figura 14 – Risco estimado para os nove estimadores nas funções: $g_6 f(z) = \Sigma(0.19, z_s)$, $g_6 g(z) = \Sigma(0.22, z_s)$, $g_6 h(z) = \Sigma(0.24, z_s)$, $g_6 i(z) = \Sigma(0.27, z_s)$, $g_6 j(z) = \Sigma(0.30, z_s)$, respectivamente.

É possível observar a partir das Figuras 12, 13 e 14 que, de um modo geral, os métodos baseados em regressão plug-in, com exceção de florestas, apresentaram os maiores riscos, evidenciando uma não adequação para esse problema. A regressão direta, na maioria dos casos, apresentou um bom desempenho quando utilizado o método de florestas aleatórias. Já no caso em que utilizamos KNN, SpAM e Séries Espectrais para estimar a função de regressão, tal método não apresentou uma boa performance.

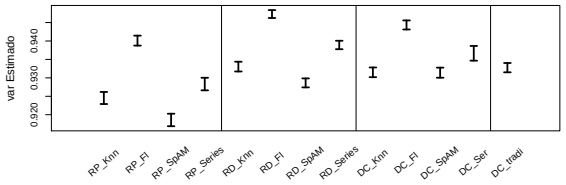
Nesse sentido, métodos baseados em densidade condicional por meio de florestas aleatórias também levaram a resultados satisfatórios. Em contrapartida, métodos via KNN, SpAM e séries espectrais não apresentaram boa execução, com riscos elevados quando comparado com os demais métodos.

Na maioria dos casos, o método de densidade condicional tradicional, que é o mais utilizado pelos astrônomos, não obteve um bom comportamento, apresentando valores altos para o risco.

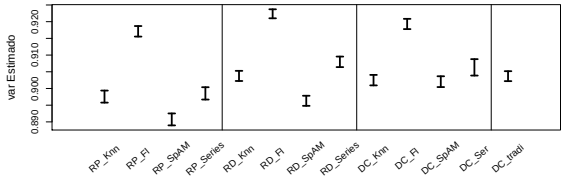
As Figuras 15 e 16 apresentam a variância estimada para cada método em cada uma das funções de interesse, bem como seus respectivos intervalos de confiança de 95% de confiança.



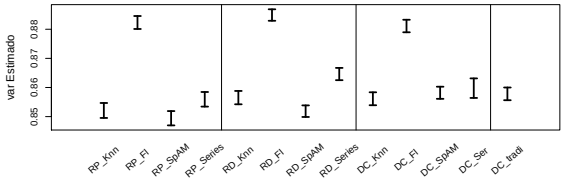
(a) Variância da função $g_6a(z) = \Sigma(0.05, z_s)$



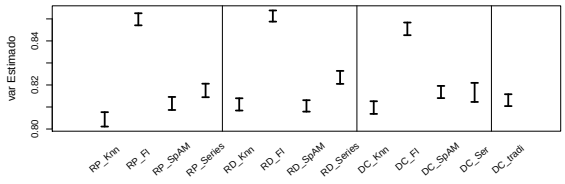
(b) Variância da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) Variância da função $g_6c(z) = \Sigma(0.11, z_s)$

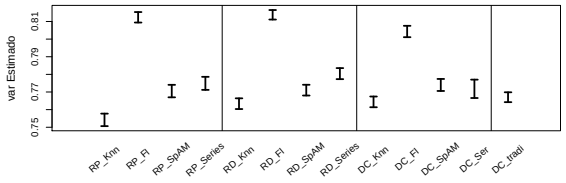


(d) Variância da função $g_6d(z) = \Sigma(0.13, z_s)$

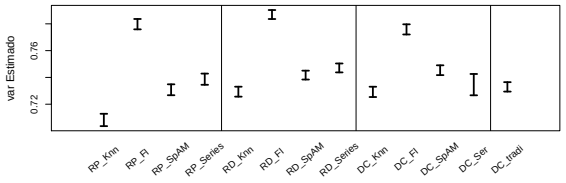


(e) Variância da função $g_6e(z) = \Sigma(0.16, z_s)$

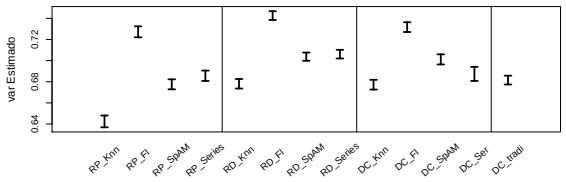
Figura 15 – Variância estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



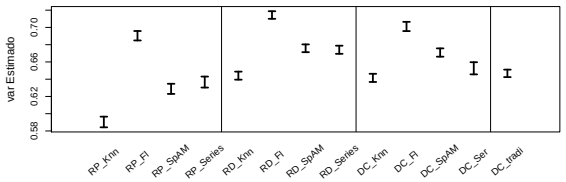
(a) Variância da função $g_6f(z) = \Sigma(0.19, z_s)$



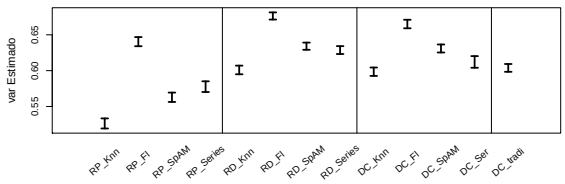
(b) Variância da função $g_6g(z) = \Sigma(0.22, z_s)$



(c) Variância da função $g_6h(z) = \Sigma(0.24, z_s)$



(d) Variância da função $g_6i(z) = \Sigma(0.27, z_s)$

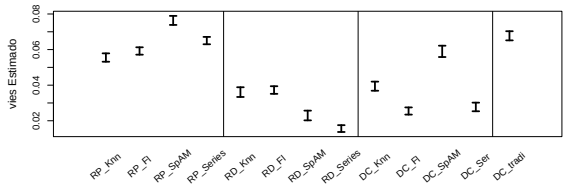


(e) Variância da função $g_6j(z) = \Sigma(0.30, z_s)$

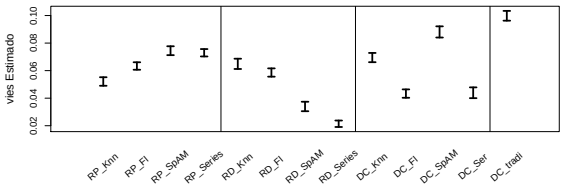
Figura 16 – Variância estimada para os nove estimadores nas funções: $g_6f(z) = \Sigma(0.19, z_s)$, $g_6g(z) = \Sigma(0.22, z_s)$, $g_6h(z) = \Sigma(0.24, z_s)$, $g_6i(z) = \Sigma(0.27, z_s)$, $g_6j(z) = \Sigma(0.30, z_s)$, respectivamente.

Em geral, foi possível notar a partir das Figuras 15 e 16 que os métodos de plug-in utilizando KNN e SpAM, para estimar a função de regressão, apresentaram valores baixos para a variância, evidenciando que esse método não é adequado, quando comparado com os demais. Neste caso, métodos baseados em densidades condicional e regressão direta que utilizam florestas levaram a bons resultados, apresentando uma medida de variância próxima de 1, o que corrobora com a teoria do risco. Para essa medida, destacamos também a regressão plug-in via florestas aleatórias. Concordando com a medida de risco, a densidade condicional tradicional, na a maioria dos casos, apresentou um valor de variância baixo, não sendo o método mais adequado para este problema.

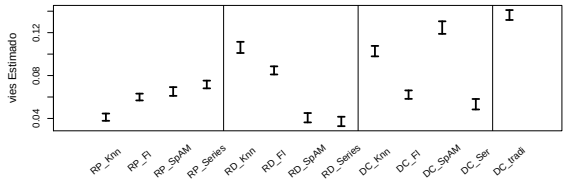
Nas Figuras 15 e 16 apresentamos o viés para cada método em cada uma das funções de interesse, bem como seus respectivos intervalos de confiança de 95% de confiança.



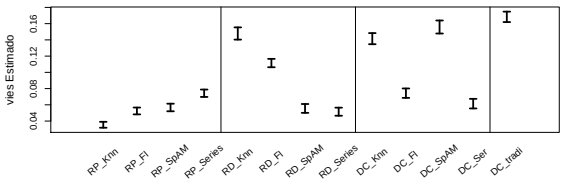
(a) Viés da função $g_6a(z) = \Sigma(0.05, z_s)$



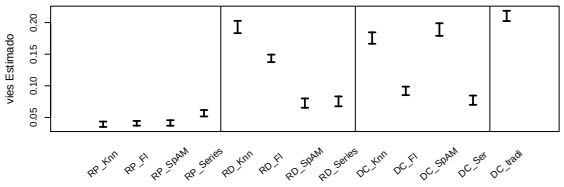
(b) Viés da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) Viés da função Viés da função $g_6c(z) = \Sigma(0.11, z_s)$

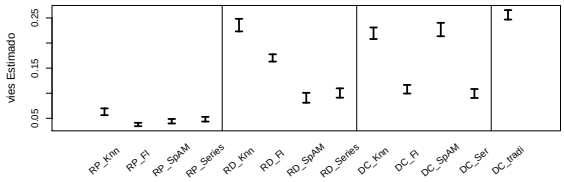


(d) Viés da função $g_6d(z) = \Sigma(0.13, z_s)$

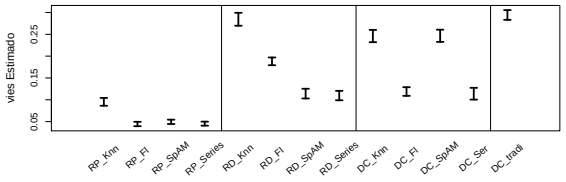


(e) Viés da função $g_6e(z) = \Sigma(0.16, z_s)$

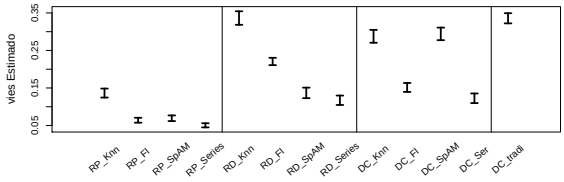
Figura 17 – Viés estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



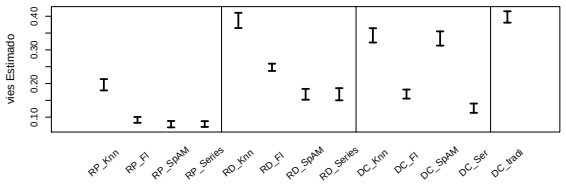
(a) Viés da função $g_6 f(z) = \Sigma(0.19, z_s)$



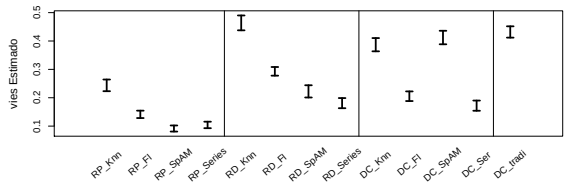
(b) Viés da função $g_6 g(z) = \Sigma(0.22, z_s)$



(c) Viés da função $g_6 h(z) = \Sigma(0.24, z_s)$



(d) Viés da função $g_6 i(z) = \Sigma(0.27, z_s)$



(e) Viés da função $g_6 j(z) = \Sigma(0.30, z_s)$

Figura 18 – Viés estimado para as nove estimadores nas funções: $g_6 f(z) = \Sigma(0.19, z_s)$, $g_6 g(z) = \Sigma(0.22, z_s)$, $g_6 h(z) = \Sigma(0.24, z_s)$, $g_6 i(z) = \Sigma(0.27, z_s)$, $g_6 j(z) = \Sigma(0.30, z_s)$, respectivamente.

Para as medidas de viés, verificamos que todos os métodos apresentaram valores próximos de zero, como se nota nas Figuras 17 e 18. Neste caso, destacamos os métodos de regressão plug-in via florestas aleatórias, séries espectrais e SpAM, que apresentaram um viés inferior aos demais. Em alguns casos, a regressão direta utilizando séries espectrais e SpAM apresentaram uma boa performance. Isso também acontece quando se trata de densidade condicional via séries espectrais e florestas aleatórias.

Para a maioria dos resultados, densidade condicional, utilizando séries espectrais, obteve uma variabilidade maior quando comparada com os demais métodos. Isso é ruim, uma vez que indica uma certa instabilidade desse método. A metodologia de densidade mais utilizado pelos astrônomos (densidade condicional tradicional) também não obteve nenhuma vantagem em relação as demais, e em geral apresentou resultados insatisfatórios.

5.5 Dados Sheldon 2012

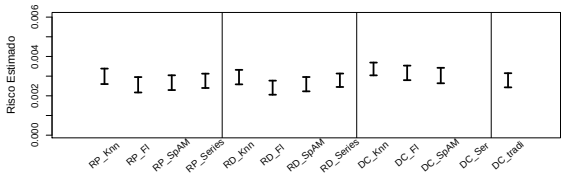
Além dos dados DEEP2, também utilizamos o conjunto Sheldon 2012 (SHELDON *et al.*, 2012; FREEMAN; IZBICKI; LEE, 2017; IZBICKI; LEE; FREEMAN, 2014) para comparar os métodos propostos neste trabalho. Tal conjunto de dados inclui 435875 galáxias com dez covariáveis fotométricas $\mathbf{x}$ e z representando o redshift de tais galáxias.

Devido ao tamanho do conjunto de dados, o processo tornou-se computacionalmente demorado, o que justifica a realização de uma amostragem de apenas duas mil galáxias no experimento. Neste caso, as amostras também foram divididas em 60% para treinamento, 20% para validação e 20% para teste.

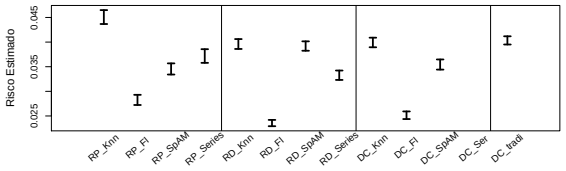
5.5.1 Resultados

As figuras a seguir apresentam os gráficos da estimativa do risco, viés e variância para as funções descritas na Seção 5.2. Esse gráficos também apresentam um intervalo de 95% de confiança para tais quantidades.

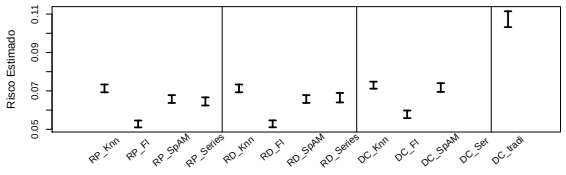
Em todos os casos, o estimador de densidade condicional baseado em séries espectrais nos trouxe resultados ruins, apresentando valores de risco, viés e variância muito discrepante dos demais métodos. Além disso, tais resultados foram extremamente instáveis, com um desvio padrão elevado e consequentemente levando a grandes intervalos de confiança. Em virtude disso, optamos por não utilizar os resultados de densidade condicional via séries espectrais nas análises gráficas.



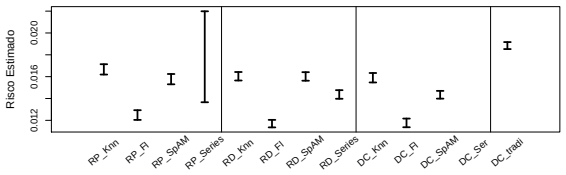
(a) Risco da função $g_1(z) = z^2$



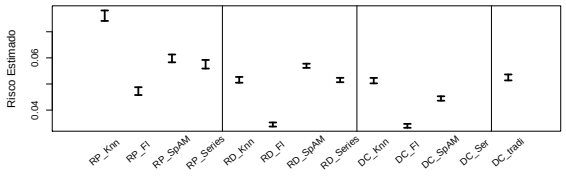
(b) Risco da função $g_2(z) = \sin(2\pi z)$



(c) Risco da função $g_3(z) = 5z + 1$

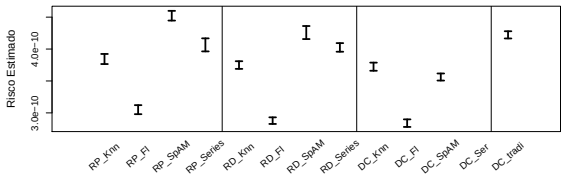


(d) Risco da função $g_4(z) = (z-1)^7$

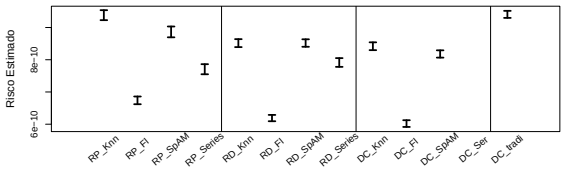


(e) Risco da função $g_5(z) = I_{(0.2;0.5)}(x)$

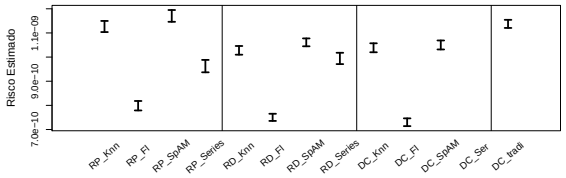
Figura 19 – Risco estimado para os nove estimadores nas funções: $g_1(z) = z^2$, $g_2(z) = \sin(2\pi z)$, $g_3(z) = 5z+1$, $g_4(z) = (z-1)^7$, $g_5(z) = I_{(0.2;0.5)}(x)$, respectivamente.



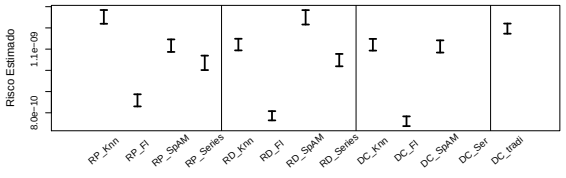
(a) Risco da função $g_6a(z) = \Sigma(0.05, z_s)$



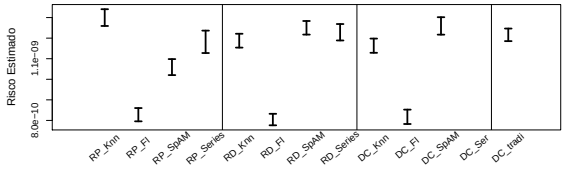
(b) Risco da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) [Risco da função $g_6c(z) = \Sigma(0.11, z_s)$

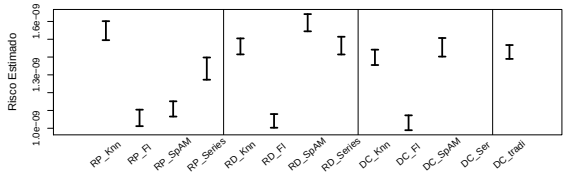


(d) Risco da função $g_6d(z) = \Sigma(0.13, z_s)$

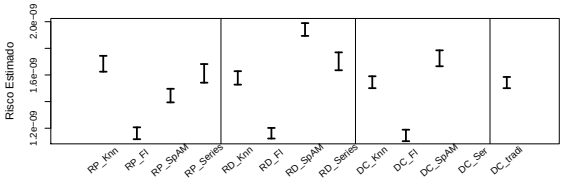


(e) Risco da função $g_6e(z) = \Sigma(0.16, z_s)$

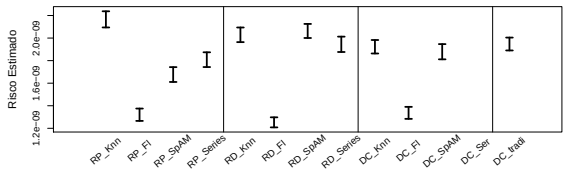
Figura 20 – Risco estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



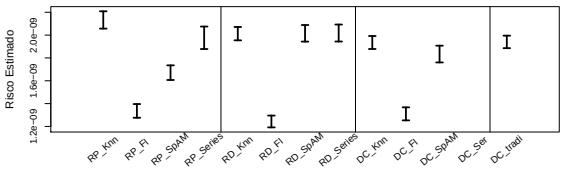
(a) Risco da função $g_6 f(z) = \Sigma(0.19, z_s)$



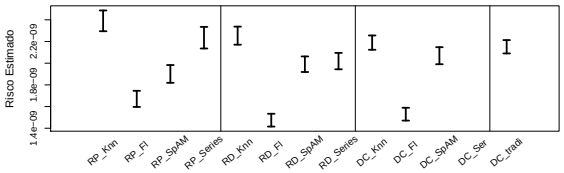
(b) Risco da função $g_6 g(z) = \Sigma(0.22, z_s)$



(c) Risco da função $g_6 h(z) = \Sigma(0.24, z_s)$



(d) Risco da função $g_6 i(z) = \Sigma(0.27, z_s)$

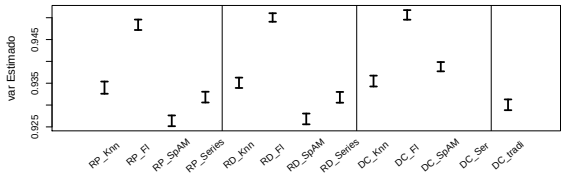


(e) Risco da função $g_6 j(z) = \Sigma(0.30, z_s)$

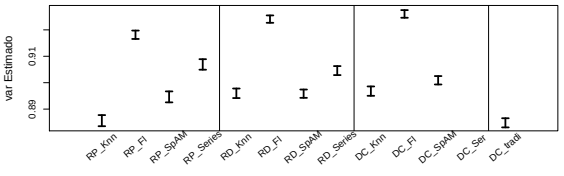
Figura 21 – Risco estimado para os nove estimadores nas funções: $g_6 f(z) = \Sigma(0.19, z_s)$, $g_6 g(z) = \Sigma(0.22, z_s)$, $g_6 h(z) = \Sigma(0.24, z_s)$, $g_6 i(z) = \Sigma(0.27, z_s)$, $g_6 j(z) = \Sigma(0.30, z_s)$, respectivamente.

Nota-se a partir das Figuras 19, 20 e 21 que exceto no caso da função de $g_1(z) = z^2$, em que os métodos se comportaram de forma semelhante, todos os demais dão destaque para regressão plug-in, direta e densidade condicional via florestas aleatórias, apresentando os menores riscos quando comparadas as demais metodologias. De forma geral, a regressão plug-in, quando se trata de KNN, apresentou os maiores riscos e, dessa forma, os piores resultados. É importante ressaltar que o método utilizado pelos astrônomos, que aqui denominamos densidade condicional tradicional, não apresentou resultados satisfatórios, com riscos elevados quando comparado as outras técnicas.

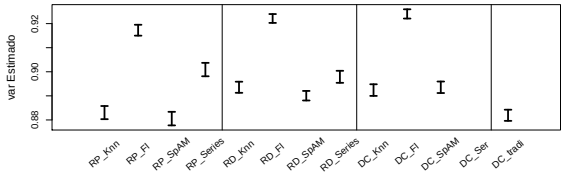
As Figuras 22 e 23 apontam os intervalos de confiança para as variâncias estimadas, com um nível de confiança de 95%.



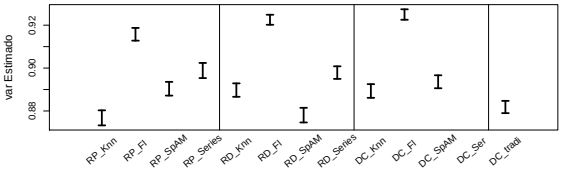
(a) Variância da função $g_6a(z) = \Sigma(0.05, z_s)$



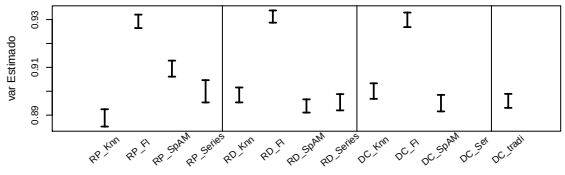
(b) Variância da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) Variância da função $g_6c(z) = \Sigma(0.11, z_s)$

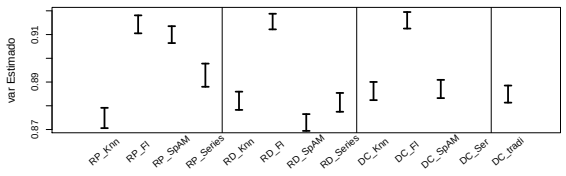


(d) Variância da função $g_6d(z) = \Sigma(0.13, z_s)$

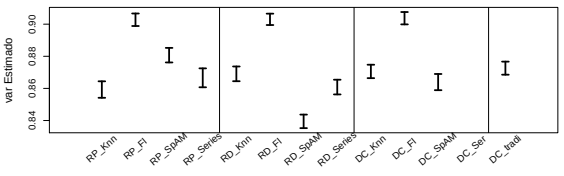


(e) Variância da função $g_6e(z) = \Sigma(0.16, z_s)$

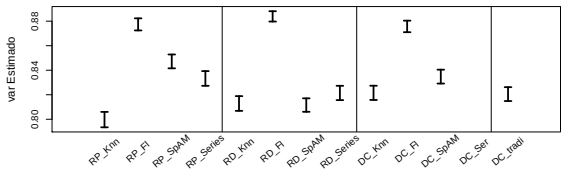
Figura 22 – Variância estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



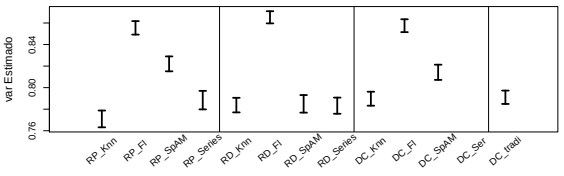
(a) Variância da função $g_6f(z) = \Sigma(0.19, z_s)$



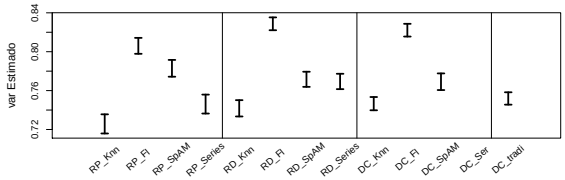
(b) Variância da função $g_6g(z) = \Sigma(0.22, z_s)$



(c) Variância da função $g_6h(z) = \Sigma(0.24, z_s)$



(d) Variância da função $g_6i(z) = \Sigma(0.27, z_s)$

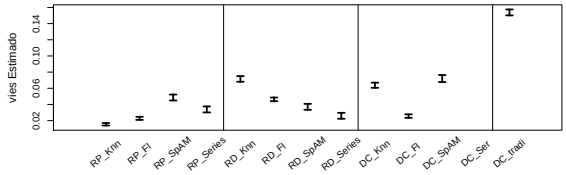


(e) Variância da função $g_6j(z) = \Sigma(0.30, z_s)$

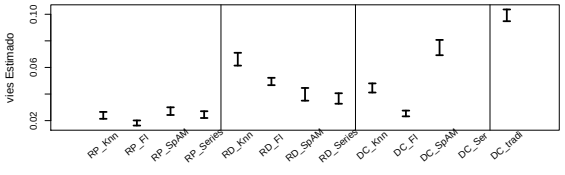
Figura 23 – Variância estimada para os nove estimadores nas funções: $g_6f(z) = \Sigma(0.19, z_s)$, $g_6g(z) = \Sigma(0.22, z_s)$, $g_6h(z) = \Sigma(0.24, z_s)$, $g_6i(z) = \Sigma(0.27, z_s)$, $g_6j(z) = \Sigma(0.30, z_s)$, respectivamente.

Todos os métodos baseados em florestas aleatórias obtiveram uma boa performance para a variância (valores próximos de um), bem como os resultados apresentados no risco estimado. Nesse mesmo sentido, a regressão plug-in, utilizando KNN, apresentou pequenos valores de variância, apontando uma não adequação do método ao problema. Na maior parte dos casos, o método de densidade condicional tradicional não obteve um bom comportamento quando comparado aos demais.

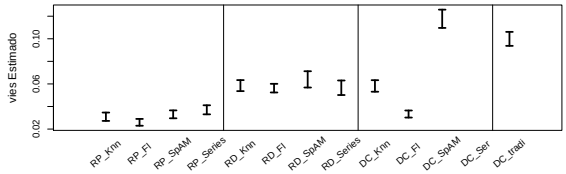
As Figuras 24 e 25 apresentam o viés para cada método e seus respectivos intervalos de confiança de 95% de confiança.



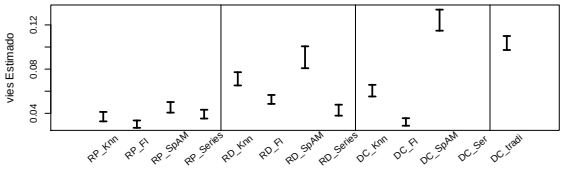
(a) Viés da função $g_6a(z) = \Sigma(0.05, z_s)$



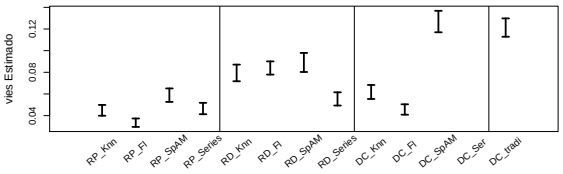
(b) Viés da função $g_6b(z) = \Sigma(0.08, z_s)$



(c) Viés da função $g_6c(z) = \Sigma(0.11, z_s)$

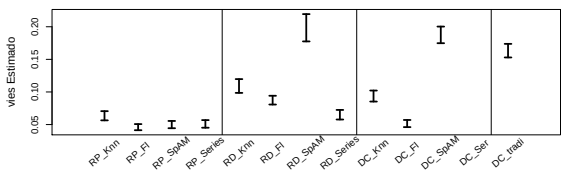


(d) Viés da função $g_6d(z) = \Sigma(0.13, z_s)$

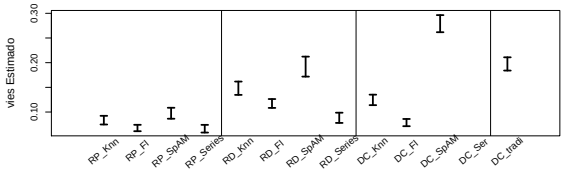


(e) Viés da função $g_6e(z) = \Sigma(0.16, z_s)$

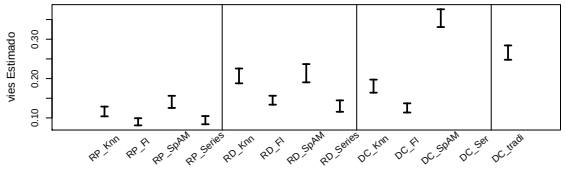
Figura 24 – Viés estimado para os nove estimadores nas funções: $g_6a(z) = \Sigma(0.05, z_s)$, $g_6b(z) = \Sigma(0.08, z_s)$, $g_6c(z) = \Sigma(0.11, z_s)$, $g_6d(z) = \Sigma(0.13, z_s)$, $g_6e(z) = \Sigma(0.16, z_s)$, respectivamente.



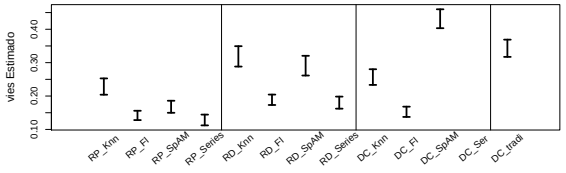
(a) Viés da função $g_6 f(z) = \Sigma(0.19, z_s)$



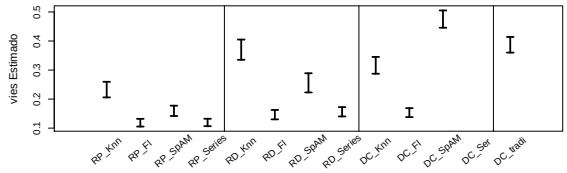
(b) Viés da função $g_6 g(z) = \Sigma(0.22, z_s)$



(c) Viés da função $g_6 h(z) = \Sigma(0.24, z_s)$



(d) Viés da função $g_6 i(z) = \Sigma(0.27, z_s)$



(e) Viés da função $g_6 j(z) = \Sigma(0.30, z_s)$

Figura 25 – Viés estimado para as nove estimadores nas funções: $g_6 f(z) = \Sigma(0.19, z_s)$, $g_6 g(z) = \Sigma(0.22, z_s)$, $g_6 h(z) = \Sigma(0.24, z_s)$, $g_6 i(z) = \Sigma(0.27, z_s)$, $g_6 j(z) = \Sigma(0.30, z_s)$, respectivamente.

Como descrito, a medida de viés é considerada satisfatória quando seus valores são próximos de zero. Nesse sentido, em geral, os métodos de densidade condicional tradicional e via SpAM apresentaram uma pior performance em comparação as outras metodologias. Em todos os casos, a regressão plug-in apresentou os melhores resultados, com menores valores de viés. De modo geral, o método de densidade condicional via florestas apresentou resultados satisfatórios. Em alguns casos, regressão direta utilizando séries também levou a bons resultados.

É importante lembrar que neste problema, o método de densidade condicional tradicional também não obteve nenhuma vantagem quando comparado com os demais. Isso acontece segundo as três medidas de performance.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS E PROPOSTAS FUTURAS

Referente ao primeiro conjunto de dados (DEEP2 EGS Region), os métodos de regressão direta, plug-in e densidade condicional quando utilizamos florestas aleatórias para estimar as funções de regressão, em geral, levaram a melhores resultados. A regressão plug-in via KNN levou a uma pior performance, apresentando os maiores riscos na maioria dos casos. A densidade condicional e regressão direta, quando utilizamos Séries Espectrais, SpAM e KNN apresentaram riscos elevados. Além disso observamos uma certa instabilidade no método de densidade condicional via Series Espectrais, com desvio padrão elevado comparados as demais metodologias. A densidade condicional tradicional também não apresentou nenhuma vantagem em comparação aos outros métodos. Isso também acontece quando observamos a medida de variância, em que todos os métodos via florestas aleatórias, obtiveram uma melhor performance comparado aos demais, com valores próximos 1. Neste caso, a regressão plug-in, em quase todas as funções, apresentou a pior performance apresentando variância perto de 0. Em termos do viés, todas as opções de regressão plug-in se destacaram quando comparamos aos outros métodos, apresentando menores valores para tal medida. Regressão direta via SpAM e Séries Espectrais e densidade condicional utilizando florestas aleatórias e Séries Espectrais obtiveram bons resultados.

De uma forma geral, no caso do conjunto de dados Sheldon 2012, os resultados se assemelham aos dos DEEP2 EGS Region. Neste caso, também damos destaque aos métodos de regressão direta, plug-in e densidade condicional via florestas aleatórias. Todas as metodologias utilizando KNN apresentaram maiores riscos e, consequentemente, uma performance ruim. Densidade condicional tradicional também apresentou riscos elevados. Para as medidas de variância, a regressão plug-in via KNN, na maioria das funções, apresentou valores próximos de 0, evidenciando uma não adequação para este problema. Quando falamos em viés, destacamos a regressão plug-in em todos os casos, bem como no primeiro conjunto de dados. Também

apresentaram valores baixos para o viés os métodos de regressão direta utilizando séries espectrais e densidade condicional via florestas aleatórias. Ainda para este conjunto, o método de densidade condicional baseado em series espectrais, apresentou valores discrepantes em relação aos demais métodos e em virtude disso, optamos por não utilizar os resultados obtidos.

Na maioria das funções testadas neste trabalho, os dois conjuntos de dados apresentaram um destaque para os métodos de regressão direta, plug-in e densidade condicional via florestas aleatórias, levando em conta as três medidas: risco, variância e viés.

A metodologia mais utilizada na astronomia (densidade condicional tradicional), de um modo geral, não obteve uma boa performance, evidenciando que existem métodos melhores para a obtenção estimativas de funções do resdhift.

O estimador de regressão direta precisa ser ajustado cada vez que utilizamos uma nova função g de interesse, enquanto que o estimador de densidade condicional possui a vantagem de só precisar ser calculado uma única vez e, então, pode ser utilizado para todas as funções g's de interesse.

REFERÊNCIAS

BALL, N.; BRUNNER, R. J. Data mining and machine learning in astronomy. **International Journal of Modern Physics D**, v. 19, p. 1049–1106, 2010. Citado na página 15.

BENEDETTI, J. K. On the nonparametric estimation of regression functions. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 248–253, 1977. Citado na página 21.

CUNHA, C.; LIMA, M.; OYAIZU, H.; FRIEMAN, J.; LIN, H. Estimating the redshift distribution of photometric galaxy samples — II. Applications and tests of a new method. **Monthly Notices of the Royal Astronomical Society**, n. 396, p. 2379–2398, 2009. Citado na página 15.

FREEMAN, P. E.; IZBICKI, R.; LEE, A. B. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 468, n. 4, p. 4556–4565, 2017. Citado na página 43.

IZBICKI, R.; LEE, A.; FREEMAN, P. Photometric redshift prediction under selection bias, submitted. 2014. Citado na página 43.

IZBICKI, R.; LEE, A. B. Nonparametric conditional density estimation in a high-dimensional regression setting. **Journal of Computational and Graphical Statistics**, Taylor & Francis, n. just-accepted, 2016. Citado na página 15.

IZBICKI, R.; LEE, A. B. Converting high-dimensional regression to high-dimensional conditional density estimation. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 11, n. 2, p. 2800–2831, 2017. Disponível em: <http://dx.doi.org/10.1214/17-EJS1302>. Citado na página 25.

IZBICKI, R.; LEE, A. B.; FREEMAN, P. E. Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. **Annals of Applied Statistics**, 2017. Citado nas páginas 15 e 25.

IZBICKI, R.; SANTOS, T. M. d. **Machine Learning sob a ótica estatística**. [S.l.]: Universidade Fedaral de São Carlos, 2017. Citado na página 21.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 6. Citado nas páginas 21 e 23.

KIND, M. C.; BRUNNER, R. J. Tpz: photometric redshift pdfs and ancillary information by using prediction trees and random forests. **Monthly Notices of the Royal Astronomical Society**, v. 432, n. 2, p. 1483–1501, 2013. Citado na página 15.

LEE, A.; IZBICKI, R. A spectral series approach to high-dimensional nonparametric regression, submitted. 2014. Citado na página 24.

LEE, A. B.; IZBICKI, R. A spectral series approach to high-dimensional nonparametric regression. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 10, n. 1, p. 423–463, 2016. Citado nas páginas 21 e 24.

MANDELBAUM, R.; SELJAK, U.; HIRATA, C.; BARDELLI, S.; BOLZONELLA, M.; BONGIORNO, A.; CAROLLO, M.; CONTINI, T.; CUNHA, C.; GARILLI, B. *et al.* Precision photometric redshift calibration for galaxy–galaxy weak lensing. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 386, n. 2, p. 781–806, 2008. Citado nas páginas 16, 30, 35 e 37.

RAU, M. M.; SEITZ, S.; BRIMIOULLE, F.; FRANK, E.; FRIEDRICH, O.; GRUEN, D.; HOYLE, B. Accurate photometric redshift probability density estimation-method comparison and application. **arXiv preprint arXiv:1503.08215**, 2015. Citado nas páginas 15 e 16.

RAVIKUMAR, P.; LAFFERTY, J.; LIU, H.; WASSERMAN, L. Sparse additive models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 71, n. 5, p. 1009–1030, 2009. Citado nas páginas 21 e 24.

SHELDON, E.; CUNHA, C.; MANDELBAUM, R.; BRINKMANN, J.; WEAVER, B. Photometric redshift probability distributions for galaxies in the SDSS DR8. **The Astrophysical Journal Supplement Series**, v. 201, n. 2, 2012. Citado nas páginas 15 e 43.

WEINER, B. J.; PHILLIPS, A. C.; FABER, S.; WILLMER, C. N.; VOGT, N. P.; SIMARD, L.; GEBHARDT, K.; IM, M.; KOO, D.; SARAJEDINI, V. L. *et al.* The deep groth strip galaxy redshift survey. iii. redshift catalog and properties of galaxies. **The Astrophysical Journal**, IOP Publishing, v. 620, n. 2, p. 595, 2005. Citado na página 37.