Universidade Estadual de Londrina

JOÃO DEBASTIANI NETO

ESTIMAÇÃO DOS REDSHIFT DE GALÁXIAS UTILIZANDO DADOS DE FOTOMETRIA:

UMA ABORDAGEM GAMLSS

Londrina
2023

JOÃO DEBASTIANI NETO

ESTIMAÇÃO DOS REDSHIFT DE GALÁXIAS UTILIZANDO DADOS DE FOTOMETRIA:

UMA ABORDAGEM GAMLSS

Dissertação de mestrado apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

Orientador: Prof. Dr. Rodrigo Rossetto Pescim

Londrina
2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Debastiani Neto, Joao .
Estimação dos redshift de galáxias utilizando dados de fotometria: uma abordagem GAMLSS / Joao Debastiani Neto. - Londrina, 2023.
101 f.

Orientador: Rodrigo Rossetto Pescim.
Dissertação (Mestrado em Matemática Aplicada e Computacional) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Matemática Aplicada e Computacional, 2023.
Inclui bibliografia.

1. Modelos Aditivos Generalizados para Locação, Escala e Forma - Tese. 2. Modelos de Regressão - Tese. 3. Modelagem Estatística - Tese. 4. Análise de Componentes Principais - Tese. I. Rossetto Pescim, Rodrigo . II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Matemática Aplicada e Computacional. III. Título.

CDU 51

JOÃO DEBASTIANI NETO

ESTIMAÇÃO DOS REDSHIFT DE GALÁXIAS UTILIZANDO DADOS DE FOTOMETRIA:

UMA ABORDAGEM GAMLSS

Dissertação de mestrado apresentada ao Departamento de Matemática da Universidade Estadual de Londrina, como requisito parcial para a obtenção do Título de MESTRE em Matemática Aplicada e Computacional.

BANCA EXAMINADORA

Prof. Dr. Rodrigo Rossetto Pescim
Universidade Estadual de Londrina – UEL

Prof. Dra. Mariana Ragassi Urbano
Universidade Estadual de Londrina – UEL

Prof. Dr. Luiz Ricardo Nakamura
Universidade Federal de Lavras – UFLA

Londrina, 24 de fevereiro de 2023.

AGRADECIMENTOS

Em princípio, devo a Deus o meu agradecimento maior, pela minha vida, pela minha saúde e pela sua proteção em todos as etapas de minha trajetória.

Aos meus pais João e Maria e irmãs Gisele e Jesiane pelo carinho, apoio, e compreensão em momentos de dificuldades durante a elaboração desta dissertação.

Ao meu orientador Professor Doutor Rodrigo Rossetto Pescim que, mais do que orientar um trabalho, acreditou em meu potencial proporcionando no decorrer desta dissertação, muitos momentos de aprendizado.

Aos professores da banca examinadora de defesa de Mestrado, pela disponibilidade para a leitura deste estudo e contribuir com sugestões, enriquecendo grandemente esta pesquisa.

Ao Programa de Pós-graduação em Matemática Aplicada e Computacional da UEL como um todo, por ter me oportunizado esses dois anos de crescimento acadêmico, profissional e pessoal.

Aos amigos Altair Santos de Oliveira Tosti, Carolina Lupifierio de Queiroz, Felinto Junior da costa e Maria Graziela da Silva Fernandes que proporcionaram momentos de troca de saberes, diálogos, aprendizagem e a conquista de grandes amizades.

Pelo muito que lhes devo, obrigado.

DEBASTIANI, João Neto. **Estimação dos redshift de galáxias utilizando dados de fotometria:** uma abordagem GAMLSS. 2023. 101 f. Dissertação (Mestrado em Matemática Aplicada e Computacional) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

Cosmologia é um ramo da astronomia que busca por interpretar as origens do Universo, bem como investigar os objetos nela presentes. Compreender como os elementos celestes interagem e quais os fatores que influenciam para tal aspecto, são alguns dos anseios dos pesquisadores que se debruçam nestas questões. Desde o início do século XX, pesquisadores referem-se a expansão constante do Universo, de maneira que galáxias e estrelas estão, em geral, se afastando da Terra a uma certa velocidade. Astrônomos e pesquisadores desta área são capazes de identificar tal afastamento (ou aproximação), por meio de uma medida denominada redshift, que refere-se ao deslocamento da luz originária destes objetos celestes para o infravermelho baseado em seu comprimento de onda. Algumas técnicas possibilitam obter uma estimativa destes redshift, dentre os quais se destacam o redshift espectroscópico e o redshift fotométrico. Embora a primeira destas técnicas seja mais apurada no que se refere aos valores estimados, o segundo método propõe uma diminuição de tempo e de recursos, sendo assim, a mais considerada. Diversas alternativas na estimação de redshift fotométricos se mostraram extremamente eficazes e altamente utilizadas, dentre os quais se destacam modelos estatísticos vinculadas a técnicas de Machine Learning e Decision Tree. Buscando-se apresentar uma nova alternativa para tal problemática, foi proposta a presente pesquisa, cujo objetivo consiste na implementação de um Modelo Aditivo Generalizado para Locação, Escala e Forma (GAMLSS) visando a estimativa de desvios para o vermelho fotométricos de galáxias, segundo a fotometria de diferentes comprimentos de ondas (bandas). Entende-se que devido a natureza mais robusta e flexível dos GAMLSS, pode-se obter resultados mais satisfatórios do que os encontrados na literatura para os Modelos Lineares Generalizados (GLM), bem como uma alternativa viável para pesquisas fundamentadas em redes neurais e decision tree. Para tanto, considerou-se para a análise e interpretação dos dados o software R, de maneira que o conjunto de dados utilizado foi proveniente do pacote CosmoPhotoz, em particular, o conjunto denominado PHoto-z Accuracy Testing (PHAT0). Devido a elevada quantidade de observações contidas nesta base de dados (169520 dados), foi estabelecido, para análise desta pesquisa, um total de 8476 observações (5% da base PHAT0), sendo composta 12 variáveis (redshift fotométrico e 11 magnitudes de filtros). Por meio da análise realizada, observou-se que as variáveis explicativas são altamente correlacionadas, sendo necessário a utilização da técnica de análise de componentes principais (PCA). O modelo GAMLSS ajustado que apresentou melhores resultados contou com suavizadores (thin plate spline $s(\cdot)$) para os quatro parâmetros da distribuição Box-Cox t (BCTo). Em síntese, a classe de modelos GAMLSS é uma alternativa eficaz para estimação de redshift fotométrico, apresentando-se como uma opção interessante para modelagem de dados desta natureza.

Palavras-chave: análise de componentes principais; astronomia; GAMLSS; redshift fotométrico; suavizadores.

DEBASTIANI, João Neto. **Estimation of redshift of galaxies using photometry data:** a GAMLSS approach. 2023. 101 p. Dissertation (Master in Applied and Computational Mathematics) – State University of Londrina, Londrina, 2023.

ABSTRACT

Cosmology is a branch of astronomy that seeks to interpret the origins of the Universe, as well as investigate the objects present on it. Understanding how the celestial elements interact and which factors influence this aspect are some of the concerns of researchers who address on these issues. Since the beginning of the 20th century, researches refer to the constant expansion of the Universe, in a way that galaxies and stars are, in general, moving away from the Earth at a certain speed. Astronomers and researchers in this area are able to identify such distance (or approximation) by means of a measure called redshift, which refers to the displacement of the light originating from these celestial objects towards the infrared based on its wavelength. In some techniques it is possible to obtain an estimate of these redshift, among which the spectroscopic and photometric redshift stand out. Although the first of these techniques is more accurate with regard to estimated values, the second method proposes a decrease in time and resources, and is therefore the most considered. Several alternatives for estimating photometric redshift proved to be extremely effective and highly used, among which stand out statistical models linked to Machine Learning and Decision Tree techniques. Seeking to present a new alternative to that problem, this research was proposed, which aims to implement a Generalized Additive Model for Location, Shape and Scale (GAMLSS) to estimate the photometric redshift of galaxies, according to the photometry of different wavelengths (bands). It is understood that due to the more robust and flexible nature of GAMLSS, more satisfactory results can be obtained than those models in the literature such as the Generalized Linear Models (GLM), as well as a viable alternative for research based on neural networks and decision tree. Therefore, the R software was used for data analysis and interpretation, so the dataset used was provided by the CosmoPhotoz package, in particular, the set called PHoto-z Accuracy Testing (PHAT0). Due to the high number of observations contained in that database (169520 data), a total of 8476 observations were established for the analysis of this research (5% of the PHAT0 base), comprising 12 variables (photometric redshift and 11 filter magnitudes). Through the analysis carried out, it was observed that the explanatory variables are highly correlated, requiring the use of the principal components analysis technique (PCA). The fitted model had smoothers (thinplate spline $s(\cdot)$) for the four parameters of the Box-Cox t distribution (BCTo). In summary, the GAMLSS class of models is an effective alternative for estimating and predicting photometric redshift, presenting itself as an interesting option for modeling data of this nature.

Key words: principal component analysis; astronomy; GAMLSS; photometric redshift; smoothing.

LISTA DE FIGURAS

Figura 3.1 - Representação dos três ciclos dos algoritmos GAMLSS 29
Figura 3.2 - Variações na curva da densidade da distribuição Box-Cox t 35
Figura 3.3 - Estratégia A - Método de seleção de covariáveis Stepwise 38
Figura 3.4 - Worm plot de um modelo GAMLSS com distribuição BCT 42
Figura 3.5 - Padrões sistemáticos dos resíduos quantílicos em um worm plot..... 43
Figura 3.6 - Gráfico indicando o coeficiente de correlação entre as variáveis 49
Figura 4.1 - Contribuição de cada PC na variância do sistema 54
Figura 4.2 - Contribuição dos 11 filtros nas duas PC´s 56
Figura 4.3 - Histograma e Box plot da variável redshift fotométrico 57
Figura 4.4 - Histograma e Box plot da PC1 59
Figura 4.5 - Histograma e Box plot da PC2 59
Figura 4.6 - Gráficos de dispersão das componentes principais versus redshift..... 62
Figura 4.7 - Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição BCTo 67
Figura 4.8 - Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição BCT 67
Figura 4.9 - Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição GB2 68
Figura 4.10 - Half Normal plot dos resíduos do modelo GAMLSS BCTo 70
Figura 4.11 - Half Normal plot dos resíduos do modelo GAMLSS BCT 70
Figura 4.12 - Half Normal plot dos resíduos do modelo GAMLSS GB2..... 71
Figura 4.13 - Worm plot do modelo GAMLSS BCTo 72
Figura 4.14 - Worm plot do modelo GAMLSS BCT 72
Figura 4.15 - Worm plot do modelo GAMLSS GB2..... 73
Figura 4.16 - Comportamento das funções de suavização para o parâmetro μ 74
Figura 4.17 - Comportamento da função de suavização para o parâmetro σ..... 76
Figura 4.18 - Comportamento da função de suavização para o parâmetro ν..... 77
Figura 4.19 - Comportamento da função de suavização para o parâmetro τ 78
Figura A.1 - Gráficos dos resíduos do modelo ajustado para BCTo 86
Figura A.2 - Worm plot do modelo ajustado para a distribuição BCTo..... 86
Figura A.3 - Half Normal Plot dos resíduos do modelo 87
Figura A.4 - Gráficos dos resíduos do modelo ajustado para BCT 88

Figura A.5 - Worm plot do modelo ajustado para a distribuição BCT 88
Figura A.6 - Half Normal Plot dos resíduos do modelo 89
Figura A.7 - Gráficos dos resíduos do modelo ajustado para GB2 90
Figura A.8 - Worm plot do modelo ajustado para a distribuição GB2 90
Figura A.9 - Half Normal Plot dos resíduos do modelo 91
Figura B.1 - Gráficos dos resíduos do modelo LM ajustado 92
Figura B.2 - Worm plot do modelo LM ajustado 92
Figura B.3 - Half Normal Plot dos resíduos do modelo LM 93
Figura B.4 - Gráficos dos resíduos do modelo GLM GAMA 94
Figura B.5 - Worm plot do modelo GLM GAMA 94
Figura B.6 - Half Normal Plot dos resíduos do modelo GLM GAMA 95
Figura B.7 - Gráficos dos resíduos do modelo GLM NORMAL INVERSA 96
Figura B.8 - Worm plot do modelo GLM NORMAL INVERSA 96
Figura B.9 - Half Normal Plot dos resíduos do modelo GLM NORMAL INVERSA 97
Figura B.10 - Gráficos dos resíduos do modelo GAM GAMA 98
Figura B.11 - Worm plot do modelo GAM GAMA 98
Figura B.12 - Half Normal Plot dos resíduos do modelo GAM GAMA 99
Figura B.13 - Gráficos dos resíduos do modelo GAM NORMAL INVERSA 100
Figura B.14 - Worm plot do modelo GAM NORMAL INVERSA 100
Figura B.15 - Half Normal Plot dos resíduos do modelo GAM NORMAL INVERSA 101

LISTA DE TABELAS

Tabela 3.1 - Famílias de distribuições contínuas implementadas no software R 31
Tabela 3.2 - Famílias de distribuições discretas implementadas no software R 31
Tabela 3.3 - Parâmetros distribucionais de BCTo 34
Tabela 3.5 - Diferentes formatos do gráfico worm plot e suas interpretações 44
Tabela 4.3 - Resumo das medidas do redshift fotométrico 58
Tabela 4.4 - Resumo das medidas das duas componentes principais 58
Tabela 4.5 - Distribuições apresentadas no comando fitDist(·) 61
Tabela 4.6 - Distribuições apresentadas por meio da função chooseDist(·) 61
Tabela 4.7 - Resumo das medidas dos modelos ajustados 65
Tabela 4.8 - Resumo das medidas dos modelos GAMLSS ajustados 66
Tabela 4.9 - Resumo das medidas dos resíduos quantílicos dos modelos GAMLSS ajustados 69
Tabela A.1 - Resumo das medidas dos resíduos quantílicos 87
Tabela A.2 - Resumo das medidas do modelo ajustado 87
Tabela A.3 - Resumo das medidas dos resíduos quantílicos 89
Tabela A.4 - Resumo das medidas do modelo ajustado 89
Tabela A.5 - Resumo das medidas dos resíduos quantílicos 91
Tabela A.6 - Resumo das medidas do modelo ajustado 91
Tabela B.1 - Resumo das medidas dos resíduos quantílicos 93
Tabela B.2 - Resumo das medidas do modelo ajustado 93
Tabela B.3 - Resumo das medidas dos resíduos quantílicos 95
Tabela B.4 - Resumo das medidas do modelo ajustado 95
Tabela B.5 - Resumo das medidas dos resíduos quantílicos 97
Tabela B.6 - Resumo das medidas do modelo ajustado 97
Tabela B.7 - Resumo das medidas dos resíduos quantílicos 99
Tabela B.8 - Resumo das medidas do modelo ajustado 99
Tabela B.9 - Resumo das medidas dos resíduos quantílicos 101
Tabela B.10 - Resumo das medidas do modelo ajustado 101

LISTA DE QUADROS

Quadro 3.4 - Resumo das propriedades matemáticas e funções da distribuição BCTo 36
Quadro 3.6 - Banco de dados - pacote CosmoPhotoz 45
Quadro 3.7 - Descrição das variáveis contidas no dados do pacote CosmoPhotoz.... 46
Quadro 3.8 - Medidas resumo das 12 variáveis do conjunto de dados PHAT0 46
Quadro 3.9 - Medidas resumo das 12 variáveis da amostra aleatória 47
Quadro 3.10 - Coeficiente de correlação das covariáveis 49
Quadro 4.1 - Proporção de Variância explicada pelos componentes principais 54
Quadro 4.2 - Contribuição dos 11 filtros nas duas componentes principais 55

SUMÁRIO

1 INTRODUÇÃO 14

2 REVISÃO DE LITERATURA 17

2.1 REDSHIFT FOTOMÉTRICO 17

2.2 CONTEXTUALIZAÇÃO DOS MODELOS DE REGRESSÃO 18

3 MATERIAIS E MÉTODOS 25

3.1 MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA 25

3.1.1 Definição de um GAMLSS 25

3.1.2 Estimação do Modelo 27

3.1.3 Família de Distribuições nos Modelos GAMLSS 30

3.1.4 Seleção de Modelos 30

3.1.5 Análise Diagnóstica de um Modelo Ajustado 40

3.2 RECURSOS COMPUTACIONAIS 42

3.3 DESCRIÇÃO DO CONJUNTO DE DADOS 44

3.4 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA) 50

4 RESULTADOS E DISCUSSÃO 53

4.1 COVARIÁVEIS - COMPONENTES PRINCIPAIS OBTIDAS (PC) 53

4.2 ANÁLISE DESCRITIVA DOS DADOS 57

4.3 SELEÇÃO DO MODELO 60

4.3.1 Distribuição da Variável Resposta e Funções de Ligação 60

4.3.2 Termos Utilizados nos Preditores dos Parâmetros da Distribuição 61

4.4 MODELO PROPOSTO E SUAS ESPECIFICAÇÕES 64

4.4.1 Análise de Resíduos e Diagnóstico dos Modelos Propostos 66

4.5 EFEITO DAS FUNÇÕES DE SUAVIZAÇÃO NOS PREDITORES DOS PARÂMETROS DISTRIBUCIONAIS 73

4.5.1 Parâmetro μ (Mediana) 74

4.5.2 Parâmetro σ (Coeficiente de Variação) 75

4.5.3 Parâmetro ν (Coeficiente da Assimetria) 76

4.5.4 Parâmetro τ (Coeficiente da Curtose) 77

5 CONSIDERAÇÕES FINAIS 79

REFERÊNCIAS 80

APÊNDICES 86

APÊNDICE A: RESULTADOS OBTIDOS COM OS PREDITORES INDICADOS PELAS FUNÇÕES ADDTERM(·) E STEPGAICALL.A(·) 86

APÊNDICE B: RESULTADOS OBTIDOS COM OS AJUSTES DOS 5 MODELOS DE REGRESSÃO: 1 MODELO LM, 2 MODELOS GLM E 2 MODELOS GAM 92

1 INTRODUÇÃO

Nas últimas décadas, é possível observar a utilização de métodos estatísticos em diversas áreas da ciência, que de maneira geral, busca-se por melhores análises e interpretações de características dos fenômenos considerados. Impulsionados com o desenvolvimento computacional, a estatística sofreu forte desenvolvimento, principalmente no que se refere ao seu carácter empírico/experimental, uma vez que, de maneira geral, com *software* computacionais, cálculos extremamente complexos foram realizados de maneira simples [16].

Neste sentido, pode-se compreender a estatística como uma ciência multidisciplinar, em que diversos profissionais das inúmeras áreas de estudo podem fundamentar-se na análise estatística de dados [18]. Exemplificando algumas das aplicações da estatística em outras áreas da ciência, tem-se a análise de desempenho na área dos esportes; otimização de recursos econômicos; previsões e predições de informações visando diminuição de custos; exploração de *big data* objetivando-se a construção de modelos estatísticos, entre outras.

Uma questão bastante importante indicada em [14], refere-se ao fato de que a estatística não se limita a análises e interpretações de dados experimentais, como por exemplo o lançamento de um dado ou a coleta de uma determinada amostra. Em diversas situações, os estudos considerados não são experimentais, mas sim estudos observacionais, tais como informações provenientes da astronomia ou cosmologia. Nesta área, tem-se uma grande dificuldade em repetir experimentos, sendo que a quantidade de dados, embora extremamente grande, são limitados.

Para tanto, ainda que se tenha essa limitação imposta devido a natureza dos dados, por meio de modelos de previsão e predição estatísticos, um vasto universo de informações podem ser analisadas. Em particular, uma área com amplo desenvolvimento de pesquisas refere-se a cosmologia, em particular estudos direcionados a aproximação ou afastamento de objetos no Universo [45].

Em geral, em [39], tem-se que as galáxias que podem ser observadas, estão se afastando da Via Láctea. Desta maneira, com este afastamento, a luz originária destas galáxias são deslocadas para o infravermelho baseado em seu comprimento de onda. A medida que considera este deslocamento para o vermelho é denominada de *redshift*, e é denotada por *photoz*. Tendo em vista que o valor do *redshift* relaciona-se com o distanciamento das galáxias, compreende-se que o *photoz* de uma determinada galáxia pode ser entendido como uma medida de distância, possibilitando a determinação de suas características.

Diversos trabalhos buscam elucidar investigações envolvendo o conceito de *redshift*, além de como a estatística apresenta alternativas bastante satisfatórias para a problemática considerada. O trabalho de [1] pesquisou distinções entre cinco modelos fundamentados em estruturas de *Machine Learning* com redes neurais, visando a estimação de *redshift* cosmológico. Outros trabalhos como em [5]; [10] e [71] apresentaram técnicas envolvendo redes

neurais para estimar o *redshift* fotométrico de uma amostra de galáxias.

Nesta mesma perspectiva, pesquisas como as de [4]; [28]; [13]; [24] também buscaram por apresentar modelos para estimação de *redshift*, considerando uma estrutura fundamentada no aprendizado de máquinas aliada às redes neurais, utilizando técnicas como o *k-nearest neighbour* (kNN) e dos algoritmos de árvores de regressão e decisão para a modelagem dos dados.

Embora exista esta diversificação de técnicas que buscam por implementar modelos de estimação de *redshift* de galáxias, questões como o custo de processamento computacional, bem como a complexidade de estruturas implementadas, se tornam entraves que devem ser otimizados [20].

Considerando estes aspectos, a pesquisa de [20] apresenta uma alternativa mais rápida e bastante eficaz para a estimação de *redshift*. Estes autores, buscando possibilidades distintas ao aprendizado de máquinas, propõem um modelo linear generalizado (GLM), considerando a distribuição gama na estimativa dos desvios para o vermelho fotométricos de galáxias, a partir da fotometria de vários comprimentos de onda. Esta foi uma ideia bastante interessante para problemas desta natureza, uma vez que os GLM's possibilitam a interpretabilidade de seus resultados, bem como sua aplicação em uma diversidade de conjuntos de dados de astronomia [20].

No que tange a possíveis vantagens dos GLM's para a estimação dos *redshift*, segundo [21], tem-se a melhoria no tempo necessário para o ajuste e treinamento do modelo, já que o tamanho do conjunto de treinamento pode influenciar nos métodos empíricos. Contudo, [56] ressalva que os GLM's possuem um certo engessamento devido à algumas restrições que são provenientes de sua estrutura, dentre as quais destacam-se: $[i]$ a distribuição de probabilidade considerada para a variável resposta, que deve pertencer à família exponencial, de maneira que a média é modelada em função das variáveis explicativas e de sua variância; $[ii]$ para as distribuições da família exponencial, a curtose e a assimetria da variável resposta são escritas em função da média e do parâmetro de dispersão. Isto significa que a variância, a curtose e a assimetria são modelados em função da média.

Buscando por uma maior flexibilidade nesta estrutura, [56], apresentaram uma classe de modelos que relaxam as condições observadas nos GLM's, permitindo assim, explicar não somente a média da variável reposta em função das variáveis explicativas, mas também a descrição de sua variabilidade, sua assimetria e curtose em função das covariáveis. Esta classe é denominada de Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS).

Pode-se verificar em [63] que nos GAMLSS, ao invés de uma única estrutura de regressão, são obtidas diversas estruturas de regressão, sendo que cada uma delas possui um conjunto de variáveis explicativas individuais. Além disso, uma grande flexibilização existente se refere a distribuição assumida para a variável resposta, já que não necessariamente pertence à família exponencial de distribuições.

Considerando a estrutura observada nos GAMLSS, que segundo [63] possi-

bilita diversas flexibilidades frente aos GLM's, além de ser uma proposta alternativa em relação às redes neurais e algoritmos de *decision tree* (DTR), o objetivo deste trabalho consiste na implementação de um modelo GAMLSS visando a estimativa de desvios para o vermelho fotométricos de galáxias *(redshift)*, segundo a fotometria de diferentes comprimentos de ondas (bandas). Entende-se que devido a natureza mais robusta e flexível dos GAMLSS, podem-se obter resultados mais satisfatórios do que os encontrados na literatura para os GLM's, bem como uma alternativa bastante viável quando consideram-se pesquisas fundamentadas em redes neurais e DTR.

Para a descrição pormenorizada desta pesquisa, além deste primeiro capítulo referente à introdução do trabalho, tem-se que a mesma foi divida em mais 4 capítulos: No segundo capítulo é apresentada uma revisão da literatura, no qual inicia-se por apontamentos referentes ao conceito de *redshift* fotométrico, bem como procedimentos para sua estimação. Em seguida, realiza-se uma discussão sobre alguns modelos de regressão, dentre os quais, se destacam os Modelos de Regressão Clássicos, perpassando pelos Modelos Lineares Generalizados e, por fim, os Modelos Aditivos Generalizados (GAM).

O terceiro capítulo apresenta informações sobre a metodologia adotada para a presente pesquisa. Aqui, é realizada uma discussão mais aprofundada dos modelos GAMLSS, sendo descrita sua definição, os métodos para estimação, as famílias de distribuição, procedimentos para seleção de modelos e análise de resíduos e diagnóstico de um modelo ajustado. Ademais, exibe-se uma descrição sobre a técnica da análise de componentes principais, além do banco de dados considerado nesta pesquisa. O quarto capítulo expõe aspectos relativos sobre a preparação do conjunto de dados visando sua modelagem, os resultados obtidos com o modelo ajustado, bem como uma discussão sobre a análise de resíduos e diagnóstico do modelo proposto. Por fim, são apresentadas as considerações finais sobre a dissertação, buscando repostas para a problemática exposta, além de reflexões sobre sobre os resultados obtidos neste trabalho.

2 REVISÃO DE LITERATURA

Neste capítulo são descritas, de maneira introdutória, a conceitualização de *redshift* fotométrico, bem como procedimentos para sua estimação. Além disso, buscando realizar uma contextualização com os conceitos e as estruturas dos modelos de regressão, são apresentadas as definições e propriedades do Modelo Clássico de Regressão Linear (LM), dos GLM's e dos Modelos Aditivos Generalizados. Esta seção constitui-se como fundamento para a discussão sobre os GAMLSS, que é realizada no capítulo de materiais e métodos.

2.1 REDSHIFT FOTOMÉTRICO

Estudos relativos a astrofísica e a cosmologia indicam a existência de inúmeras galáxias distintas à Via Láctea, bem como a constante expansão do Universo [49]. Contudo, estas pesquisas não são recentes. Tais descobertas ocorreram entre os anos de 1923 e 1929, quando o astrônomo americano Edwin Powell Hubble (1889-1953) conseguiu visualizar e medir individualmente estrelas da galáxia de Andrômeda. Além disso, este pesquisador foi capaz de observar o deslocamento para o vermelho nas linhas de espectro de galáxias, medindo suas distâncias e concluindo que estas estavam se afastando da Terra a uma certa velocidade [50].

Considerando tal fato, a luz emitida por estes entes celestiais sofrem um deslocamento para o vermelho, acarretando no aumento do comprimento de onda emitido pela galáxia. Esse deslocamento para o vermelho é denominado de *redshift*. Segundo [23], caso ocorra a aproximação destes objetos, o comprimento de onda diminui, e neste caso diz-se que ocorre um *blueshift*, ou desvio para o azul. Assim sendo, tanto o *redshift* quanto o *blueshift* são referências de deslocamentos para as partes vermelhas (de comprimentos de onda mais longos) e azul (de comprimentos de onda mais curtos) do espectro [23].

Como observado em [50], a medida do *redshift* relaciona-se com a velocidade com que determinadas galáxias se afastam, bem como a distância que se encontram do planeta Terra. Neste sentido, há a possibilidade de que valor do *redshift* de uma galáxia, seja entendido como uma medida de distância, permitindo estabelecer algumas características da galáxia [39].

Em [49] infere-se que existem técnicas que possibilitam obter uma estimativa do *redshift* de uma galáxia. O método mais preciso, denominado de *redshift* espectroscópico (RE), consiste na observação de linhas do espectro, em que os comprimentos de onda medido são comparados com os comprimentos de onda das linhas em repouso. Outra técnica bastante usual, denominada de *redshift* fotométrico (*photoz*), fundamenta-se no uso da fotometria de um objeto em diferentes bandas, de maneira que pode ser compreendida como uma construção aproximada da Distribuição de Energia Espectral (SED).

Embora seja considerado um espectro de baixa resolução, apresentando grandes incertezas devido a escassez de dados precisos sobre a SED, tem-se em [49], que essa

técnica possibilita que sejam medidas inúmeras vezes mais distâncias do que as medidas mais precisas utilizando o RE. Além disso, em [39], afirma-se que a fotometria possibilita uma diminuição de tempo e de recursos, uma vez que podem ser obtidas estimativas de *redshift* de inúmeros objetos simultaneamente.

A grande desvantagem dos *photoz*, quando comparados com os RE, consiste em sua menor precisão na estimativa da média, bem como na existência dos erros catastróficos, que são valores estimados de *redshift* muito distintos dos valores obtidos espectroscopicamente. Em geral, tem-se que de 1% a 10%, dos valores de *redshift* fotométrico podem ser classificados como erros catastróficos [60].

2.2 CONTEXTUALIZAÇÃO DOS MODELOS DE REGRESSÃO

Entende-se por modelo, como uma representação aproximada de um determinado fenômeno [3]. Diversas áreas buscam por modelar situações do cotidiano, dentre as quais se destacam a matemática e a estatística. Embora estas ciências busquem descrever e analisar objetos similares, a maneira como estruturam esse processo é distinta. Em ambos os campos científicos, a organização consiste em estabelecer uma relação de dependência entre variáveis.

Contudo, enquanto na matemática um modelo apresenta seus componentes fixos, na estatística existe no mínimo um componente aleatório, inserindo assim, uma variável aleatória no modelo, relacionando-se com uma distribuição de probabilidade [8].

Uma classe em particular de modelos estatísticos é o denominado Modelo Clássico de Regressão Linear, que é definido em [55] por

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i, \tag{2.1}$$

com $i = 1, 2, \ldots, n$, em que Y_i são os valores observados atribuídos à variável resposta; $(x_{1i}, ..., x_{pi})$ representam as variáveis explicativas; p é o número de variáveis preditoras; β_a, com $a = 1, ..., p$, são os parâmetros do modelo, tal que β_0 é denominado de intercepto. Os ε_i são os erros aleatórios associados, de maneira que $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, isto é, são independentes entre si e identicamente distribuídos segundo uma distribuição normal com média zero e variância constante igual a σ^2.

Em [44] o modelo apresentado em (2.1), é reescrito de maneira matricial da seguinte maneira:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

em que $\mathbf{Y} = (y_1, y_2, ..., y_n)^T$ é o vetor cujos elementos são os valores da variável resposta Y, tal que $\boldsymbol{Y} \overset{iid}{\sim} N(\boldsymbol{\mu}; \boldsymbol{I}\sigma^2)$, com $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ e $\boldsymbol{I}$ denotando a matriz identidade de ordem n; $\mathbf{X} = (\mathbf{1}, \boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ é a matriz cujos elementos são os valores das variáveis explicativas, denominada de matriz de delineamento; $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ é o vetor de parâmetros a serem

estimados e $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ é o vetor de erros, em que $\boldsymbol{\varepsilon} \overset{iid}{\sim} N(\mathbf{0}; \boldsymbol{I}\sigma^2)$.

Ainda em [44], os autores mostram uma grande vantagem ao se trabalhar com expressões matriciais em modelos de regressão linear, isto é, referem-se ao fato de que o método de mínimos quadrados utilizado para a estimação do vetor de parâmetros $\boldsymbol{\beta}$ em (2.2), pode ser construído de maneira geral, e refere-se à minimização da forma quadrática

$$\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2. \tag{2.3}$$

Realizando-se cálculos matriciais, obtém-se a expressão para os estimadores de mínimos quadrados para $\boldsymbol{\beta}$, que é dada em [44] por

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{2.4}$$

Entretanto, pode-se observar que os modelos lineares clássicos apresentam uma estrutura bastante rígida quando se consideram seus pressupostos. Podem ser destacados alguns aspectos, tais como, a obrigatoriedade da variável resposta seguir distribuição normal, bem como não proporcionar maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear η. Estes pressupostos tornam os modelos lineares clássicos pouco satisfatórios quando implemantados em situações que envolvam dados binários, contagens ou proporções [15].

Buscando alternativas para tal situação, John Nelder e Robert Wedderburn [48], propuseram uma classe de modelos estatísticos que visaram a generalização dos Modelos Lineares Clássicos. A essa teoria unificadora de modelagem estatística, denominaram de Modelos Lineares Generalizados. Em [11], infere-se que os GLM's envolvem uma variável resposta univariada Y relacionada com um conjunto de covariáveis $x_1, x_2, ..., x_p$. Nesse sentido, supondo uma amostra com n observações $(y_i, \boldsymbol{x}_i)$, em que $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ é o vetor coluna de covariáveis, existem três componentes associados aos GLM's, a saber, o componente aleatório, o componente sistemático e a função de ligação.

• **Componente aleatório**: É a variável resposta, constituído por um conjunto de variáveis aleatórias independentes $Y_1, ..., Y_n$, obtidas de uma mesma distribuição que pertence à uma família paramétrica, denominada família exponencial de distribuições [11].

A família exponencial uniparamétrica é caracterizada por uma função (de densidade ou probabilidade) que depende de um parâmetro desconhecido θ, descrita em [11], na forma

$$f(x; \theta) = c(x) \cdot \exp[\zeta(\theta) \cdot t(x) - \psi(\theta)], \tag{2.5}$$

em que as funções $\zeta(\theta)$, $\psi(\theta)$, $t(x)$ e $c(x)$ tem por domínio o conjunto dos números reais e não apresentam unicidade.

Em [11], tem-se que a família exponencial pode ser analisada considerando as funções $\zeta(\theta)$ e $t(x)$ como as funções identidades, isto é, $\zeta(\theta) = \theta$ e $t(x) = x$. É então denominada de família exponencial na forma canônica e pode ser reescrita como

$$f(x;\theta) = c(x) \cdot \exp[\theta x - \psi(\theta)], \tag{2.6}$$

em que θ é denominado de parâmetro canônico. Esta abordagem possui grande importância, já que é possível calcular os dois primeiros momentos em termos de derivadas da função $\psi(\theta)$ em relação ao parâmetro canônico θ.

Embora se tenha essa vantagem, nos GLM's a variância não é mais constante. Em [11] tem-se que é necessário um parâmetro para capturar essa variabilidade. Nesse sentido, quando se observa o componente aleatório de um GLM, deve-se assumir que a função de densidade (ou de probabilidade) de Y possa ser expressa na forma

$$f(y;\theta,\phi) = \exp\{\phi^{-1}[y\theta - \psi(\theta)] + v(y,\phi)\}, \tag{2.7}$$

em que $\phi > 0$ é um parâmetro de dispersão, que é uma medida de dispersão da distribuição. Além disso, quando se tem ϕ conhecido, a família de distribuições apresentada em (2.7) é idêntica a família exponencial na forma canônica, indicada em (2.6). O modelo apresentado por meio da função (2.7) é denominado de família exponencial linear ou família exponencial de dispersão na forma canônica [11].

Quando considera-se a família exponencial de dispersão, a média e a variância da variável reposta Y são dadas em [11] por

$$\mathrm{E}(Y) = \mu = \psi'(\theta) \quad e \quad \mathrm{Var}(Y) = \phi \cdot \psi''(\theta), \tag{2.8}$$

em que a função da média μ na variância é representada por $\psi''(\theta) = V(\mu)$, que é denominada de função da variância.

Alguns exemplos de distribuições que pertencem à família exponencial de dispersão são: a distribuição normal (logo, um modelo linear clássico é um caso particular dos GLM's), as distribuições gama e normal inversa (que são distribuições assimétricas, com $Y > 0$), as distribuições Poisson e binomial (distribuições discretas), entre outras [11].

Este é um grande avanço frente ao modelo linear clássico, já que foi retirado o forte pressuposto de que a variável resposta segue a distribuição normal, sendo agora possível atribuir qualquer distribuição que pertença à família exponencial de dispersão. Isso possui muitas vantagens, uma vez que é possível, nos GLM's, a modelagem de dados positivos, assimétricos, dados de contagem, proporção, entre outros [26].

• **Componente sistemático**: Em [11], tem-se que é no preditor linear do modelo que são inseridas as variáveis explicativas, por meio de uma combinação linear dos

parâmetros desconhecidos, ou seja,

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \boldsymbol{x}_i^T\boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \quad \text{com } i = 1, ..., n, \tag{2.9}$$

em que $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T$ é a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ é o vetor de parâmetros desconhecidos e $\boldsymbol{\eta} = (\eta_1.\eta_2, ..., \eta_n)^T$ é o preditor linear.

• **Função de Ligação**: Em [11] observa-se que a função de ligação é uma função que relaciona o componente aleatório com o componente sistemático, isto é, associa a média com o preditor linear. Nesse sentido,

$$\eta_i = g(\mu_i), \tag{2.10}$$

sendo $g(\cdot)$ uma função monótona e diferenciável.

Por meio das equações (2.9) e (2.10), pode-se obter

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \implies \mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}), \tag{2.11}$$

ou seja, em um GLM temos uma função da média que é relacionada com o preditor linear e, que pode ser escrita como uma combinação linear das covariáveis com os parâmetros, adicionado ao β_0 (intercepto).

Para se estimar o vetor de parâmetros $\boldsymbol{\beta}$, pode-se utilizar o Método de Máxima Verossimilhança (MV), o Método Bayesiano (MB), entre outros. Considerando o Método de MV para estimar o vetor de parâmetros $\boldsymbol{\beta}$, utiliza-se o Algoritmo da Estimação para os GLM's que é representado em [11] por

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \tag{2.12}$$

em que m indica o número da iteração do algoritmo; $\mathbf{X}$ é a matriz do modelo; $\mathbf{W}$ é uma matriz diagonal de pesos que capta a informação sobre a distribuição e a função de ligação utilizadas; $\mathbf{z} = \boldsymbol{\eta}^{(m)} + \mathbf{G}^{(m)}(\mathbf{y} - \boldsymbol{\mu}^{(m)})$ é denominada de variável dependente ajustada, em que $\mathbf{G}$ é uma matriz diagonal formada pelas derivadas de primeira ordem da função de ligação; $\mathbf{y}$ é o vetor de dados e $\boldsymbol{\mu}$ o vetor de médias [11].

Ainda segundo [11], uma observação bastante importante, se refere ao fato de que o algoritmo apresentado em (2.12), quando avaliado em um modelo de regressão linear clássico (com distribuição normal), reduz no estimador de mínimos quadrados ponderados, em que $\mathbf{W}^{(m)} = \boldsymbol{I}$ (matriz identidade) e $\mathbf{z}^{(m)} = y$.

Em síntese, pode-se observar que os GLM's apresentam flexibilizações em relação aos modelos lineares de regressão clássicos, destacando-se a possibilidade de diversas

distribuições de probabilidade para a variável resposta (desde que pertençam à família exponencial de dispersão), não restringindo somente à distribuição normal. Além disso, tem-se a possibilidade de uso da função de ligação para modelar a conexão entre a média com o preditor linear.

Ainda que apresentem diversas vantagens quando comparado com os modelos clássicos de regressão linear, em [17] tem-se que os GLM's dispõe de uma estrutura bastante restrita, principalmente pelo aspecto de considerar uma relação linear entre a média com as variáveis explicativas. Um alternativa viável para superar estas dificuldades, refere-se a utilização das regressões não-paramétricas ou semi-paramétricas. Em [17] tem-se que nestes modelos, a função da média é estimada sem relação a uma forma funcional pré-existente, possibilitando maior flexibilidade e versatilidade para tais regressões.

Os modelos aditivos generalizados são uma classe de modelos que foram introduzidos por Hastie e Tibshirani em [27], em que a inovação consistiu em inserir técnicas de suavização que não existiam nos GLM's. Nesse sentido, observa-se em [67] que um GAM corresponde a um Modelo Linear Generalizado, com preditores lineares apresentando somas de funções suavizadoras de variáveis explicativas, isto é

$$\begin{aligned} g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + h_1(x_{1i}) + h_2(x_{2i}) + h_3(x_{3i}, x_{4i}) + \ldots \implies \\ g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + \sum_j h_j(x_{ji}), \end{aligned} \tag{2.13}$$

em que Y é a variável resposta; Y_i segue alguma distribuição que pertence à família exponencial, apresentando como vetor de médias $\boldsymbol{\mu}$ e parâmetro de dispersão ϕ; $\mathbf{X}_i^*$ é a i-ésima linha da matriz de delineamento para qualquer componente paramétrico do modelo; $\boldsymbol{\theta}$ é o vetor de parâmetros correspondente à matriz de delineamento; h_j são as funções suavizadoras não-paramétricas das covariáveis x_j que, nesta pesquisa, serão consideradas as funções denominadas de *splines*.

Embora este modelo permita um ajuste por meio de funções suavizadoras, evitando alguns modelos complexos em sua manipulação, o mesmo apresenta algumas questões que requerem cuidado, dentre os quais se destacam a representação destas funções de suavização, bem como a escolha de quão suaves elas são [67].

Com relação a estimação das funções suavizadoras h, em [67] infere-se que pode ser feita escolhendo-se uma base definida no espaços das funções, de maneira que a função h seja um dos elementos da base. Assim, por meio de penalizações nessas bases, em [68] verifica-se que cada função pode ser representada por

$$h_j(x_{ji}) = \sum_{k=1}^{k_j} \beta_{jk} b_{jk}(x_{ji}),$$

em que $b_{jk}(x_{ji})$ são funções de base conhecidas, que são escolhidas por terem propriedades convenientes, enquanto β_{jk} são parâmetros que devem ser estimados.

Por meio de reparametrizações, o modelo apresentado em (2.13) pode ser reescrito centrado matricialmente para cada termo de suavização, da seguinte maneira

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \tag{2.14}$$

em que $\mathbf{X} = [X^* : X_1 : X_2 : ...]$ e $\boldsymbol{\beta}^T = [\theta^T, \beta_1^T, \beta_2^T, ...]$. Nesse sentido, o modelo (2.14) refere-se a uma GLM, podendo-se escrever o logarítmo da sua função de verossimilhança $l(\beta)$.

Em [67] observa-se que nos GAM's, tem-se a existência de penalizações objetivando-se o ajuste de mínimos quadrados. Por exemplo, o modelo (2.14) pode ser ajustado minimizando

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \int_0^1 [h''(x)]^2 dx, \tag{2.15}$$

em que λ é denotado de parâmetro de suavização; e a integral da segunda derivada de h, penaliza os modelos. Quem realiza a compensação entre o ajuste do modelo e sua suavidade é o parâmetro de suavização. Como h é linear nos parâmetros β_i, com $i = 1, \ldots, n$, tem-se que a penalidade sempre pode ser reescrita como uma forma quadrática em β, isto é,

$$\int_0^1 [h''(x)]^2 dx = \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}, \tag{2.16}$$

em que $\mathbf{S}$ é uma matriz de coeficientes conhecida. Assim, por meio das equações (2.15) e (2.16), o ajuste da regressão por *splines* penalizados, se reduz a minimizar

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}, \tag{2.17}$$

em relação a β, sendo λ conhecido.

Em [67] verifica-se que a expressão formal para minimizar (2.17) fundamenta-se na estimação de mínimos quadrados penalizados de β, especificada por

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S}\right)^{-1} \cdot \mathbf{X}^T\mathbf{y}.$$

Assim, conhecida uma medida de oscilação para cada função, pode-se obter o logarítmo da função de verossimilhança penalizada para o modelo, que é dada em [67] por

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j\boldsymbol{\beta}^T\mathbf{S}_j\boldsymbol{\beta}, \tag{2.18}$$

em que os λ_j são os parâmetros de suavização, que realizam uma regulação entre a qualidade de ajuste do modelo e a sua suavidade. Nesse sentido, considerando os valores para os parâmetros de suavização, pode-se maximizar o logarítmo da função de verossimilhança penalizada do modelo, com intuito de obter $\hat{\beta}$. No entanto, os λ_j devem ser estimados por métodos computacionais. Este método de estimação dos GAM's é denominado de mínimos quadrados penalizados

iterativamente reponderados (IRLS) [67].

Em síntese, nos modelos GAM's tem-se uma grande flexibilidade ao inserir as funções suavizadoras, uma vez que possibilita que o relacionamento entre a variável resposta e as variáveis explicativas não seja obrigatoriamente linear, como nos GLM's. Contudo, uma ressalva em se introduzir as funções suavizadoras refere-se ao fato de não se ter uma interpretabilidade simples como se observa nos GLM's. Além disso, tanto para os GAM's quanto para os GLM's, ainda se tem a restrição da média (μ) ser modelada necessariamente por distribuições pertencentes à família exponencial de dispersão.

Buscando-se por alternativas para situações como a supracitada, Rigby e Stasinopoulos em [56], propuseram uma classe de modelos de regressão denominada de Modelos Aditivos Generalizados para Locação, Escala e Forma. Na seção a seguir, são apresentadas em detalhes sua estrutura e propriedades.

3 MATERIAIS E MÉTODOS

Para uma compreensão pormenorizada dos aspectos metodológicos desta pesquisa, são apresentadas a definição e as propriedades referentes a classe de modelos de regressão utilizada, bem como procedimentos para seleção e análise diagnóstica de modelos.

3.1 MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA

3.1.1 Definição de um GAMLSS

Em [56], pode-se observar que nos Modelos Aditivos Generalizados para Locação, Escala e Forma, termos aditivos, paramétricos e aleatórios podem ser considerados para a modelagem dos q parâmetros $\boldsymbol{\theta}^T = (\theta_1, \theta_2, ..., \theta_q)$, de uma função (densidade) de probabilidade $f(y|\boldsymbol{\theta})$.

Este modelo assume que para todo $i = 1, ..., n$, as y_i variáveis respostas são independentes, condicionais a $\boldsymbol{\theta^i}$, com função de probabilidade (densidade) $f(y_i|\boldsymbol{\theta}^i)$, em que $\boldsymbol{\theta}^{iT} = (\theta_{i1}, \theta_{i2}, ..., \theta_{iq})$ é o vetor de q parâmetros relacionado com as variáveis explicativas e efeitos aleatórios.

Assim sendo, considerando $\mathbf{y}^T = (y_1, y_2, ..., y_n)$ o vetor de observações da variável resposta, e $g_t(\cdot)$ uma função de ligação monótona, com $t = 1, ..., q$, tal que relacione θ_t com as variáveis explicativas e efeitos aleatórios por meio de um modelo aditivo dado por

$$g_t(\boldsymbol{\theta}_t) = \boldsymbol{\eta}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \sum_{j=1}^{J_t} \boldsymbol{Z}_{jt}\boldsymbol{\gamma}_{jt}, \tag{3.1}$$

em que $\boldsymbol{\theta}_t$ é o vetor referente ao $t-$ésimo parâmetro da distribuição considerada para a variável resposta $\mathbf{Y}$, referente ao vetor $\boldsymbol{\eta}_t$ (seu preditor linear); $\mathbf{X}_t$ é a matriz das covariáveis de tamanho $n \times J_t'$; $\boldsymbol{\beta_t}^T = (\beta_{1t}, \beta_{2t}, ..., \beta_{J't})$ é um vetor paramétrico dos efeitos fixos de tamanho J_t'; $\mathbf{Z}_{jt}$ é uma matriz de termos aditivos; γ_{jt} são vetores de variáveis aleatórias que podem ser combinados em um único vetor $\boldsymbol{\gamma}_t$, associado a uma matriz $\mathbf{Z}_{(t)}$. Em [56], o modelo definido em (3.1) é denominado de GAMLSS.

Como já observado, os vetores $\boldsymbol{\gamma}_{jt}$, podem ser manipulados e combinados em um único vetor $\boldsymbol{\gamma}_t$ com uma única matriz $\mathbf{Z}_t$ [17]. Contudo, a formulação apresentada em (3.1) é mais apropriada, uma vez que facilita a utilização do algoritmo de auto ajuste para estimar os parâmetros (*backfitting*), além de possibilitar que combinações de diferentes termos aditivos e de efeitos aleatórios sejam facilmente implementados no modelo.

Observe que para o caso em que não existem termos aditivos associados aos parâmetros da distribuição ($J_t = 0$), o modelo (3.1) se reduz a um modelo linear estritamente paramétrico, dado por

$$g_t(\boldsymbol{\theta}_t) = \boldsymbol{\eta}_t = \mathbf{X}_t\boldsymbol{\beta}_t.$$

Para o caso em que $\mathbf{Z}_{jt} = \mathbf{I}_n$, em que $\mathbf{I}_n$ é uma matriz identidade de ordem n, e $\boldsymbol{\gamma}_{jt} = \mathbf{h}_{jt} = h_{jt}(\mathbf{x}_{jt})$ para todas as combinações de j e t no modelo (3.1), tem-se que

$$g_t(\boldsymbol{\theta}_t) = \boldsymbol{\eta}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \sum_{j=1}^{J_t} h_{jt}(\mathbf{x}_{jt}), \tag{3.2}$$

em que $\mathbf{x}_{jt}$, para $j = 1, 2, ..., J_t$ e $t = 1, 2, ..., q$, são vetores de tamanho n. A função h_{jt} é uma função desconhecida das variáveis explicativas $\mathbf{X}_{jt}$ e $\mathbf{h}_{jt} = h_{jt}(\mathbf{x}_{jt})$ é um vetor que avalia a função h_{jt} em $\mathbf{x}_{jt}$. Para essa situação, considera-se que os vetores x_{jt} sejam conhecidos, e o modelo (3.2) é um caso especial do modelo (3.1) que é denominado em [56] por GAMLSS semiparamétrico.

Além disso, em [57] observa-se que o modelo (3.1) pode ser estendido, possibilitando que termos paramétricos não lineares sejam incluídos nos parâmetros da distribuição, sendo apresentado na forma

$$g_t(\boldsymbol{\theta}_t) = \boldsymbol{\eta}_t = \omega_t(\mathbf{X}_t\boldsymbol{\beta}_t) + \sum_{j=1}^{J_t} h_{jt}(\mathbf{x}_{jt}), \tag{3.3}$$

em que ω_t, com $t = 1, ..., q$ são funções não lineares conhecidas e $\mathbf{X}_t$ é uma matriz de variáveis explicativas conhecidas, de ordem $n \times J_t''$. O modelo em (3.3) é denominado de GAMLSS semiparamétrico não linear. Para $J_t = 0$, ou seja, se os parâmetros da distribuição não possuir termos aditivos, então o modelo (3.3) é reduzido a um modelo GAMLSS paramétrico não linear, dado por

$$g_t(\boldsymbol{\theta}_t) = \boldsymbol{\eta}_t = \omega_t(\mathbf{X}_t\boldsymbol{\beta}_t). \tag{3.4}$$

Assim, para modelar uma variável resposta Y que segue uma distribuição de probabilidade com quatro parâmetro, isto é $q = 4$, verifica-se em [12] que a classe de modelos GAMLSS semiparamétricos pode ser representada por $\boldsymbol{Y} \sim \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$, de maneira que cada parâmetro distribucional é estimado por um preditor particular, dado por

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + h_{11}(\mathbf{x}_{11}) + \ldots + h_{1J_1}(\mathbf{x}_{1J_1}),$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + h_{21}(\mathbf{x}_{21}) + \ldots + h_{2J_2}(\mathbf{x}_{2J_2}),$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + h_{31}(\mathbf{x}_{31}) + \ldots + h_{3J_3}(\mathbf{x}_{3J_3}),$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + h_{41}(\mathbf{x}_{41}) + \ldots + h_{4J_4}(\mathbf{x}_{4J_4}), \tag{3.5}$$

em que $\mathcal{D}(\mu, \sigma, \nu, \tau)$ é a distribuição considerada para a variável resposta $\mathbf{Y}$; $g_1(\cdot)$, $g_2(\cdot)$, $g_3(\cdot)$ e $g_4(\cdot)$ são as funções de ligação dos parâmetros da distribuição (μ, σ, ν, τ), respectivamente; $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ e $\mathbf{X}_4$ são as matrizes referentes as variáveis explicativas (parte paramétrica do modelo); $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ e $\boldsymbol{\beta}_4$ são os vetores dos coeficientes lineares da parcela paramétrica do modelo e, para $t = 1, 2, 3, 4$ e $j = 1, 2, \ldots, j_t$, $h_{tj}(\mathbf{x}_{tj})$ são funções suavizadoras das variáveis explicativas x_{tj} inseridas aditivamente nos preditores dos parâmetros da distribuição considerada para o modelo.

Por meio das obras de [55], [56] e [12], observa-se que as funções suavizadoras podem ser reescritas como $h(\mathbf{x}) = \mathbf{Z}_{\boldsymbol{\gamma}}$, em que $\mathbf{Z}$ é a matriz constituída por funções geradoras da função suavizadora e γ é o vetor dos coeficientes a serem estimados, relacionados com uma penalização que é expresso na forma quadrática $\boldsymbol{\lambda}^*\boldsymbol{\gamma}^T\mathbf{G}\boldsymbol{\gamma}$, tal que $\boldsymbol{\lambda}^*$ é o parâmetro de suavização, $\mathbf{G}$ é a matriz de penalização que é expressa na forma $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ e $\mathbf{D}$ é uma matriz de diferenças.

Sob essa perspectiva, a estrutura apresentada em (3.5) pode ser generalizada por

$$
\begin{aligned}
g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\
g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\
g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 &= \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\
g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 &= \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4},
\end{aligned} \tag{3.6}
$$

de maneira que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_4^T)$ são os vetores dos parâmetros dos efeitos fixos e $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}^T, \boldsymbol{\gamma}_{12}^T, \ldots, \boldsymbol{\gamma}_{1J_1}^T, \boldsymbol{\gamma}_{21}^T, \ldots, \boldsymbol{\gamma}_{4J_4}^T)$ são os coeficientes dos efeitos aleatórios.

Assim, nota-se que a classe de modelo GAMLSS apresentada em (3.1) é mais geral que os modelos Lineares Clássicos, GLM e GAM, uma vez que a distribuição da variável resposta não restringe-se à família exponencial, bem como todos os parâmetros (e não apenas a média) são modelados considerando termos de efeitos aleatórios e/ou fixos.

3.1.2 Estimação do Modelo

Ao considerar a estimação do modelo, verifica-se em [56], que para o ajuste dos componentes aditivos em uma estrutura GAMLSS faz-se fundamental o algoritmo *back-fitting* (retroajuste), bem como penalidades quadráticas no logarítmo da função de verossimi-

lhança. Nesse sentido, a estimação resultante utiliza matrizes suavizadas como componente do algoritmo de retroajuste.

Outro aspecto importante no GAMLSS, refere-se sobre a hipótese de independência entre os diferentes vetores γ_{jt} de efeitos aleatórios. Contudo, se para um valor particular de t, dois ou mais vetores de efeitos aleatórios não são indenpendentes, eles podem ser combinados de modo a obter um único vetor, com matriz de correspondência, $\mathbf{Z}_{jt}$, única para que a condição de independência seja satisfeita.

Nesse sentido, tem-se em [63] que para o ajuste do modelo (3.6), é necessário estimar os vetores dos parâmetros dos efeitos fixos $\boldsymbol{\beta}$, os coeficientes dos efeitos aleatórios $\boldsymbol{\gamma}$ e também o vetor de hiperparâmetros $\boldsymbol{\lambda}^*$.

Em [56] verifica-se que pode-se estimar os $\beta_{t's}$ e os $\gamma_{jt's}$ considerando o vetor de hiperparâmetros $\boldsymbol{\lambda}^*$ fixo ou, de uma maneira mais geral, estimar tal vetor. Analisando-se primeiramente para o caso em que λ^*_{jt} é fixo, segundo estes mesmos autores, os $\beta_{t's}$ e os $\gamma_{jt's}$ podem ser estimados no GAMLSS pela maximização do logarítmo da função de verossimilhança penalizada, l_{pen}, dada por

$$l_{pen} = l - \frac{1}{2}\sum_{t=1}^{q}\sum_{j=1}^{J_t}\lambda_{jt}\gamma_{jt}^T\boldsymbol{G}_{jt}\gamma_{jt}, \tag{3.7}$$

em que $l = \sum_{i=1}^{n}\log f(y_i|\boldsymbol{\theta^i})$ é o logarítmo da função de verossimilhança dos dados condicionais a $\boldsymbol{\theta^i}$, com $i = 1, ..., n$.

Visando-se a maximização do logarítmo da função de verossimilhança penalizada em (3.7), no *software* R, é possível utilizar dois algoritmos distintos: o algoritmo CG (Cole e Green), e o algoritmo RS (Rigby e Stasinopoulos), ambos generalizações de modelos apresentados em [9] e [54], respectivamente. Em [64], tem-se que o algoritmo iterativo RS maximiza o logarítmo da função de verossimilhança (penalizada) sobre cada um dos parâmetros, realizando ciclos até sua convergência. Por ser mais rápido e mais estável, é o método padrão utilizado no *software* R. Este algoritmo apresenta três partes interrelacionadas: as iterações externas; as iterações internas ou GLIM e o algoritmo de *backfitting*, cada um relacionado com o anterior. A Figura 3.1 ilustra a estrutura do algoritmo RS que, também se aplica ao algoritmo CG.

Embora a estrutura dos algoritmos RS e CG sejam similares, como indicado na Figura 3.1, o algoritmo CG apresenta um grande distinção ao RS, fundamentalmente ao que se refere no uso de valores esperados das derivadas cruzadas do logarítmo da função de verossimilhança em relação a cada um dos parâmetros da distribuição, possibilitando a atualização conjunta de todos os vetores de parâmetros. Contudo, para algumas funções de densidade de probabilidade $f(y|\boldsymbol{\theta})$, os parâmetros $\boldsymbol{\theta}$ são ortogonais, acarretando em valores esperados das derivadas cruzadas do logarítmo da função de verossimilhança serem iguais a zero [64].

Nestes casos, o algoritmo CG não é recomendado. Em contrapartida, o al-

Figura 3.1: Representação dos três ciclos dos algoritmos GAMLSS.

Fonte: Adaptado do trabalho de [64].

goritmo CG é adequado ao considerar distribuições que apresentam estimadores de parâmetros extremamente correlacionados, uma vez que, em [64] infere-se que nestes casos, o algoritmo RS pode ser moroso, além de possibilidade de convergência para um máximo local do logarítmo da função de verossimilhança ao invés do máximo global.

Por outro lado, de uma maneira mais geral, pode-se estimar os hiperparâmetros de suavização $\boldsymbol{\lambda}^*$. Para tanto, em [63] tem-se que existem diversas estratégias para este processo. Ainda segundo os mesmos autores, estas podem ocorrer:

• Globalmente: quando o método para estimar o hiperparâmetro de suavização é estabelecido fora dos algoritmos RS ou CG.

• Localmente: quando o método para estimar cada λ_{jt}^* do hiperparâmetro de suavização é aplicado internamente ao algoritmo de *backfitting* dos algoritmos RS ou CG.

Além disso, nas palavras de [63] existem pelo menos três métodos distintos para se estimar os hiperparâmetros, à saber, os métodos fundamentados na verossimilhança (ML/REML), o critério de informação de Akaike generalizado (GAIC); e técnicas de validação como a validação cruzada generalizada (GCV).

Em [63], tem-se que os métodos locais são os mais rápidos e geralmente produzem resultados similares aos métodos globais. Embora os métodos globais possam apresentar resultados mais confiáveis, estes são computacionalmente intensivos.

3.1.3 Família de distribuições nos modelos GAMLSS

Como observado na definição dos GAMLSS, diferentemente dos GLM's e dos GAM's, sua função densidade de probabilidade $f(y|\boldsymbol{\theta})$ não mais restringe-se a família exponencial de dispersão, sendo ampliada para outras estruturas. Contudo, existe uma limitação para a implementação do modelo GAMLSS. Nas palavras de [58], a única imposição existente é que a função $f(y|\boldsymbol{\theta})$ e suas primeiras derivadas com relação a cada um dos parâmetros devem ser computáveis. Ademais, as derivadas explícitas são preferíveis a derivadas numéricas, sendo que estas últimas também podem ser utilizadas, contudo com potencial de acarretar em velocidade computacional prejudicada.

Existem atualmente mais de 100 distribuições implementadas no *software* R, junto ao pacote **gamlss.dist**, de maneira que a distribuição a ser considerada depende da natureza da variável resposta Y. Em [59] tem-se que as distribuições encontradas no pacote **gamlss.family** (*software* R) pertencem a três tipos distintos de distribuições, à saber, distribuições de probabilidade para variáveis aleatórias discretas, contínuas e mistas. As Tabelas 3.1 e 3.2 apresentam exemplos de um número reduzido de famílias de distribuições contínuas e discretas que podem ser encontradas no *software* R.

Tendo em mente a grande quantidade de distribuições que estão implementadas neste *software*, na próxima subseção são apresentadas características e propriedades para a seleção do modelo considerado nos GAMLSS.

3.1.4 Seleção de Modelos

Os GAMLSS podem ser representados por $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L}\}$, em que $\mathcal{D}$ especifica a distribuição da variável resposta; $\mathcal{G}$ representa o conjunto de funções de ligação $(g_1, g_2, ..., g_q)$ para os parâmetros $(\theta_1, \theta_2, ..., \theta_q)$ da distribuição; $\mathcal{T}$ caracteriza os termos nos preditores $(z_1, z_2, ..., z_q)$ para os preditores $(\eta_1, \eta_2, ..., \eta_q)$; $\mathcal{L}$ define os parâmetros de suavização associados a cada um dos parâmetros da distribuição da variável resposta [55].

Note que, devido a grande flexibilidade que os modelos GAMLSS oportunizam, diversas combinações entre os quatro componentes de $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L}\}$ podem ser utilizadas no processo de construção de um modelo. Uma das alternativas para tal comparação é o Critério de Informação de Akaike Generalizado (GAIC), dado em [63] por

$$\text{GAIC} = -2\hat{l} + (\kappa\ df),$$

em que $\hat{l} = \sum_{i=1}^{n} \log f(y_i|\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ é o logarítmo da função de verossimilhança ajustada, κ uma penalidade exigida e df são os graus de liberdade do modelo proposto.

Observe que, casos especiais do GAIC são especificados dependendo do valor de κ. Se $\kappa = 2$, tem-se o Critério de Informação de Akaike (AIC); se $\kappa = \log(n)$ tem-se o Critério Bayesiano de Schwarz (SBC) [2].

Tabela 3.1: Famílias de distribuições contínuas implementadas no *software* R.

Distribuição	Família	Suporte	Funções de ligação de			
			μ	σ	ν	τ
Beta	BE	(0,1)	logit	logit	-	-
Box-Cox Cole-Green	BCCG	$(0, \infty)$	ident.	log	ident.	-
Box-Cox Power Expon.	BCPE	$(0, \infty)$	ident.	log	ident.	log
Box-Cox t	BCT	$(0, \infty)$	ident.	log	ident.	log
Exponencial	EXP	$(0, \infty)$	log	-	-	-
Exponencial Gaussiana	exGAUS	$(-\infty, \infty)$	ident.	log	log	-
Gamma	GA	$(0, \infty)$	log	log	-	-
Gamma Generalisada	GG	$(0, \infty)$	log	log	ident.	-
Gumbel	GU	$(-\infty, \infty)$	ident.	log	-	-
Gamma Inversa	IGAMMA	$(0, \infty)$	log	log	-	-
Johson's SU	JSU	$(-\infty, \infty)$	ident.	log	ident.	log
Johson's SU original	JSUo	$(-\infty, \infty)$	ident.	log	ident.	log
Logística	LO	$(-\infty, \infty)$	ident.	log	-	-
Log Normal	LOGNO	$(0, \infty)$	ident.	log	-	-
Normal	NO	$(-\infty, \infty)$	ident.	log	-	-
Power Exponencial	PE	$(-\infty, \infty)$	ident.	log	log	-
Skew Normal Tipo 1	SN1	$(-\infty, \infty)$	ident.	log	ident.	-
Skew Normal tipo 2	SN2	$(-\infty, \infty)$	ident.	log	log	-
Skew t Tipo 1	ST1	$(-\infty, \infty)$	ident.	log	ident.	log
Skew t Tipo 2	ST2	$(-\infty, \infty)$	ident.	log	ident.	log
Weibull	WEI	$(0, \infty)$	log	log	-	-

Fonte: Adaptado da obra [59].

Tabela 3.2: Famílias de distribuições discretas implementadas no *software* R.

Distribuição	Família	Suporte	Funções de ligação de			
			μ	σ	ν	τ
Binomial	BI	{0,1,...,n}	logit	-	-	-
Geométrica	GEOM	{ 0,1,2,... }	log	-	-	-
Logarítmica	LG	$1, 2, ..., \infty$	logit	-	-	-
Multinomial	MULTIN	$1, 2, ..., n$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	-
Binomial Negativa Tipo 1	NBI	{ 0,1,2,... }	log	log	-	-
Binomial Negativa Tipo 2	NBII	{ 0,1,2,... }	log	log	-	-
Poisson	PO	{ 0,1,2,... }	log	-	-	-
Poisson Gaussiana Inversa	PIG	{ 0,1,2,... }	log	log	-	-

Fonte: Adaptado da obra [59].

Outra possibilidade para a escolha de um modelo adequado é por meio da

utilização do coeficiente de determinação generalizado (R^2), dado em [46] por

$$\mathrm{R}^2 = 1 - \exp\left[-\frac{2}{n}\{l(\hat{\beta}) - l(0)\}\right] = 1 - \left[\frac{L(0)}{L(\hat{\beta})}\right]^{\frac{2}{n}},$$

em que $l(0) = \log L(0)$ e $l(\hat{\beta}) = \log L(\hat{\beta})$ representam o logarítmo das funções de verossimilhança do modelo nulo com apenas o termo constante, e do modelo ajustado, respectivamente. O modelo que apresentar o menor valor de GAIC e/ou maior valor para o R^2 generalizado será selecionado.

Seleção da distribuição $\mathcal{D}$

No que se refere a seleção da distribuição, deve-se observar inicialmente o suporte da variável resposta Y, ou seja, se os dados inidicarem valores de Y estritamente positivos, é razoável que se pense em uma distribuição que possua suporte positivo. Da mesma maneira, se Y for uma variável aleatória discreta, é razoável pensar em distribuições discretas.

A partir do suporte da variável resposta, diversas distribuições podem ser tomadas como candidatas para o processo de implementação do modelo. Para verificar aquela que melhor se ajusta aos dados, em [56] infere-se que é selecionada aquela que apresenta o menor valor do GAIC.

Para colaborar no processo de seleção da distribuição da variável resposta Y, verifica-se em [61], que o *software* R apresenta a função **fitDist(**$\cdot$**)**. Essa função permite o ajuste de distribuições paramétricas da família GAMLSS, especificadas pelos diferentes argumentos utilizados. Destaca-se aqui, o argumento **type(**$\cdot$**)** pertencente a função **fitDist(**$\cdot$**)**, uma vez que aqui é indicado o suporte da variável resposta, tal como *realline*, *realplus* ou *realAll*.

Segundo [62], a opção *realline* selecionará distribuições em que Y varia entre todos os valores reais necessariamente $(-\infty < Y < +\infty)$. Já a opção *realplus* selecionará distribuições de maneira que $Y > 0$, necessariamente. Por fim, tem-se que *realAll* é uma combinação das opções anteriores, considerando em um único argumento ambos os ajustes de distribuições. A distribuição marginal final definida é aquela selecionada pelo GAIC, com penalidade κ. O *default* do *software* R é $\kappa = 2$, ou seja, AIC.

Outra função que possibilita o ajuste de distribuições paramétricas da família GAMLSS, é a função **chooseDist(**$\cdot$**)** [12]. Diferentemente da função **fitDist(**$\cdot$**)**, que indica as melhores distribuições marginais para a variável resposta, não considerando as covariáveis do modelo a ser ajustado, a função **chooseDist(**$\cdot$**)** informa as melhores distribuições marginais a partir de um modelo GAMLSS já ajustado, considerando para tanto, as covariáveis inseridas neste modelo preliminar.

Assim como para a função **fitDist(**$\cdot$**)**, na função **chooseDist(**$\cdot$**)**, a distribuição marginal escolhida é selecionada pelo Critério Informação de Akaike Generalizado, com penalidade κ. O *default* do *software* R é $\kappa = 2$, ou seja, AIC [12].

Considerando as diversas distribuições implementadas no *software* R, a seguir é apresentada uma discussão introdutória sobre a distribuição Box Cox t, uma vez que a mesma foi utilizada neste trabalho.

• Distribuição Box- Cox t (BCT)

Em [6], Box e Cox propuseram algumas transformações nas variáveis consideradas para modelagem, visando-se a construção de uma nova família de distribuições, à qual foi denominada de distribuição normal de potência truncada.

Diferentemente da distribuição normal, as distribuições que pertencem à esta família, possuem propriedades menos restritivas, destacando-se pela não necessidade da normalidade dos dados, de sua independência e da variância constante. Isso se justifica, uma vez que em [6], verifica-se que tais suposições não são necessárias inicialmente para estas variáveis, mas sim, após algumas transformações adequadas realizadas. Nesta família, se encontram algumas distribuições, dentre as quais tem-se a distribuição normal e a log-normal [12].

A distribuição BCT é uma generalização da família de distribuições normal propostas por Box e Cox em [6], no qual apresenta importantes características no processo de modelagem de dados com assimetria e elevada curtose [57]. Essa distribuição foi desenvolvida em [57] por Rigby e Stasinopoulos, no qual apresenta quatro parâmetros, visando a modelagem de variáveis respostas $Y > 0$. Sua representação é dada por BCT(μ, σ, ν, τ) [57].

Ressata-se ainda que, a distribuição BCT apresenta diversas vantagens frente à outras distribuições, dentre as quais tem-se a possibilidade da produção de um modelo flexível para assimetria e leptocurtose, além de possuir a fdp contínua, com derivadas de ordens primárias e superiores contínuas [57].

Por meio do formalismo matemático, considere Y uma variável aleatória contínua positiva, que segue a distribuição BCT(μ, σ, ν, τ), definida por meio de uma variável aleatória transformada Z, dada em [59] por

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], & \text{se } \nu \neq 0, \\ \text{ou} & \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), & \text{se } \nu = 0, \end{cases} \tag{3.8}$$

com $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ e Z segue uma distribuição truncada t, com $\tau > 0$ graus de liberdade.

Nesse sentido, a fdp de Y é dada em [59] por,

$$f_Y(y \mid \mu, \sigma, \nu, \tau) = \frac{y^{\nu-1} f_T(z)}{\mu^{\nu}\sigma F_T\left(\frac{1}{\sigma|\nu|}\right)}, \tag{3.9}$$

em que Z é dada pela equação (3.8) e $f_T(t)$ e $F_T(t)$ são, respectivamente, a fdp e a fda de uma variável aleatória T que segue distribuição t padrão, com graus de liberdade $\tau > 0$, isto é,

$T \sim t[0,1]$, dadas por ([59])

$$f_T(z) = \frac{\Gamma[(\tau+1)/2]}{\Gamma(1/2)\Gamma(\tau/2)\tau^2}\left[1+\frac{z^2}{\tau}\right]^{-\frac{(\tau+1)}{2}} = \frac{1}{B\left(\frac{1}{2},\frac{\tau}{2}\right)\tau^{\frac{1}{2}}}\left[1+\frac{z^2}{\tau}\right]^{-\frac{(\tau+1)}{2}}$$

e

$$F_T(z) = \frac{1}{2} + z\Gamma[(\tau+1)/2] \times \frac{2^{F_1\left(\frac{1}{2},\frac{(\tau+1)}{2};\frac{3}{2};-\frac{z^2}{\tau}\right)}}{\Gamma(1/2)\Gamma(\tau/2)\tau^2},$$

em que Z é definida em (3.8), 2^{F_1} é a função hipergeométrica, $\Gamma(\cdot)$ e $B(\cdot)$ são as funções gama e beta, respectivamente.

De acordo com [59], tem-se que para a distribuição BCTo, sua função densidade é também dada pela equação (3.9), de maneira que difere da distribuição BCT unicamente pela função de ligação do parâmetro de locação (μ), uma vez que para a distribuição BCT a função de ligação é a identidade, enquanto que para a distribuição BCTo é a função logarítmica.

Fundamentados na obra de Rigby e Stasinopoulos [59], é apresentada a Tabela 3.3 que exibe o espaço paramétrico de cada um dos quatro parâmetros da distribuição BCTo, bem como o suporte da variável aleatória Y.

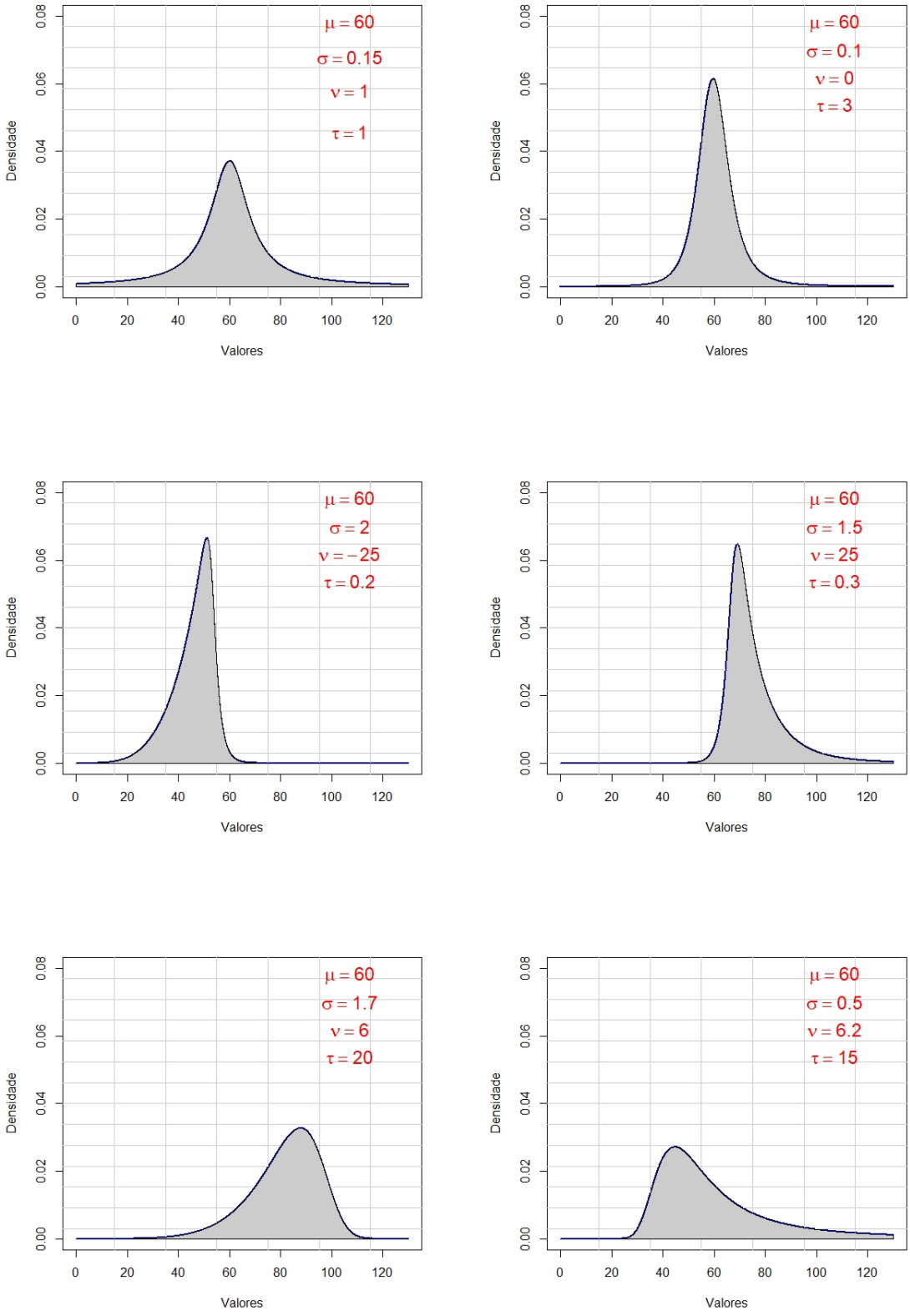
Tabela 3.3: Parâmetros distribucionais de BCTo.

Parâmetros	**Variação**
Y (variável aleatória)	$0 < y < \infty$
μ (parâmetro de escala)	$0 < \mu < \infty$
σ (coeficiente de variação aproximado)	$0 < \sigma < \infty$
ν (parâmetro de assimetria)	$-\infty < \nu < \infty$
τ (parâmetro de curtose)	$0 < \tau < \infty$

Fonte: Adaptado de [59].

Para o estudo do comportamento da curva de densidade da distribuição BCTo, é apresentada a Figura 3.2, que ilustra o perfil assumido por tais curvas para diferentes valores dos parâmetros distribucionais.

Figura 3.2: Variações na curva da densidade da distribuição Box-Cox t .



Fonte: Próprio autor.

Visando-se a complementação da caracterização da distribuição BCTo, o Quadro 3.4 apresenta um resumo de suas principais propriedades. Para mais detalhes, ver [59].

Quadro 3.4: Resumo das propriedades matemáticas e funções da distribuição BCTo.

Medidas e Funções da Distribuição	
Média	-
Mediana	μ
Moda	-
Variância	-
Coeficiente de Assimetria	ν
Coeficiente de Curtose	τ
Função geratriz de momento	-
Função densidade de probabilidade (fdp)	$\frac{y^{\nu-1} f_T(z)}{\mu^{\nu}\sigma F_T\left(\frac{1}{\sigma\lvert\nu\rvert}\right)}$, em que $T \sim t_\tau = TF[0,1]_{(\tau)}$
Função de distribuição acumulada (fda)	$F_T(z)$, em que $T \sim t_\tau$ e z é dado em (3.8)
Inversa da fda (y_p)	$\begin{cases} \mu\left(1+\sigma\nu t_{p,\tau}\right)^{1/\nu}, & \text{se } \nu \neq 0 \\ \mu\exp\left(\sigma t_{p,\tau}\right), & \text{se } \nu = 0 \\ \text{em que } t_{p,\tau} = F_T^{-1}(p) \text{ e } T \sim t_\tau \end{cases}$

Fonte: Adaptado de [59].

Seleção das Funções de Ligação $\mathcal{G}$

Após a seleção da distribuição para a variável resposta, deve-se analisar quais e quantos parâmetros que esta apresenta, para que assim, se verifique as funções de ligação possíveis para cada parâmetro. No pacote do *software* R existem funções que são apresentadas por *default* e, em geral, se respeita tal característica.

Contudo, verifica-se em [63], que para a alteração das funções de ligação dos parâmetros, deve-se respeitar o seu espaço paramétrico, isto é, dependendo dos valores assumidos para um parâmetro, a função de ligação deve ser definida no espaço paramétrico do parâmetro. Por exemplo, ao selecionar a distribuição Box-Cox t (BCTo), com função de ligação logarítmica para o parâmetro μ para representar a variável resposta Y, observa-se a existência de 4 parâmetros, à saber μ, σ, ν e τ. Note que o parâmetro μ assume valores positivos na reta real e, dessa maneira, basta utilizar a função logarítmica como a função de ligação para este parâmetro e, assim, o espaço paramétrico estaria sendo respeitado [59]. Ademais, no *software* R, para a distribuição BCTo, a função de ligação *default* para o parâmetro μ é a função logarítmica.

Uma questão natural que emerge é: caso se tenham duas ou mais funções de ligação que respeitem o espaço paramétrico de um determinado parâmetro, como comparar qual a função mais adequada, segundo o objetivo da pesquisa? Para situações semelhantes à proposta, pode-se utilizar um critério de qualidade de ajuste ou de análise de diagnóstico (GAIC e de gráficos de resíduos).

As Tabelas 3.1 e 3.2 apresentam as funções de ligação, dadas por *default*, para cada parâmetro de algumas distribuições de probabilidade implementadas no R.

Seleção das covariáveis do modelo $\mathcal{T}$

Diferententemente dos modelos anteriormente discutidos (clássico, GLM e GAM), enquanto observava-se um conjunto de características que explicavam simplesmente a média ou uma função da média, nos GAMLSS existe um conjunto de termos que buscam explicar cada um dos parâmetros da distribuição de probabilidade selecionada. Em [63], afirma-se que para uma determinada distribuição de probabilidade para a variável resposta Y, a seleção destes termos deve ser feita para todos os parâmetros da distribuição assumida, e não apenas para o parâmetro de locação (μ).

Distintas abordagens podem ser utilizadas por meio da biblioteca **gamlss** do *software* R. Segundo [63], algumas funções como a **addterm**(·), **dropterm**(·), **stepGAIC**(·) e **stepGAICAll.A**(·) apresentam diversas estratégias para selecionar as covariáveis que fazem parte da estrutura do preditor de um parâmetro distribucional. Ainda segundo estes mesmos autores, tem-se que cada umas das funções supracitadas podem ser utilizadas com os seguintes objetivos:

• **Função dropterm**(·)**:** Quando se busca, a partir de um modelo ajustado com todas as covariáveis, retirar aquelas que não são significativas para a explicação da variável resposta. Para tanto, a covariável que possuir maior $p-$valor será retirada do processo de modelagem.

• **Função addterm**(·)**:** Quando se busca, a partir de um modelo nulo ajustado (sem as covariáveis), adicionar aquelas que são significativas para a explicação da variável resposta, segundo a medida de seus $p-$valores.

Em [63], pode-se verificar dois aspectos extremamente importantes ao se utilizar as estratégias supracitadas. O primeiro deles refere-se ao fato de ser possível realizar a seleção de covariáveis para todos os preditores dos parâmetros da distribuição selecionado, ao se alterar o argumento **parameter**(·) (*default* é para o parâmetro μ). O segundo relaciona-se com a velocidade de processamento de seleção das covariáveis, ou seja, por meio do argumento **paralel**(·), é possível realizar cálculos paralelos, assumindo que a máquina utilizada possui diversos CPUs. Isso apresenta vantagens ao se utilizar grandes conjuntos de dados.

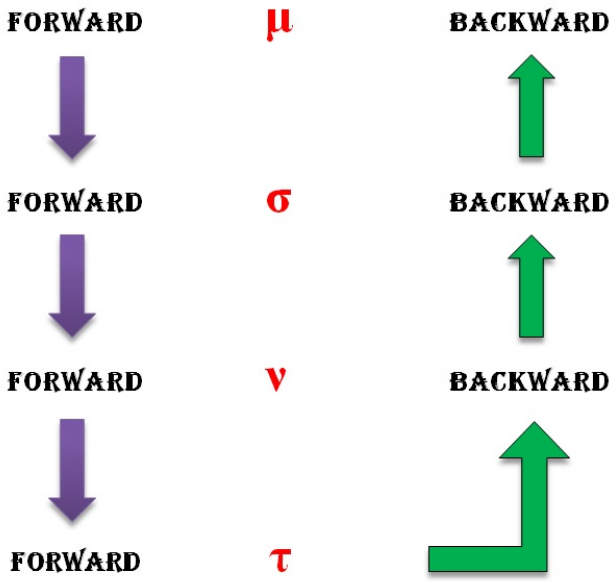
• **Função stepGAIC**(·)**:** Esta função pode ser utilizada para construir um modelo para qualquer um dos parâmetros de distribuição, fundamentado nos procedimentos *backward*, *forward* ou *stepwise*, utilizando o GAIC [63].

Procedimentos tais como, o método *forward*, o método *backward* e o método *stepwise* são aqueles que podem ser aplicados para cada parâmetro ou para todos os parâmetros da distribuição. Em conformidade com [43], o método de seleção *forward* inicia o processo com o modelo nulo, e gradativamente insere covariáveis significativas, segundo medidas do AIC. Já

o método de seleção *backward* realiza o procedimento de maneira inversa, iniciando-se um modelo com todas as covariáveis possíveis e, gradativamente remove as variáveis menos significativas. O método *stepwise* integra as estruturas dos métodos *forward* e *backward*, existindo no R a função **stepGAIC**(·) que possibilita seu uso, de maneira que o argumento **direction**(·) permite selecionar qual o método utilizado. Por *default* o método é o *backward*.

• **Função stepGAICAll.A(·):** A função **stepGAICAll.A**(·) integra a estrutura da função **stepGAIC**(·) para ambos os métodos (*forward* e *backward*) simultaneamente [63]. A estratégia adotada por este método, pode ser visualizada por meio da Figura 3.3. A seguir é apresentado este procedimento. Para mais detalhes, ver [63].

Figura 3.3: Estratégia A - Método de seleção de covariáveis *Stepwise*.



Fonte: Adaptado de [63].

- Inicia-se a seleção sem covariável em todos os parâmetros e, a partir daí, adiciona-se covariáveis para o parâmetro μ, sendo considerado constantes os demais parâmetros σ, ν e τ. A comparação dos modelos, com as covariáveis selecionadas para o parâmetro μ é realizada por meio do AIC.
- Quando obter o melhor modelo possível para o parâmetro μ, considera-se este modelo para este parâmetro e realiza-se o mesmo procedimento para o parâmetro σ, considerando ν e τ constantes.
- Este procedimento é repetido para os parâmetros ν e τ.
- Quando finalizado o procedimento *Forward* para o parâmetro τ, inicia-se o Método *Backward* neste parâmetro (retirar covariáveis), mantendo os demais parâmetros como constantes.

- O procedimento é repetido para os demais parâmetros, mantendo fixas as covariáveis selecionadas para os parâmetros em etapas anteriores.

O modelo final é obtido, contendo as covariáveis selecionadas para os preditores de cada parâmetro da distribuição. Ressalta-se que não se tem, necessariamente, as mesmas covariáveis para os preditores de cada parâmetro da distribuição, de maneira que, assim como nos modelos GAM's, estas estruturas podem ser compostas por funções suavizadoras.

Estas funções de suavização podem ser inseridas nos preditores dos parâmetros de distribuições por meio do pacote **gamlss** ou do pacote **mgcv**. A primeira alternativa apresentada possibilita o uso de diversos suavizadores, dentre os quais se destacam os *splines* cúbicos (**cs(**$\cdot$**), scs(**$\cdot$**)**), os p-*splines* (**pb(**$\cdot$**), ps(**$\cdot$**)**), os efeitos aleatórios (*random effects*) (**re(**$\cdot$**), random(**$\cdot$**)**), entre outros. Para a segunda opção (pacote **mgcv**), a inserção das funções é realizada por meio de uma interface da função **gam(**$\cdot$**)**, em que diversos suavizadores estão disponíveis e podem ser utilizados nos modelos GAMLSS [63].

Ainda segundo [63], algumas funções suavizadoras tais como, os *thin plate* (**s(**$\cdot$**)**) ou os produtos tensor (**ti(**$\cdot$**)** ou **te(**$\cdot$**)**) podem ser inseridas nos preditores dos parâmetros da distribuição do modelo, contudo, é necessário o uso da função **ga(**$\cdot$**)**, que é quem possibilita o acesso para o uso das funções suavizadoras do pacote **mgcv** no GAMLSS.

Sob uma perspectiva computacional, [63] afirma que a função **pb(**$\cdot$**)** do pacote **gamlss** e a função **s(**$\cdot$**)** do pacote **mgcv**, quando utilizadas em modelos de regressão, apresentam resultados similares no ajuste. Assim, neste trabalho, optou-se pela função **s(**$\cdot$**)** (*thin plate spline*) devido ao menor tempo de processamento no ajuste dos modelos implementados.

Seleção dos parâmetros de suavização do modelo (Hiperparâmetros) $\mathcal{L}$

Existem diversas metodologias para selecionar os parâmetros de suavização do modelo que, por sua vez, segundo [63] podem ser fixos ou estimados. Segundo [27], a forma de fixação dos hiperparâmetros pode ser realizada fixando o grau de liberdade. Outra possibilidade refere-se a escolha de diferentes graus de penalização e comparar esses elementos a partir do GAIC. Para além destas duas maneiras de seleção de hiperparâmetros, o *software* R possibilita sua estimação. Os três métodos tradicionais para estimar os parâmetros de suavização são: Validação cruzada generalizada (GCV), GAIC e Método de máxima verossimilhança.

Cada um dos três métodos supracitados pode ser realizado segundo duas maneiras distintas, a saber, o Método Local, que refere-se quando o método é aplicado dentro do algoritmo GAMLSS iterativo; ou o Método Global, que indica quando o método é aplicado fora do algoritmo GAMLSS iterativo. Em suas palavras, [63] infere que os Métodos Locais são mais rápidos e geralmente produzem resultados semelhantes aos métodos globais.

3.1.5 Análise diagnóstica de um modelo ajustado

Um aspecto de fundamental importância para a análise da qualidade e adequabilidade do modelo ajustado, refere-se ao estudo e a interpretação diagnóstica de seus resíduos [19]. Entende-se por resíduos ordinários como sendo as diferenças entre os valores observados por um modelo e seus valores estimados, isto é, $\text{Res}_i = y_i - \hat{y}_i$, com $i = 1, ..., n$. Diferentemente do modelo clássico que utiliza resíduos ordinários, estudentizados ou padronizados, e dos GLM's, que utiliza por exemplo, os resíduos de Pearson ou a componente da deviance como resíduo, em [63] observa-se que nos GAMLSS são considerados os resíduos quantílicos (aleatorizados) normalizados, que foram introduzidos por [19].

A grande vantagem em se utilizar os resíduos quantílicos, refere-se a possibilidade destes elementos sempre possuírem uma distribuição normal padronizada, quando o modelo estiver bem ajustado, independentemente da distribuição de probabilidade da variável resposta [63]. Isso não ocorre para os resíduos ordinários ou os de Pearson, devido à limitações impostas quanto a diversidade de parâmetros ajustados e, à natureza dos dados que podem apresentar curvas com elevado índice de curtose e assimetria.

Considerando que a função densidade $f(y; \boldsymbol{\theta})$ é ajustada às observações y_i, com $i = 1, 2, ..., n$, os resíduos quantílicos normalizados ajustados, em [19], são representados por $\hat{r}_i = \boldsymbol{\Phi}^{-1}(\hat{u}_i)$, em que $\boldsymbol{\Phi}^{-1}(\cdot)$ é a função quantílica da distribuição normal padronizada, e os $\hat{u}_i$ são resíduos quantílicos definidos diferentemente para variáveis resposta contínuas e discretas.

Dado que y seja uma observação de uma variável resposta contínua, e considere $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ a função de distribuição acumulada do modelo ajustado. Supondo que este modelo seja especificado corretamente, u tem distribuição uniforme no intervalo unitário. Neste sentido, o resíduo quantílico normalizado para um GAMLSS é dado em [63] por

$$r_i = \Phi^{-1}\{F(y_i|\hat{\boldsymbol{\theta}}_i)\}, \quad \text{com } \ i = 1, ..., n,$$

em que $\boldsymbol{\Phi}^{-1}(\cdot)$ é a função quantílica da distribuição normal padronizada, $F(y|\hat{\boldsymbol{\theta}})$ é a função de distribuição estimada e $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ são as estimativas dos parâmetros do modelo.

Para o caso em que y seja uma observação de uma variável resposta discreta, novamente se exige a tranformação em uma variável aleatória u, com distribuição uniforme no intervalo $(u_1, u_2) = [F(y-1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})]$. Assim, neste intervalo é selecionado aleatoriamente u_i e o resíduo quantílico normalizado é dado em [63] por

$$r_i = \Phi^{-1}(u_i).$$

Assim, os resíduos quantílicos normalizados oferecem um ferramental bastante eficiente para a análise de resíduos de um modelo. Outro aspecto de destaque quando se consideram os resíduos quantílicos, consiste na exploração gráfica oferecida para o ajuste do

modelo GAMLSS. Algumas funções implementadas no *software* R, por meio do pacote **gamlss**, auxiliam nesta observação e análise, dentre as quais se destacam **plot(·)** e **wp(·)**.

• **Função plot(·)**

Esta função pode ser utilizada desde que o seu primeiro argumento seja um modelo ajustado. A função **plot(·)** produz quatro gráficos com o objetivo de verificação dos resíduos quantílicos normalizados de um modelo GAMLSS ajustado [34]. Ainda segundo estes autores, os quatro gráficos apresentados são

- resíduos *versus* valores ajustados do parâmetro μ.
- resíduos *versus* uma variável explicativa especificada.
- uma estimativa da densidade dos resíduos.
- um gráfico QQ-normal dos resíduos.

Ressalta-se que se o modelo proposto apresentar um bom ajuste, então $r_i \sim \mathbf{N}(0, 1)$, apresentando simetria e achatamento similar a da curva da distribuição normal (mesocúrtica).

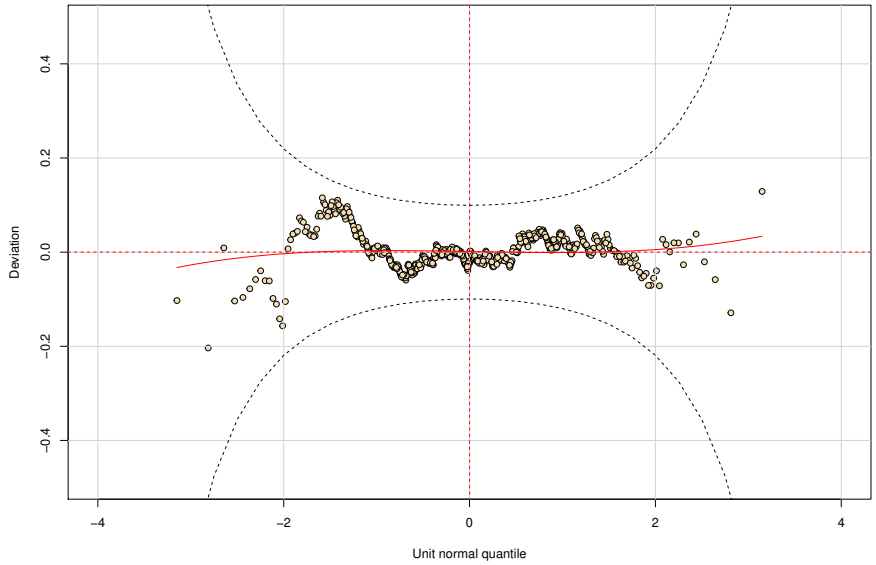
• **Função Worm plot(·)**

Uma das principais técnicas gráficas para diagnóstico de ajuste do modelo, utilizando os resíduos dos quantis aleatórios normalizados, foi introduzida por [7], e é denominada de gráfico *worm plot* (gráfico de minhoca). Os *worm plots* dos resíduos foram implementados visando identificar regiões de uma variável explicativa dentro das quais o modelo não atende adequadamente os dados (chamado "violação do modelo") [63]. O gráfico de *worm* é um QQ-plot sem tendência e o nome vem da aparência de minhoca dos pontos plotados.

Para exemplificar os elementos deste gráfico, considere a Figura 3.4 que elucida um *worm plot* de um modelo GAMLSS ajustado do conjunto de dados *abdom*, utilizando a distribuição Box Cox t. Esta situação é proveniente da obra [63].

- Os pontos de cor dourada que se apresentam no gráfico, indicam o quão distantes os resíduos estão de seus valores esperados, representados pela reta pontilhada $y = 0$, na cor vermelha. Uma situação ideal seria quando todos os pontos dourados sobrepusessem a reta vermelha pontilhada.
- As curvas elípticas indicam regiões de confiança de 95%. Isso significa que, para o modelo implementado apresentar um bom ajuste, espera-se que aproximadamente 95% dos pontos dourados estarão entre as duas curvas elípticas e 5% nas regiões complementares. Qualquer valor acima destes 5% nas regiões complementares à região entre as curvas elípticas indica que o ajuste realizado não está adequado para a explicação da variável resposta.

Figura 3.4: *Worm plot* de um modelo GAMLSS com distribuição BCT.



Fonte: Adaptado de [63].

- A curva vermelha é um ajuste de um polinômio cúbico aos pontos do gráfico *worm plot*. Esse ajuste reflete diferentes inadequações no modelo implementado, podendo ser uma assimetria e/ou curtose existentes nos resíduos, bem como média e/ou variância mal ajustadas para estes mesmos elementos.

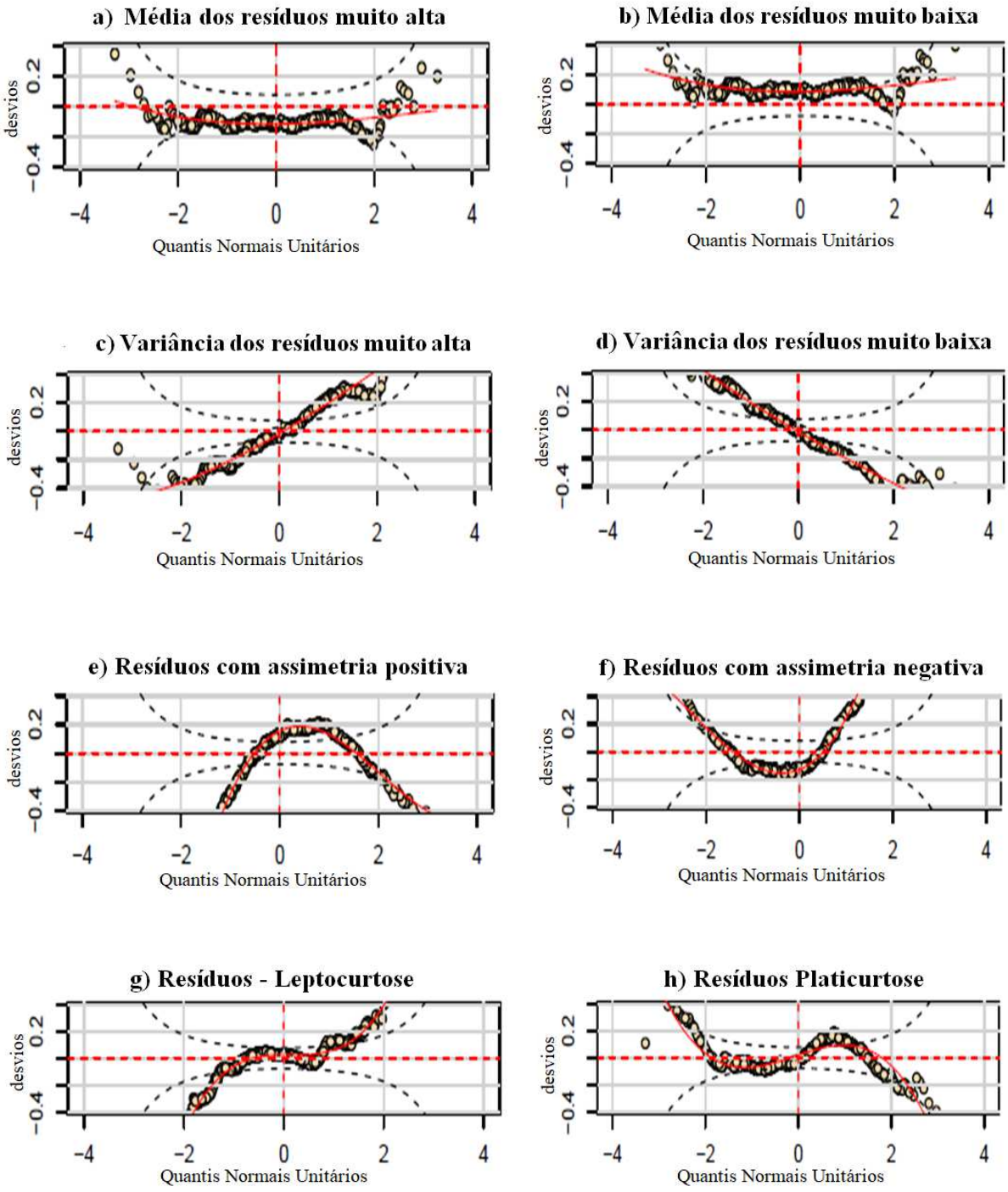
Isso posto, uma vez que todas as observações pertencem a região que se localiza entre as duas curvas elípticas e nenhuma forma específica é detectada nos pontos, o modelo pode ser considerado adequado. Para um análise mais detalhada, é apresentada a Figura 3.5, contendo oito padrões sistemáticos de afastamento dos resíduos quantílicos da linha horizontal de referência em um *worm plot*. A Tabela 3.5 apresenta as interpretações das curvas mostradas na Figura 3.5.

3.2 Recursos computacionais

Para o processamento computacional de ajuste de modelos, análises gráficas e obtenção de medidas visando a comparação e seleção de modelos, esta pesquisa utilizou o *software* R [52]. Este programa foi desenvolvido por Robert Gentleman e Ross Ihaka do Departamento de Estatística da Universidade de Auckland, de maneira que possui uma plataforma livre e oferece uma grande diversidade de técnicas estatísticas e de análises gráficas [33].

Além das facilidades e vantagens supracitadas, observa-se em [52] que a funcionalidade de seu ambiente oportuniza uma grande simplicidade para a manipulação de dados, cálculos computacionais envolvendo matrizes, ferramentas para análise, armazenamento e in-

Figura 3.5: Padrões sistemáticos dos resíduos quantílicos em um *worm plot*.



Fonte: Adaptado de [63].

terpretação de informações, aliada a uma linguagem de programação de fácil acesso.

Para esta pesquisa, o *software* R é de vital importância, já que foi utilizado um modelo GAMLSS objetivando a estimativa de desvios para o vermelho fotométricos de galáxias (redshift), segundo a fotometria de diferentes comprimentos de ondas (bandas). Logo, por meio dos pacotes **gamlss** e **CosmoPhotoz**, ambos implementados neste *software*, é possível

Tabela 3.5: Diferentes formatos do gráfico *worm plot* e suas interpretações.

Casos	**Formato do gráfico worm plot**	**Resíduos**	**Diagnóstico do modelo ajustado**
Caso a)	Nível: acima da origem	Média muito alta	Locação ajustada baixa
Caso b)	Nível: abaixo da origem	Média muito baixa	Locação ajustada alta
Caso c)	Reta: inclinação positiva	Variância muito alta	Escala ajustada muito baixa
Caso d)	Reta: inclinação negativa	Variância muito baixo	Escala ajustada muito alta
Caso e)	Parábola: concavidade voltada para cima	Assimetria positiva	Coeficiente de assimetria ajustado muito baixo
Caso f)	Parábola: concavidade voltada para baixo	Assimetria negativa	Coeficiente de assimetria ajustado muito alto
Caso g)	Em $\mathbb{S}$ com extremidade esquerda para baixo	Leptocurtose	Caudas da distribuição ajustada muito leves
Caso h)	Em $\mathbb{S}$ com extremidade esquerda para cima	Platicurtose	Caudas da distribuição ajustada muito pesadas

Fonte: Adaptado de [63].

selecionar distribuições, verificar seus ajustes, identificar possíveis inadequações buscando-se um melhor ajuste possível.

3.3 DESCRIÇÃO DO CONJUNTO DE DADOS

Considerando o objetivo da pesquisa, bem como a busca pela apresentação de um modelo com dados que sejam acessíveis e publicamente disponíveis, foi adotado como base de dados o *PHoto-z Accuracy Testing* (PHAT). O PHAT foi uma iniciativa internacional, idealizada para a busca/comparação de métodos promissores envolvendo *redshift* fotométrico, buscando melhorias em metodologias para medição de galáxias [29].

A estrutura do PHAT apresenta distintos ambientes, dentre os quais se destacam aqueles referentes a testes padronizados para pesquisadores da área, com catálogos fotométricos simulados (PHAT0) ou observados (PHAT1), contendo materiais explicativos, com Modelos de Distribuição de Energia Espectral, curvas de filtro, entre outros [29].

Observando-se para os catálogos fotométricos disponibilizados pelo PHAT, em [29] tem-se que o PHAT0 é fundamentado em uma simulação altamente idealizada representando um caso fácil para testar os elementos mais básicos da estimativa de *photoz*. Já o PHAT1 é pautado em dados reais oriundos do *Great Observatories Origins Deep Survey* [25], e que segundo [20], seus dados não são abertos publicamente. Assim sendo, o catálogo PHAT0 apresenta um total de 169.520 galáxias simuladas com *redshift* variando de $z = 0,02$ à $z = 2,24$. Em conjunto aos dados mencionados, são apresentadas magnitudes em 11 filtros distintos, denominado por: u, g, r, i, z, Y, J, H, K, IRAC1 e IRAC2.

Para maior praticidade na visualização e exploração dos dados, foi possível utilizar o *sofware* R, em particular, o pacote **CosmoPhotoz**, que possui funções de modelagem

da variável *redshift* em função dos comprimentos de onda (11 filtros), segundo os GLM´s [20].

Um aspecto importante referente ao pacote **CosmoPhotoz**, é o fato de possuir o catálogo PHAT0 implementado em sua totalidade de observações, isto é, com as 169.520 galáxias simuladas, com seu *redshift* e os 11 filtros. Contudo, este catálogo foi divido em dois conjuntos de dados, à saber, o **PHAT0train** e o **PHAT0test**.

Ambos os conjuntos de dados pertencem ao pacote **CosmoPhotoz**, contendo informações de 12 variáveis (magnitudes de 11 filtros e o *redshift* fotométrico), sendo que o catálogo PHAT0train apresenta dados de 8.478 galáxias, enquanto que o PHAT0test possui de 161.042 galáxias.

Os Quadros 3.6 e 3.7 exibem as informações dos dados utilizados nesta pesquisa, bem como expõe as 12 variáveis com suas características, respectivamente.

Quadro 3.6: Banco de dados - pacote **CosmoPhotoz**.

Catálogos	**Variáveis**	**número de galáxias**
PHAT0train	redshift, u, g, r, i, z, Y, J H, K, IRAC1, IRAC2	8.478
PHAT0test	redshift, u, g, r, i, z, Y, J H, K, IRAC1, IRAC2	161.042
TOTAL (PHAT0)	redshift, u, g, r, i, z, Y, J H, K, IRAC1, IRAC2	169.520

Fonte: Próprio autor.

Como informado anteriormente, o pacote **CosmoPhotoz** foi desenvolvido visando a modelagem do *redshift* fotométrico em função dos 11 filtros, por meio dos GLM. Logo, a estrutura PHAT0train e PHAT0test é estabelecida como se verifica no Quadro 3.6, não podendo serem alterados a divisão dos dados para estes catálogos. Nesse sentido, em consonância com o objetivo e problemática desenvolvidos, nesta pesquisa construiu-se um banco de dados único, com as 169.520 galáxias.

Pode-se observar que o banco de dados **PHAT0** apresenta um número elevado de dados, contando com observações de 169.520 gálaxias. Sem perda de generalidade, uma alternativa adotada para esta pesquisa foi considerar uma amostra aleatória de 5% dos dados do conjunto **PHAT0** (total de 8.476 galáxias), para assim realizar as etapas de seleção da distribuição da variável resposta, das funções de ligação, dos preditores lineares dos parâmetros da distribuição considerada, e finalmente ajustar um modelo GAMLSS.

Visando assegurar uma aproximação das características entre a amostra aleatória considerada para a pesquisa, com a totalidade dos dados da base **PHAT0**, são apresentados os Quadros 3.8 e 3.9, que ilustram as medidas resumo de todas as variáveis envolvidas no ajuste.

Por meio dos quadros supracitados, pode-se observar que as medidas resumo da amostra aleatória de 5% dos dados considerados, apresentam uma proximidade bastante significativa com as medidas de resumo dos dados presentes na base **PHAT0**. Ressalta-se que

Quadro 3.7: Descrição das variáveis contidas no dados do pacote **CosmoPhotoz**.

Variáveis	**Descrição**	**Tipo de variável**	**Domínio**	**Intervalo**
redshift	O desvio para o vermelho da galáxia	Quantitativa contínua dependente	$\mathbb{R}_+$	0,02 - 2,24
u	A magnitude da galáxia na banda u	Quantitativa contínua independente	$\mathbb{R}_+$	15,37 - 28,94
g	A magnitude da galáxia na banda g	Quantitativa contínua independente	$\mathbb{R}_+$	14,88 - 26,60
r	A magnitude da galáxia na banda r	Quantitativa contínua independente	$\mathbb{R}_+$	14,70 - 24,14
i	A magnitude da galáxia na banda i	Quantitativa contínua independente	$\mathbb{R}_+$	14,46 - 24,18
z	A magnitude da galáxia na banda z	Quantitativa contínua independente	$\mathbb{R}_+$	14,29 - 23,97
Y	A magnitude da galáxia na banda Y	Quantitativa contínua independente	$\mathbb{R}_+$	14,12 - 23,92
J	A magnitude da galáxia na banda J	Quantitativa contínua independente	$\mathbb{R}_+$	14,07 - 23,74
H	A magnitude da galáxia na banda H	Quantitativa contínua independente	$\mathbb{R}_+$	13,84 - 23,70
K	A magnitude da galáxia na banda K	Quantitativa contínua independente	$\mathbb{R}_+$	14,03 - 23,89
IRAC1	A magnitude da galáxia na banda IRAC1	Quantitativa contínua independente	$\mathbb{R}_+$	14,80 - 26,37
IRAC2	A magnitude da galáxia na banda IRAC2	Quantitativa contínua independente	$\mathbb{R}_+$	15,22 - 26,37

Fonte: Próprio autor.

Quadro 3.8: Medidas resumo das 12 variáveis do conjunto de dados PHAT0.

Medidas	***redshift***	**up**	**gp**	**rp**	**ip**	**zp**	**Y**	**J**	**H**	**K**	**IRAC1**	**IRAC2**
Mínimo	0,02	15,37	14,88	14,70	14,46	14,29	14,12	14,07	13,84	14,03	14,80	15,22
1º Quartil	0,24	23,50	22,99	22,34	21,93	21,73	21,54	21,28	21,02	20,89	21,38	21,77
Mediana	0,40	24,16	23,74	23,13	22,71	22,53	22,39	22,18	22,02	21,95	22,78	23,27
Média	0,42	24,05	23,48	22,82	22,45	22,29	22,14	21,93	21,76	21,69	23,35	23,69
3º Quartil	0,58	24,64	24,20	23,61	23,25	23,11	23,00	22,84	22,72	22,69	26,37	26,37
Máximo	2,24	28,94	26,60	24,14	24,18	23,97	23,92	23,74	23,70	23,89	26,37	26,37
Desvio Padrão	0,25	1,18	1,10	1,06	1,07	1,10	1,14	1,18	1,23	1,27	2,44	2,33
Assimetria	0,76	-0,32	-1,43	-1,59	-1,41	-1,33	-1,27	-1,19	-1,02	-0,94	0,07	-0,08
Curtose	4,09	5,61	6,24	6,22	5,79	5,48	5,23	4,80	4,20	3,83	1,66	1,68

Fonte: Próprio autor.

para a amostra aleatória, a variável resposta *redshift* apresenta valor de desvio padrão $(0,25)$, e alto valor de curtose $(4,41)$, com assimetria positiva $(0,80)$. Além disso, seus valores mínimo e máximo são $0,02$ e $2,08$, respectivamente. Isso sugere, que a distribuição da variável resposta

Quadro 3.9: Medidas resumo das 12 variáveis da amostra aleatória.

Medidas	*redshift*	up	gp	rp	ip	zp	Y	J	H	K	IRAC1	IRAC2
Mínimo	0,02	16,29	15,83	15,14	14,73	14,50	14,24	14,07	13,84	14,03	14,80	15,22
1º Quartil	0,24	23,47	22,96	22,31	21,92	21,73	21,53	21,27	21,03	20,90	21,42	21,80
Mediana	0,40	24,14	23,72	23,11	22,69	22,52	22,39	22,18	22,01	21,95	22,80	23,32
Média	0,42	24,03	23,46	22,81	22,45	22,29	22,14	21,93	21,76	21,70	23,38	23,72
3º Quartil	0,58	24,63	24,19	23,60	23,25	23,12	23,00	22,85	22,72	22,69	26,37	26,37
Máximo	2,08	28,46	26,43	24,05	24,09	23,89	23,86	23,86	23,64	23,88	26,37	26,37
Desvio Padrão	0,25	1,16	1,09	1,05	1,07	1,10	1,14	1,17	1,22	1,26	2,44	2,32
Assimetria	0,80	-0,35	-1,38	-1,54	-1,38	-1,30	-1,25	-1,17	-1,02	-0,95	0,06	-0,10
Curtose	4,41	5,56	6,06	6,10	5,74	5,43	5,22	4,82	4,25	3,90	1,65	1,68

Fonte: Próprio autor.

deve apresentar suporte $\mathbb{R}_+$, com características que permitam explicar o elevado índice para o coeficiente de curtose e assimetria positiva (cauda mais pesada à direita na curva de distribuição de probabilidade).

Outro aspecto interessante, refere-se às pequenas diferenças obtidas entre as medidas das 12 variáveis do conjunto **PHAT0** e a amostra aleatória. Pode-se observar que para o desvio padrão, a maior diferença entre suas medidas ocorre para a variável **up**, e vale $0,02$. Além disso, o valor do desvio padrão da variável *redshift* é a mesma para os dois conjuntos $(0,25)$. Já para a assimetria, observa-se que a maior diferença de valores ocorre para as variáveis **gp** e **rp**, sendo tal discrepância $0,05$. Ressalta-se que somente a variável **H** apresenta o mesmo valor de assimetria $(-1,02)$ para os dois conjuntos. Além disso, a distância dos valores para a variável *redshift* fotométrico é de $0,04$.

Ademais, os valores de curtose dos dois conjuntos de dados apresentados nos Quadros 3.8 e 3.9, são os que possuem maiores diferenças entre as 12 variáveis. Pode-se observar que a variável *redshift* fotométrico é a que exibe maior diferença entre os valores $(0,32)$, enquanto que a variável **IRAC2** apresenta os mesmos valores $(1,68)$. Embora exista essa diferença entre os valores das medidas dos dois conjuntos, compreende-se que tais distâncias são pequenas e não acarretam em prejuízos no processo de modelagem do problema proposto. Nesse sentido, sem perda de generalidade, considerar-se-á a amostra aleatória para o ajuste de um modelo GAMLSS.

De posse das 8.476 observações (amostra aleatória), prosseguiu-se a investigação considerando um pressuposto fundamental para os modelos de regressão: a necessidade das covariáveis não apresentarem altos valores de correlação, isto é, que não sejam altamente correlacionadas [43], para assim, evitar o problema denominado de multicolinearidade. Segundo [22], multicolinearidade é a presença de um elevado nível de correlação entre as variáveis explicativas, de maneira que apresentem uma dependência linear entre si.

Várias são as consequências da presença da multicolinearidade entre as variáveis independentes. Em [30], infere-se que, dentre elas, destacam-se: as variâncias e cova-

riâncias das estimativas dos parâmetros do modelo serão bastantes altas, ou seja, estas estimativas podem apresentar erros elevados, tornando difícil de analisar a influência das variáveis explicativas no modelo; considerando a alta correlação entre as covariáveis, pode-se supor em eliminá-las do ajuste do modelo, acarretando em um grave erro, já que tais elementos podem ser extremamente importantes para a explicação da variável resposta.

Nesse sentido, identificar a multicolinearidade entre as covariáveis, perpassa pela identificação de seus coeficientes de correlação. Para tanto, considerando um conjunto de n pares (x_i, y_i) de observações, o coeficiente de correlação linear de Pearson, cl_{prs}, (calculado por *default* no R) entre duas variáveis X e Y, é dado em [44] por

$$cl_{prs} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{\frac{1}{2}}}.$$

Os valores admitidos para cl_{prs} variam no intervalo real de -1 a 1, isto é, $-1 \leq cl_{prs} \leq 1$, de maneira que quanto mais próximo de -1 ou 1, mais fortemente correlacionadas linearmente as variáveis estarão. Por outro lado, se o valor de cl_{prs} estiver próximo de zero, pode-se entender que as variáveis não são correlacionadas [44]. Em geral, quando se realizam experimentos com dados reais, as variáveis apresentam altos índices de correlação [44].

Visando-se identificar essa medida para os dados da pesquisa, o *software* R permite encontrar tais valores por meio da função **cor**$(\cdot)$. O Quadro 3.10 apresenta a matriz de correlação das variáveis contidas na base de dados desta pesquisa. Buscando-se auxiliar na visualização dos valores dos coeficientes de correlação, também é apresentada a Figura 3.6 indicando, por meio da intensidade das cores azul e vermelho, o relacionamento das covariáveis.

Considerando as medidas deste coeficiente, pode-se verificar diversas covariáveis que apresentam altos valores de correlação. Exemplifcando tal situação, o filtro K, com excessão aos filtros u e g, exibe elevados índices de relacionamento com os demais filtros, sendo todos os valores maiores do que 0,71, isto é, $cl_{prs} > 0,71$. Além disso, são observados diversos índices de correlação $cl_{prs} = 0,99$ entre as covariáveis, tais como $cl_{prs}(z, Y)$, $cl_{prs}(z, i)$, $cl_{prs}(J, H)$, entre outros. Nesse sentido, constata-se alto nível de correlação entre as covariáveis consideradas nesta pesquisa, acarretando no risco da multicolinearidade.

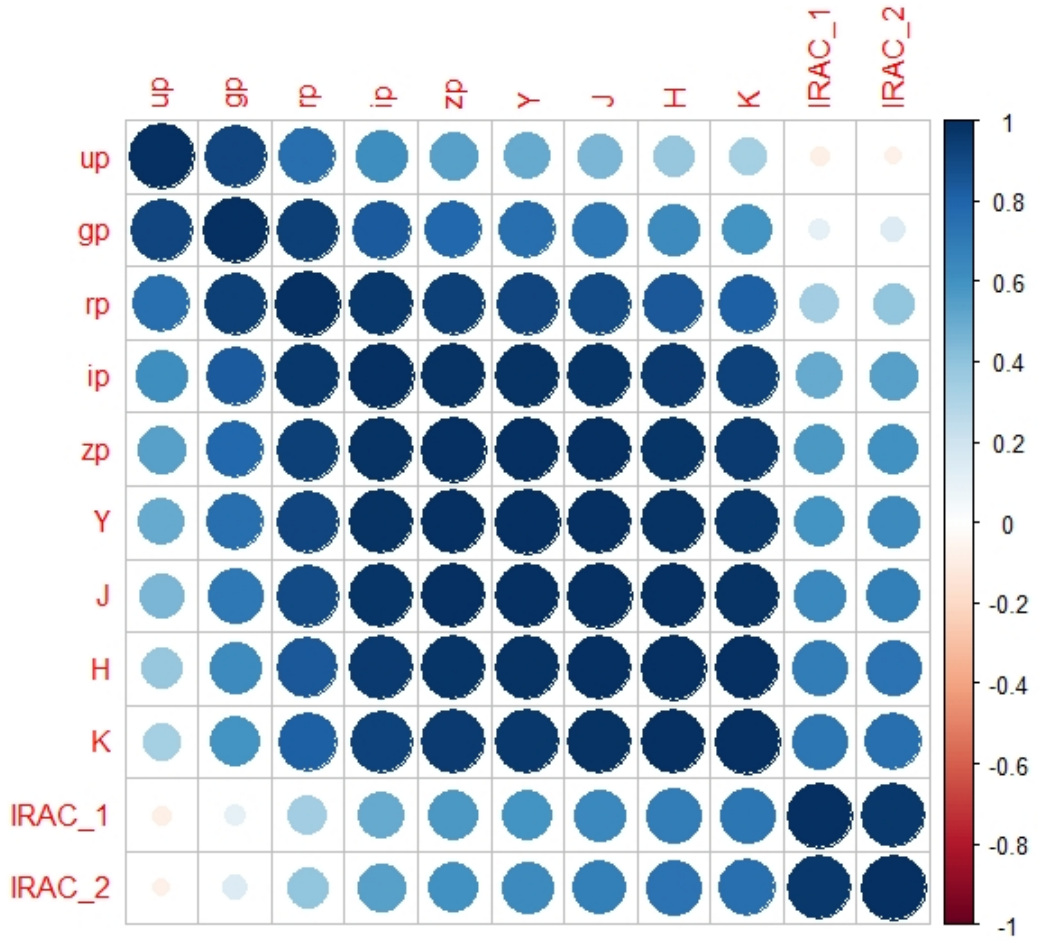
Assim sendo, existem algumas alternativas visando a superação desta situação. Dentre elas, e a que será adotada nesta pesquisa, tem-se a análise de componentes principais, que embora não possibilite uma fácil interpretabilidade do modelo ajustado frente as variáveis explicativas originais, permite considerar todas as covariáveis para o ajuste do modelo, sem eliminar elementos que podem ser essenciais.

Quadro 3.10: Coeficiente de correlação das covariáveis.

	u	g	r	i	z	Y	J	H	K	IRAC1	IRAC2
u	1,00	0,92	0,75	0,61	0,55	0,51	0,46	0,38	0,33	-0,08	-0,07
g	0,92	1,00	0,93	0,83	0,78	0,75	0,71	0,64	0,59	0,11	0,14
r	0,75	0,93	1,00	0,96	0,94	0,92	0,90	0,85	0,82	0,35	0,39
i	0,61	0,83	0,96	1,00	0,99	0,99	0,97	0,95	0,93	0,51	0,55
z	0,55	0,78	0,94	0,99	1,00	0,99	0,99	0,97	0,96	0,57	0,61
Y	0,51	0,75	0,92	0,99	0,99	1,00	0,99	0,97	0,96	0,57	0,61
J	0,46	0,71	0,90	0,97	0,99	0,99	1,00	0,99	0,98	0,65	0,68
H	0,38	0,64	0,85	0,95	0,97	0,98	0,99	1,00	0,99	0,70	0,73
K	0,33	0,59	0,82	0,93	0,96	0,97	0,98	0,99	1,00	0,72	0,75
IRAC1	-0,08	0,11	0,35	0,51	0,57	0,60	0,65	0,70	0,72	1,00	0,97
IRAC2	-0,07	0,14	0,39	0,55	0,61	0,64	0,68	0,73	0,75	0,97	1,00

Fonte: Próprio autor.

Figura 3.6: Gráfico indicando o coeficiente de correlação entre as variáveis



Fonte: Próprio autor.

A seguir, é apresentada uma discussão sobre a técnica de Análise de Componentes Principais considerada neste trabalho.

3.4 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

A Análise de Componentes Principais é uma das técnicas mais utilizadas dentro da estatística multivariada, uma vez que se busca pela redução da dimensão de um sistema, apresentando a menor perda possível da informação [31]. Embora esta técnica tenha sido construída por Pearson [51], foi somente com Hotelling [32] e Kendall [41] que observou sua aplicação em estruturas de regressão múltipla.

A técnica da PCA oferece uma grande vantagem para conjuntos de dados que apresentam medidas com muitas variáveis, já que é possível reduzir este número de variáveis, transformando-as em um subconjunto menor de características, chamadas de componentes principais [37]. Isso posto, segundo [31] a técnica PCA transforma um conjunto original de variáveis correlacionadas entre si, em um outro conjunto menor de variáveis (componentes principais), não correlacionadas entre si, de tal maneira que contenham, cumulativamente, toda a informação do conjunto original.

Segundo [47], as novas variáveis obtidas, que são denominadas de componentes principais, são combinações lineares das variáveis originais, construídas por meio da matriz de covariância ou correlação dos dados, de maneira que retenham, segundo a ordem em que são estimadas, o máximo de informação, em termos da variabilidade total do sistema [31].

Nos trabalhos de [42], [47] e [70], pode-se observar que dentre as diversas finalidades da PCA, seus principais objetivos consistem, primeiramente, na redução da dimensão da estrutura da modelagem, verificada por meio do menor número de componentes obtidos, e a determinação do índice de correlação entre as variáveis originais, evitando assim o problema da multicolinearidade.

No entanto, existem situações em que a técnica da PCA pode não colaborar com o processo de modelagem de um sistema. Contextos em que o número de variáveis é maior do que o número de observações, ou quando as variáveis originais são pouco correlacionadas, podem acarretar em perda de informações da variabilidade das variáveis originais (primeira situação), ou na obtenção de componentes principais iguais as variáveis originais (segundo caso) [31].

Suponha-se observados p covariáveis originais em um sistema de n observações. Logo, $\mathbf{X}$ é a matriz de dados de ordem $n \times p$. Além disso, considera-se que as p covariáveis não sejam independentes. Assim, é possível obter a matriz de covariância $\sum$ de ordem $n \times n$ [31].

Um problema que pode ocorrer, refere-se às unidades de medidas das covariáveis apresentarem diferentes escalas. Nesse sentido, em [53] verifica-se a conveniência na padronização destas medidas, em que se destacam a padronização com média zero e variância

um, bem como a padronização com variância um e média qualquer. Isso posto, após esta etapa, tem-se uma nova matriz $\mathbf{\Psi}$ de ordem $n \times p$, que segundo [65] é igual a matriz de correlação R dos variáveis originais.

A partir da matriz de correlação R, encontram-se os pares de autovalores e autovetores $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$, de tal maneira que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, ou seja, para cada autovalor λ_i, com $i = 1, \ldots, p$, existe uma autovetor e_i, dado por $e_i^T = [a_{i1} \;\; a_{i2} \;\; \ldots \;\; a_{ip}]$ [65].

Em seguida, os autovetores e_i são normalizados, ou seja, a soma dos quadrados de seus coeficientes deve ser igual a um. Isso significa que tais autovetores serão ortogonais entre si. Segundo [65], considerando λ_i o autovalor associado ao autovetor e_i, tem-se que o $i-$ésimo componente principal é

$$\text{PC}_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p. \tag{3.10}$$

É possível observar ainda em [65], que os componentes principais obtidos por meio da equação (3.10), apresentam algumas propriedades importantes para análise de seu comportamento, dentre as quais se destacam:

• Os componentes principais explicam de forma decrescente a variabilidade dos dados de um sistema, isto é, o primeiro componente principal é aquele que apresenta maior variabilidade, seguido pelo segundo componente, e assim sucessivamente. Ou seja,

$$\widehat{\text{Var}}(\text{PC}_1) > \widehat{\text{Var}}(\text{PC}_2) > \ldots > \widehat{\text{Var}}(\text{PC}_p).$$

• Os componentes principais obtidos não são correlacionados entre si, isto é,

$$\text{Cor}(\text{PC}_i, \text{PC}_j) = 0, \quad \text{com } \; i, \; j = 1, \ldots, p \; \text{ e com } i \neq j.$$

Esta é uma característica muito importante, pois tem-se novas covariáveis que não são correlacionadas entre si, proporcionando a superação do problema da multicolinearidade em modelos de regressão.

• A variância do componente principal PC_i é igual ao seu autovalor associado, isto é

$$\widehat{\text{Var}}(\text{PC}_i) = \lambda_i$$

• A variabilidade total observada nas variáveis originais do sistema é igual a variabilidade total explicada nos componentes principais [31], isto é

$$\sum_{i=1}^{p} \widehat{\text{Var}}(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \widehat{\text{Var}}(\text{PC}_i)$$

Ressalta-se que a contribuição de cada componente principal no sistema, pode

ser obtida por meio do quociente entre a variabilidade explicada individualmente por cada componente principal e variabilidade total do sistema.

Pode-se verificar em [36], que por meio da explicação da variabilidade individual de cada componente principal, é possível determinar o número de componentes a ser considerado na estrutura de modelagem. Embora não se tenha um critério bem estabelecido na literatura, segundo [36], muitas pesquisas adotam componentes que expliquem cumulativamente, 80% da variabilidade total do sistema.

Em [31], tem-se que outro critério bastante utilizado é o Critério de Kaiser [40]. Neste trabalho, infere-se que a escolha dos componentes principais a serem consideradas no sistema, está relacionada com o valor observado em seus autovalores. De maneira mais específica, consideram-se as componentes tais que seus valores próprios sejam maiores do que um $(\lambda_i > 1)$, descartando os componentes que apresentem as demais possibilidades de valores para os seus autovalores.

Outro critério utilizado é observado em [38], na qual consideram-se os componentes principais que possuem valores acumulados de variância superiores a 70% da variância total. Outra alternativa é apresentada em [47], em que se considera gráficos de contribuição acumulada dos componentes principais. Segundo este autor, a seleção do número de componentes para modelagem de um sistema fundamenta-se na inclinação do gráfico, de maneira que se observa que a partir de uma determinada componente, a curva de variabilidade acumulada explicada passa a ter inclinação aproximadamente constante e pequena. Nessa situação, consideram-se as componentes que corresponde a região de maior inclinação da curva (à esquerda).

Em síntese, embora não se verifique um critério unificado quanto ao número de componentes a serem consideradas em um sistema, diversas metodologias podem ser utilizadas. Fato é, que a técnica de PCA é uma ferramenta bastante robusta para modelagem de dados que apresentem altos valores de correlação, além de possibilitar a redução da dimensão dos dados para a modelagem de um sistema proposto [37].

4 RESULTADOS E DISCUSSÃO

Para uma compreensão pormenorizada dos resultados desta pesquisa, inicialmente são apresentadas informações referentes à construção das componentes principais, à partir do conjunto de dados considerado (11 magnitudes dos filtros). Em seguida, exibe-se a investigação realizada objetivando-se a seleção, o ajuste e o diagnóstico do modelo GAMLSS. Para tanto, explanações referentes a análise descritiva das variáveis que compõem o modelo, além da análise de seu potencial de estimação serão elucidadas.

4.1 COVARIÁVEIS - COMPONENTES PRINCIPAIS OBTIDAS (PC)

Como já discutido na seção 3.4, a técnica de análise de componentes principais tem por objetivo a redução da dimensionalidade de observações multivariadas fundamentado em sua estrutura de dependência [44]. Isso significa que, a partir de variáveis que sejam altamente dependentes, são construídas componentes principais (PC) (combinações lineares das p variáveis), de maneira que sejam independentes entre si, apresentando uma estimação em ordem das PC's, contendo o máximo de informação em relação a variação total dos dados [35].

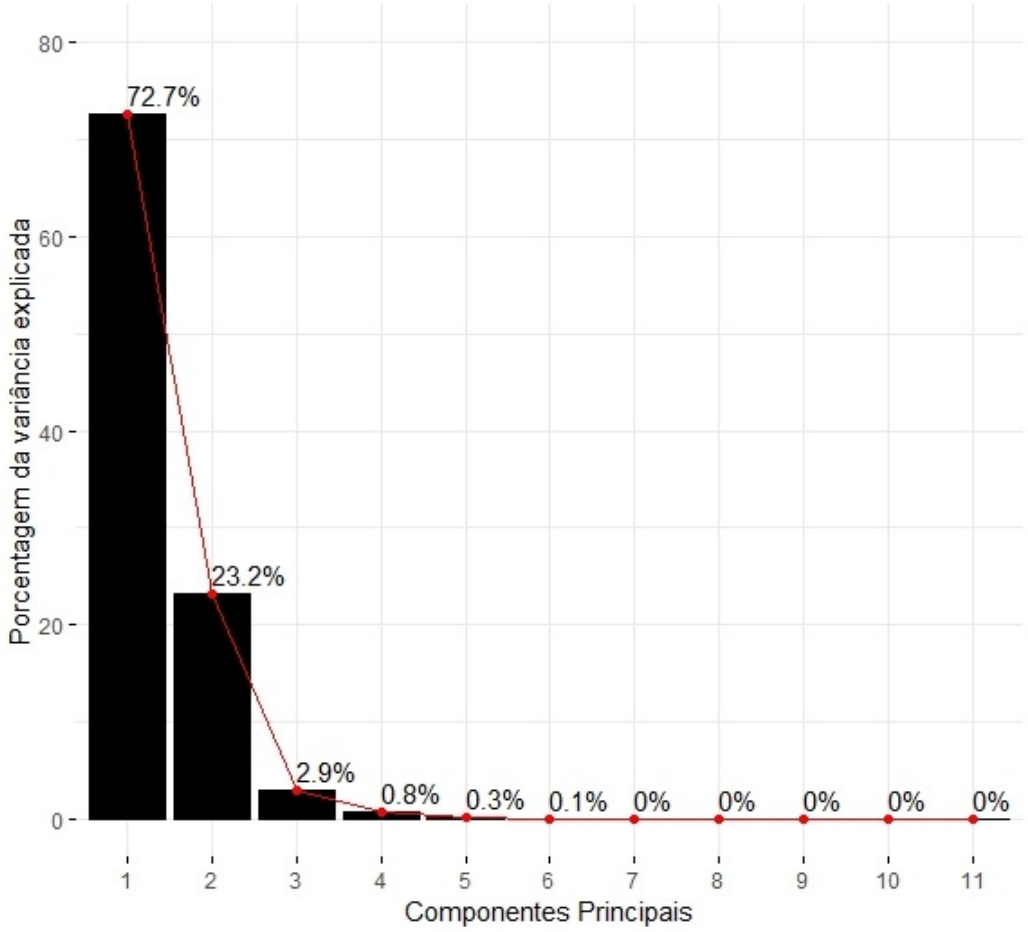
Este é um aspecto bastante importante, uma vez que as componentes principais podem substituir as variáveis originais, sendo utilizadas como novas variáveis explicativas de um modelo de regressão.

Na prática, uma questão natural a se considerar, refere-se à quantidade de componentes principais a serem tomadas para o ajuste de um modelo de regressão. Para tanto, tal escolha fundamenta-se na contribuição que cada componente principal apresenta, para uma proporção da variância total explicada. Embora não exista uma unanimidade nos modelos estatísticos para tal decisão, foi considerado o proposto na obra de [36], que sugere a escolha de um número de componentes principais, em que se explique, cumulativamente, 80% da variabilidade total do sistema.

Assim sendo, por meio da técnica de PCA, foram obtidas 11 componentes principais, a partir da combinação linear dos 11 filtros observados na amostra aleatória de 5% do banco de dados PHAT0. Para uma melhor visualização dos resultados, são apresentadas a Figura 4.1 e o Quadro 4.1, que exibem como e quanto cada PC contribui no sistema.

Pode-se observar que a primeira componente explica $72,65\%$ da variância total dos dados originais. Isso indica que, aproximadamente três quartos das informações presentes nos 11 filtros estão contidas em apenas uma única componente principal. Mais ainda, como a segunda componente explica $23,19\%$ da variância total, tem-se que com somente duas componentes principais, $95,84\%$ da variância dos dados é explicada. As demais 9 componentes explicam cumulativamente $4,16\%$. Portanto, de acordo com [36], foi escolhido o número de 2

Figura 4.1: Contribuição de cada PC na variância do sistema.



Fonte: Próprio autor.

Quadro 4.1: Proporção de Variância explicada pelos componentes principais.

	Componentes principais										
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**
Variância (autovalores)	16,7633	5.3500	0,6764	0,1829	0,0602	0,0150	0,0109	0,0057	0,0042	0,0027	0,0020
Proporção da Variância	0,7265	0,2319	0,02932	0,00793	0,00261	0,00065	0,00047	0,00025	0,00018	0,00012	0,00009
Proporção Acumulada	0,7265	0,9584	0,98771	0,99564	0,99824	0,99889	0,99937	0,99962	0,99980	0,99991	1,0000

Fonte: Próprio autor.

componentes principais para o processo de modelagem do GAMLSS.

Buscando por outras indicações nos referenciais teóricos que ateste o número de componentes adotadas nesta pesquisa, pode-se observar por meio do Critério de Kaiser [40], que somente as duas primeiras componentes principais apresentam seus autovalores maiores do que 1, isto é, $16,7633$ para a primeira componente e $5,3500$ para a segunda componente. Assim sendo, tem-se outro indício de que com duas componentes principais, a variância dos dados originais é contemplada satisfatoriamente para a implementação de um modelo GAMLSS.

Por meio da função **get_pca_var(·)** implementada no *software* R, é possível obter a contribuição de cada um dos 11 filtros para as duas componentes principais. Nesse sentido, o Quadro 4.2 e a Figura 4.2 exibem essa contribuição (em porcentagem), para a PC1 e a PC2.

Quadro 4.2: Contribuição dos 11 filtros nas duas componentes principais.

	PC1	PC2
up	0,45%	17,59%
gp	1,62%	15,41%
rp	3,31%	9,98%
ip	4,66%	6,40%
zp	5,42%	5,14%
Y	6,00%	4,62%
J	6,85%	3,68%
H	7,75%	2,41%
K	8,34%	1,78%
IRAC_1	28,66%	18,95%
IRAC_2	26,94%	14,05%

Fonte: Próprio autor.

Pode-se observar que os filtros IRAC_1 e IRAC_2 são os mais importantes para a PC1, de maneira que contribuem $28,66\%$ e $26,94\%$, respectivamente nesta componente. A covariável de menor importância é o filtro up, de maneira que sua contribuição para esta componente é de $0,45\%$.

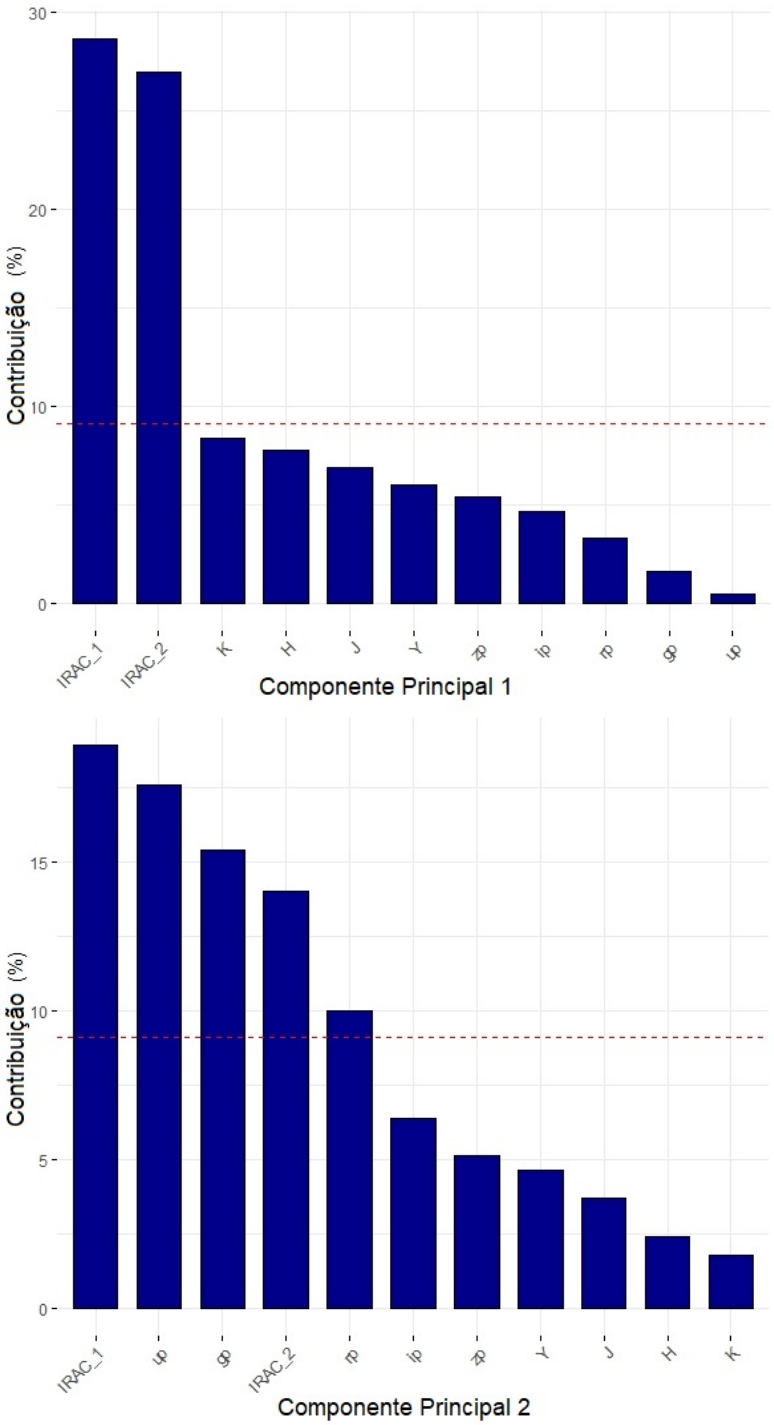
No que se refere a segunda componente, pode-se observar que as duas covariáveis que são mais importantes são os filtros IRAC_1 e up, contribuindo $18,95\%$ e $17,59\%$ para a PC2, respectivamente. Além disso, o filtro K é o menos importante, contribuindo $1,78\%$ para esta componente.

Outro aspecto interessante, refere-se disparidade observada nas contribuições destas duas componentes. Observa-se que na PC1, somente com os filtros IRAC_1 e IRAC_2, cumulativamente, a contribuição é de $55,60\%$, isto é, duas covariáveis explicam mais da metade desta componente. Isso é corroborado pelo gráfico apresentado na Figura 4.2, em que tem-se uma grande diferença na amplitude das barras que representam a contribuição destes dois filtros frente às demais covariáveis.

Essa característica não é observada para a segunda componente, uma vez que nota-se um comportamento bastante equilibrado entre as contribuições. O gráfico apresentado na Figura 4.2 ilustra tal situação, em que se tem uma queda gradativa das contribuições dos filtros, não ocorrendo uma diminuição abrupta entre esses elementos.

Nesse sentido, considerando as duas componentes principais como covariáveis do modelo a ser ajustado, bem como a variável resposta *redshift* fotométrico, a seção a

Figura 4.2: Contribuição dos 11 filtros nas duas PC´s.



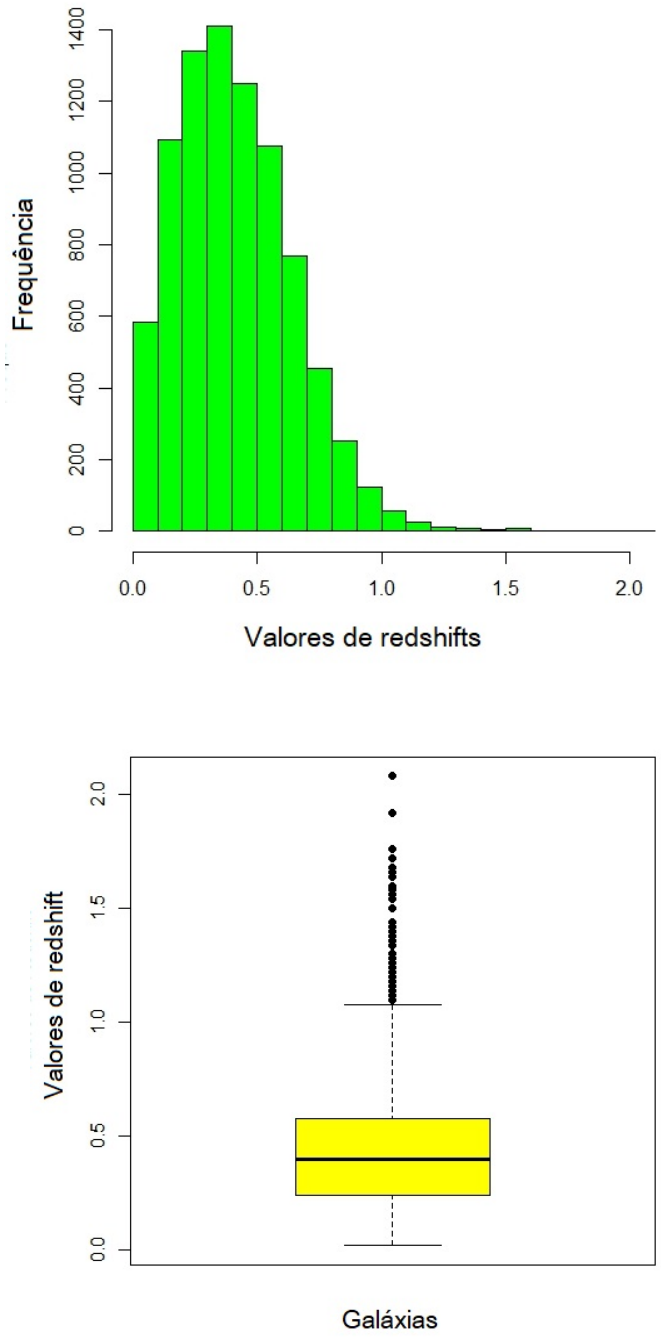
Fonte: Próprio autor.

seguir apresenta a análise descritiva destes elementos.

4.2 Análise Descritiva dos Dados

Como já informado na seção 4.1, a variável resposta é o *redshift* fotométrico das galáxias observadas. Buscando-se compreender características desta variável, é apresentada a Figura 4.3 contendo os seus gráficos de histograma e *box plot*. Para além disso, a Tabela 4.3 exibe suas principais medidas de posição e dispersão.

Figura 4.3: Histograma e *Box plot* da variável *redshift* fotométrico.



Fonte: Próprio autor.

Tabela 4.3: Resumo das medidas do *redshift* fotométrico.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Coeficiente de assimetria	Coeficiente de curtose	Coeficiente de variação
0,0200	0,2400	0,4000	0,4218	0,5800	2,0800	0,8030	4,4145	0,0553

Fonte: Próprio autor.

Ao analisar o histograma da variável resposta *redshift* fotométrico, nota-se que seus valores variam entre 0 (zero) e aproximadamente 2 (dois), com uma frequência maior para valores na faixa de $0,3$ à $0,5$. Tem-se uma diminuição significativa na frequência de valores de *redshift* a partir de $0,5$, indicando uma concentração maior de valores no início da distribuição. Nesse sentido, observa-se a existência de um histograma com assimetria positiva apresentando um único pico (unimodal), que ocorre entre valores de *redshift* sendo $0,3$ à $0,4$. Ademais, tem-se alguns valores discrepantes de *redshift* (maiores do que $1,1$), apontando para a existência de *outliers*.

Corroborando com as informações supracitadas, o *box plot* da variável resposta indica que 50% dos valores de *redshift* se encontram entre $0,3$ e $0,6$ aproximadamente (primeiro e terceiro quartis). Embora o valor de sua mediana (que é $0,40$) esteja próximo ao valor da média entre primeiro e terceiro quartis, indicando baixa variabilidade dos dados, observa-se uma grande quantidade de *outliers* cujos valores são maiores que o máximo observado nos dados (valores maiores do que $1,1$, aproximadamente).

Para explicar o comportamento do *redshift* fotométrico, foram consideradas duas componentes principais, uma vez que as onze variáveis explicativas (onze filtros de bandas) apresentaram altos índices de correlação. A Tabela 4.4 e as Figuras 4.4 e 4.5 exibem um resumo das medidas das duas componentes principais incorporadas aos preditores dos quatro parâmetros distribucionais do modelo GAMLSS ajustado, bem como seus gráficos de histogramas e *box plot*.

Tabela 4.4: Resumo das medidas das duas componentes principais.

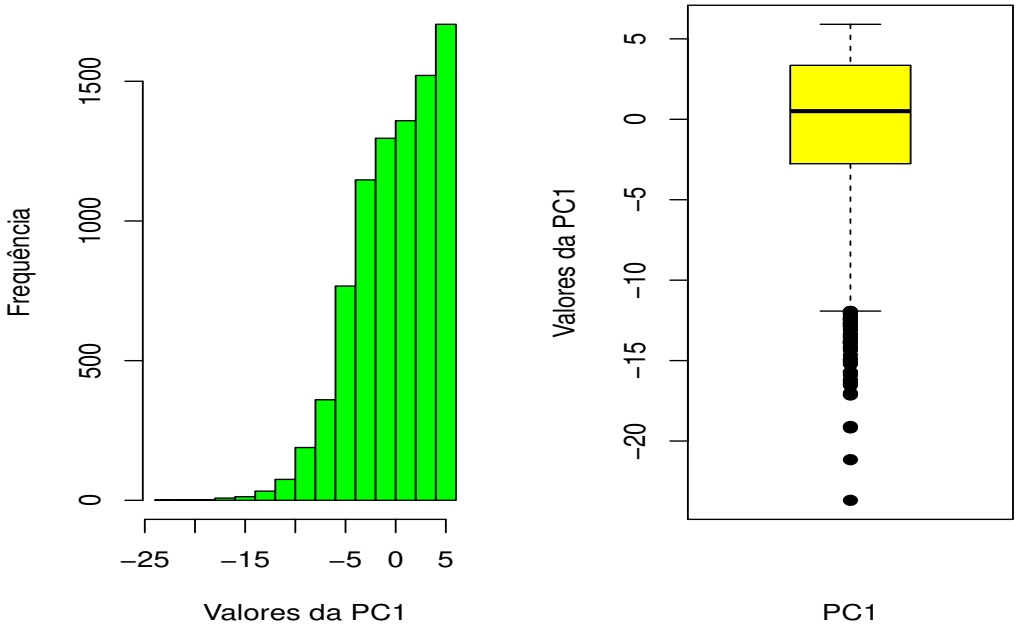
Variáveis	Valor Mínimo	1º Quartil	Mediana	Valor Médio	3º Quartil	Valor Máximo	Coef. de Assimetria	Coef. de Variação
PC1	-23,69	-2,76	0,51	0,00	3,35	5,91	-0,72	16,76
PC2	-4,8511	-1,77	-0,25	0,00	1,23	18,60	1,07	5,35

Fonte: Próprio autor.

Por meio da Tabela 4.4 e das Figuras 4.4 e 4.5, pode-se analisar à variabilidade dos dados das duas componentes principais, de maneira que a melhor medida a se observar é o coeficiente de variação. Essa medida indica a variabilidade dos dados em relação à sua média.

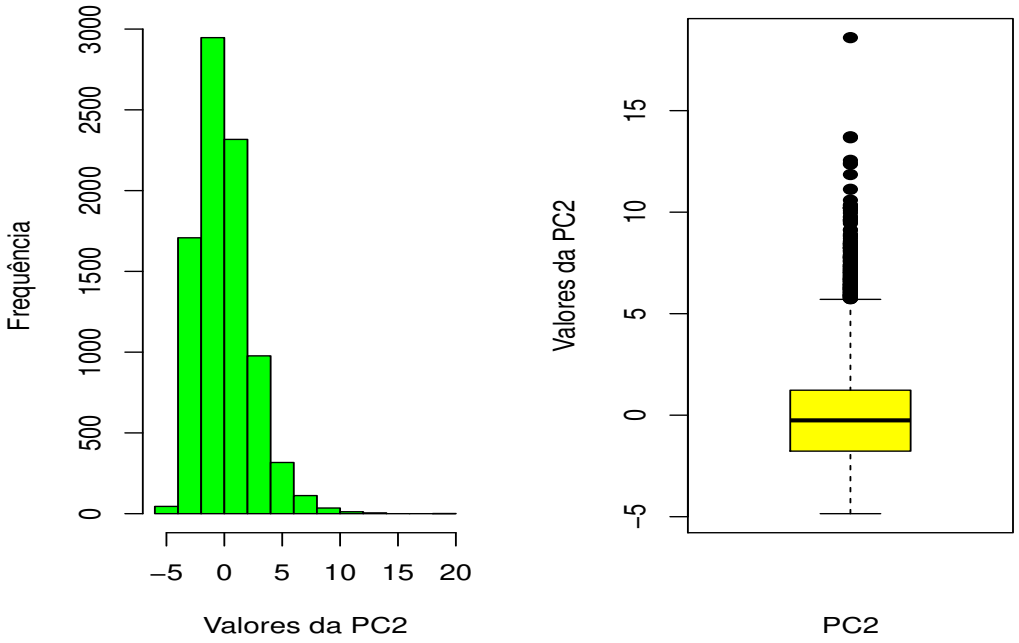
Note que ambas as componentes apresentam altos índices para o coeficiente de variação, de maneira que a PC1 possui um valor (16,76) que é aproximadamente o triplo

Figura 4.4: Histograma e *Box plot* da PC1.



Fonte: Próprio autor.

Figura 4.5: Histograma e *Box plot* da PC2.



Fonte: Próprio autor.

da PC2 (5,35). Logo, a primeira componente é a que apresenta maior variação entre seu valor mínimo e sua média. Além disso, nota-se que a PC1 é assimétrica negativa (coeficiente de assimetria menor que zero), indicando uma distribuição de frequências com uma cauda mais pesada à esquerda; enquanto que a PC2 é assimétrica positiva (coeficiente de assimetria maior que zero), indicando uma distribuição de frequências com uma cauda mais pesada à direita.

Outro aspecto interessante refere-se aos gráficos de *box plot*, em que se verifica uma simetria dos valores dos escores das componentes principais. Além disso, é possível observar a existência de uma grande quantidade de *outliers*, de maneira que para a PC1, estes extremos ocorrem para valores menores do que $-12,0$ (aproximadamente), e para a PC2 para valores maiores do que $6,0$ (aproximadamente).

A partir da análise descritiva realizada, dos materiais e métodos exibidos no Capítulo 3, será apresentada na seção seguinte, o processo de seleção da distribuição de probabilidade da variável resposta *redshift* fotométrico, bem como de que maneira cada uma das funções de suavização se evidenciam. Este fato é de vital importância, uma vez que se obtém novos indícios para a inserção destes elementos na estrutura do modelo GAMLSS proposto.

4.3 SELEÇÃO DO MODELO

4.3.1 Distribuição da variável resposta e funções de ligação

Como apresentado no Quadro 3.7, a variável resposta *redshift* é de natureza quantitativa contínua, sendo seu suporte o conjunto dos número reais positivos. Nesse sentido, para o processo de seleção da distribuição de probabilidade $(\mathcal{D})$ da variável dependente, algumas restrições são observadas.

Visando o cumprimento destas limitações quanto à escolha das distribuições, algumas funções do *software* R possibilitam identificar aquelas que melhor se ajustam aos dados da variável resposta. Nessa pesquisa, utilizou-se as funções **chooseDist(**$\cdot$**)** e **fitDist(**$\cdot$**)**, que pertencem ao pacote **gamlss**. Essas funções permitem o ajuste de distribuições paramétricas que pertencem a este pacote, de maneira que a distribuição marginal final pode ser escolhida por meio do GAIC com penalidade κ. O *default* do *software* R é $\kappa = 2$, ou seja, o AIC [61].

Embora ambas as funções permitam o ajuste de distribuições, elas possuem distinções segundo a abordagem realizada. Enquanto a função **fitDist(**$\cdot$**)** realiza o ajuste de distribuições paramétricas considerando unicamente a variável resposta, a função **chooseDist(**$\cdot$**)** procede a partir de um modelo GAMLSS já ajustado, considerando também as covariáveis indicadas no modelo [61].

Desta maneira, primeiramente utilizando-se a função **fitDist(**$\cdot$**)** com especificações particulares quanto a natureza das distribuições possíveis (distribuição contínua e utilizando o argumento *realplus* devido ao suporte da variável resposta), foram elencadas as três distribuições com melhores valores possíveis de AIC. A Tabela 4.5 apresenta a lista destas dis-

tribuições, bem como os seus números de parâmetros e os valores da estatística AIC.

Tabela 4.5: Distribuições apresentadas no comando **fitDist(·)**.

Distribuição	**AIC**	**Número de parâmetros**
Beta Generalizada - Tipo 2 (GB2)	-1661,594	Quatro
Box-Cox t (BCTo)	-1661,091	Quatro
Gama Generalizada (GG)	-1604,091	Três

Fonte: Próprio autor.

Após isso, ajustando-se um modelo GAMLSS, considerando a distribuição GB2 com uma única componente principal para cada preditor dos quatro parâmetros distribucionais do modelo, visando unicamente a utlização da função **chooseDist(·)**, foram obtidas as distribuições com melhores AIC. A Tabela 4.6 indicam estes resultados.

Tabela 4.6: Distribuições apresentadas por meio da função **chooseDist(·)**.

Distribuição	**AIC**	**Número de parâmetros**
Box-Cox t (BCTo)	-6381,734	Quatro
Box-Cox t (BCT)	-6376,765	Quatro
Beta Generalizada - Tipo 2 (GB2)	-6260.550	Quatro

Fonte: Próprio autor.

Embora a distribuição GB2 tenha sido selecionada como a melhor distribuição por meio da função **fitDist(·)**, apresentando o melhor valor de AIC $(-1661,594)$, quando utilizada a função **chooseDist(·)**, a distribuição BCTo foi a que se mostrou com melhor valor de AIC $(-6381,734)$.

Como já discutido na seção $3.1.4$, a distribuição contínua BCTo, permite a estimação de quatro parâmetros, à saber, os parâmetros de locação e dispersão (μ) e (σ), o de assimetria (ν) e o de curtose (τ) (os dois últimos relativos à forma), sendo que as funções de ligação consideradas para os preditores dos parâmetros desta distribuição, são dadas (*default*) por

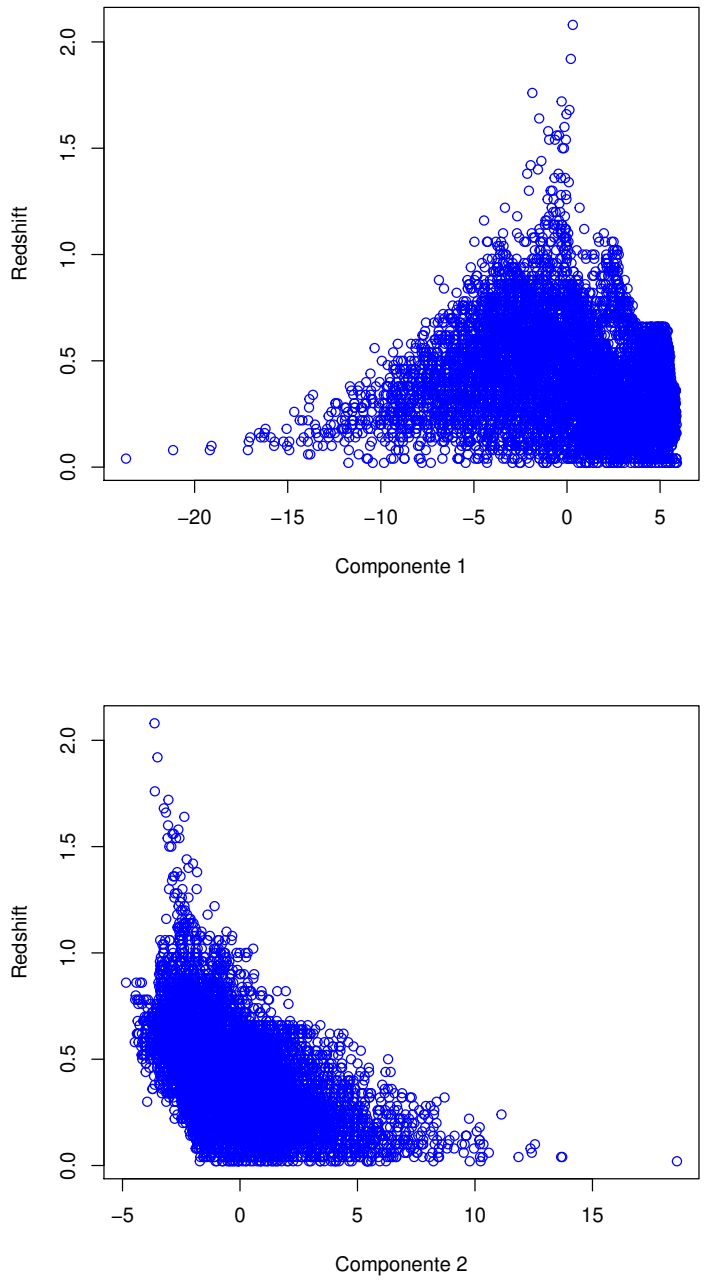
$$\begin{aligned}
\boldsymbol{\eta}_1 &= \mathbf{g}_1(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}),\\
\boldsymbol{\eta}_2 &= \mathbf{g}_2(\boldsymbol{\sigma}) = \log(\boldsymbol{\sigma}),\\
\boldsymbol{\eta}_3 &= \mathbf{g}_3(\boldsymbol{\nu}) = (\boldsymbol{\nu}),\\
\boldsymbol{\eta}_4 &= \mathbf{g}_4(\boldsymbol{\tau}) = \log(\boldsymbol{\tau}),
\end{aligned} \tag{4.1}$$

4.3.2 Termos utilizados nos preditores dos parâmetros da distribuição

Na seção 4.1 (Análise de Componentes Principais) foi possível observar que tomando-se duas componentes principais, tem-se uma explicação de $95,84\%$ da variabilidade

dos dados para a estrutura proposta. No entanto, não se sabe qual e como se dá o relacionamento de cada PC com a variável resposta *redshift*. Buscando-se analisar a questão supracitada, é apresentada a Figura 4.6, que traz luz aos gráficos de dispersão de cada componente principal *versus* a variável resposta.

Figura 4.6: Gráficos de dispersão das componentes principais *versus redshift*.



Fonte: Próprio autor.

É possível observar que as covariáveis (Componentes Principais) não se relacionam linearmente com a variável reposta *redshift*. Mais ainda, os gráficos de dispersão

apresentados na Figura 4.6, indicam que uma possibidade para explicar cada covariável em função da variável resposta, é realizar um ajuste por meio de gráfico de funções polinomiais (curvas). Assim sendo, considerando os aspectos supracitados, bem como a natureza das covariáveis (que são variáveis quantitativas contínuas), optou-se para o ajuste dos preditores dos parâmetros distribucionais, o uso de funções de suavização ou *splines*.

A partir disso, foi ajustado um modelo considerando a distribuição BCTo que apresenta quatro parâmetros ($q = 1, \ldots, 4$) sendo que todos os preditores apresentavam as mesmas especificações, isto é,

$$\boldsymbol{\eta}_q = g_q(\boldsymbol{\theta}) = \hat{\beta}_{q0} + \sum_{i=1}^{2} s_{qi}(PC_i), \text{ com } i = 1, 2.$$

No entanto, pode-se observar que este modelo apresenta as duas covariáveis (componentes principais) associadas às funções de suavização, para os quatro parâmetros distribucionais. Visando avaliar a necessidade de todos os elementos para os preditores dos parâmetros, foram utilizadas duas função do pacote **gamlss**, que possibilita selecionar as covariáveis para o ajuste do modelo, por meio do critério GAIC [61], à saber, as funções **stepGAICAll.A(·)** e **addterm(·)**. Ressalta-se que foram utilizados os dois procedimentos, afim de averiguar a existência de similaridades nas respostas obtidas para os preditores dos parâmetros da distribuição BCTo.

Após esta análise, verificou-se em ambos os procedimentos supracitados, que para os preditores dos parâmetros μ e σ da distribuição BCTo, as duas componentes suavizadas deveriam ser inseridas. Para o preditor do parâmetro ν foi indicado para não serem inseridas componentes e, por fim, para o preditor do parâmetro τ inferiu-se a necessidade da inserção de somente a primeira componente suavizada. Ou seja, a estrutura do modelo sugerido é dado por

$$\begin{aligned}
\hat{\mu} &= \hat{\beta}_{10} + s_{11}(PC_1) + s_{12}(PC_2), \\
\hat{\sigma} &= \hat{\beta}_{20} + s_{21}(PC_1) + s_{22}(PC_2), \\
\hat{\nu} &= \hat{\beta}_{30}, \\
\hat{\tau} &= \hat{\beta}_{40} + s_{41}(PC_1).
\end{aligned} \tag{4.2}$$

Embora a estrutura em (4.2) seja a indicada pelas funções **stepGAICll.A(·)** e **addterm(·)**, ao realizar o ajuste e a análise diagnóstica do modelo, observou-se gráficos de *worm plot* e *half normal plot* bastante insatisfatórios, em que os resíduos não seguiam a distribuição normal padrão. No apêndice A são indicados os resultados de diagnóstico obtidos no ajuste do modelo com a estrutura (4.2).

Assim sendo, a partir da estrutura indicada em (4.2) e considerando o comportamento dos gráficos de *worm plot*, foi construído um modelo com preditores distintos ao exibido em (4.2), de maneira que apresentou melhores resultados em sua análise diagnóstica e comportamento dos resíduos. Desta maneira, a estrutura do modelo final ajustado com distribuição BCTo, tendo como variável resposta *redshift*, e variáveis explicativas as duas componentes principais, é dada por,

$$\hat{\mu} = \hat{\beta}_{10} + s_{11}(PC_1) + s_{12}(PC_2) \Longrightarrow \hat{\mu} = \hat{\beta}_{10} + \sum_{i=1}^{2} s_{1i}(PC_i),$$

$$\hat{\sigma} = \hat{\beta}_{20} + s_{21}(PC_1),$$

$$\hat{\nu} = \hat{\beta}_{30} + s_{32}(PC_2),$$

$$\hat{\tau} = \hat{\beta}_{40} + s_{42}(PC_2). \tag{4.3}$$

Após a explanação realizada, a seção seguinte apresenta com maiores detalhes o modelo considerado, bem como suas especificações e diagnóstico.

4.4 MODELO PROPOSTO E SUAS ESPECIFICAÇÕES

Fundamentados na organização descrita neste trabalho, a estrutura do modelo ajustado com distribuição BCTo, tendo como variável resposta *redshift*, e como variáveis explicativas as duas componentes principais, é dada por,

$$\log(\hat{\mu}) = -0,94 + s_{11}(PC_1) + s_{12}(PC_2),$$

$$\log(\hat{\sigma}) = -0,09 + s_{21}(PC_1),$$

$$\hat{\nu} = -0,60 + s_{32}(PC_2),$$

$$\log(\hat{\tau}) = 1,14 + s_{42}(PC_2), \tag{4.4}$$

em que todas ambas PC's são variáveis quantitativas contínuas e os s_{qi}, com $q = 1, 2, 3, 4$ e $1 = 1, 2$ são funções suavizadoras, inseridas no modelo via plataforma dos modelos aditivos generalizados.

Dada a complexidade dos dados, buscou-se verificar o potencial do modelo GAMLSS. Para tanto, foram implementados outros cinco modelos de regressão, que objetiva-

vam explicar o *redshift* fotométrico por meio das duas componentes principais. Estes cinco modelo consistiram em: Um modelo linear clássico; Dois modelos lineares generalizados, em que foram consideradas as famílias Gama e Normal Inversa; Dois modelos aditivos generalizados, com as mesmas famílias adotadas para os GLM, sendo utilizado *thin plate splines* para cada componente principal.

A Tabela 4.7 apresenta algumas das principais métricas dos seis modelos de regressão, tais como a Deviance Global, o valor do AIC, os graus de liberdade do modelo ajustado e dos seus resíduos.

Tabela 4.7: Resumo das medidas dos modelos ajustados.

Métrica	**Valor**					
	LM	GLM - Gama	GLM - N.Inversa	GAM - Gama	GAM - N.Inversa	GAMLSS BCTo
AIC	-3447,407	-4063,298	-3775,635	-4425,473	-4564,592	-7647,624
BIC	-3419,227	-4035,118	-3747,455	-4285,306	-4441,006	-7303,93
Deviance Global	-3455,407	-4071,298	-3783,635	-4465,264	-4599,676	-7745,195
Graus de Lib. do Ajuste	4	4	4	19,8958	17,5423	48,7856
Graus de Lib. dos Resíduos	8472	8472	8472	8456,104	8458,458	8427,214
Núm. de Ciclos para ajuste	2	2	2	2	2	58

Fonte: Próprio autor.

Pode-se observar por meio da Tabela 4.7 que, embora o custo computacional seja o maior dentre todos os modelos ajustados, o modelo GAMLSS foi aquele que apresentou melhor ajuste frente aos dados considerados. Tem-se que as métricas AIC, BIC e Deviance Global do modelo GAMLSS, foram as que apresentaram os menores valores dentre todos os ajustes realizados. Nesse sentido, pode-se considerar que o modelo GAMLSS é aquele que apresenta o melhor ajuste, dentre os modelos de regressão implementados. No apêndice B são indicados os resultados de diagnóstico obtidos no ajuste dos modelos de regressão supracitados.

Tendo em vista que para os modelos de regressão, os modelos GAMLSS foram aqueles que apresentaram melhores resultados de ajuste para explicar o *redshift* fotométrico por meio das duas componentes principais, foram implementados outros dois modelos GAMLSS, com distribuições indicadas por meio da função **chooseDist**$(\cdot)$, à saber, as distribuições BCT e a GB2.

De maneira similar à distribuição BCTo, as distribuições BCT e GB2 apresentam quatro parâmetros distribucionais. Nesse sentido, foram considerados os mesmos preditores lineares para os parâmetros destas duas distribuições. Assim, é apresentada a Tabela 4.8 com as principais métricas dos três modelos GAMLSS ajustados.

Por meio da Tabela 4.8, pode-se supor que o modelo ajustado com a distribuição BCT é aquele que possui o melhor ajuste dentre os três modelos GAMLSS, haja vista que apresenta menores valores de AIC, BIC e Deviance Global. Contudo, ao se realizar a análise de resíduos e diagnóstico dos modelos, isso não foi confirmado. Pode-se suspeitar que dentre os motivos para esta situação, tem-se o fato da função de ligação do parâmetro μ da distribuição BCT ser a identidade, embora se tenha que $\mu > 0$. Além disso, observa-se um custo computa-

Tabela 4.8: Resumo das medidas dos modelos GAMLSS ajustados.

Métrica	Valor		
	GAMLSS - BCTo	GAMLSS - BCT	GAMLSS - GB2
AIC	-7647,624	-9679,572	-6723,289
BIC	-7303,93	-9335,578	-6416,736
Deviance Global	-7745,195	-9777,228	-6810,317
Graus de Lib. do ajuste	48,7856	48,8281	43,5137
Graus de Lib. dos resíduos	8427,214	8427,172	8432,486
Núm. de ciclos para ajuste	58	805	581

Fonte: Próprio autor.

cional significativamente menor do modelo GAMLSS - BCTo, em que com 58 ciclos obteve-se a convergência para o seu ajuste.

A seguir, é apresentada a análise de resíduos e diagnóstico dos modelos propostos, elucidando qual destes modelos apresentou melhores resultados.

4.4.1 Análise de resíduos e diagnóstico dos modelos propostos

Como destacado na seção 3.1.5, existem diversas ferramentas para a análise de resíduos de um modelo GAMLSS proposto. Para tanto, tal investigação repousa seus pressupostos no comportamento apresentado pelos resíduos quantílicos normalizados [63].

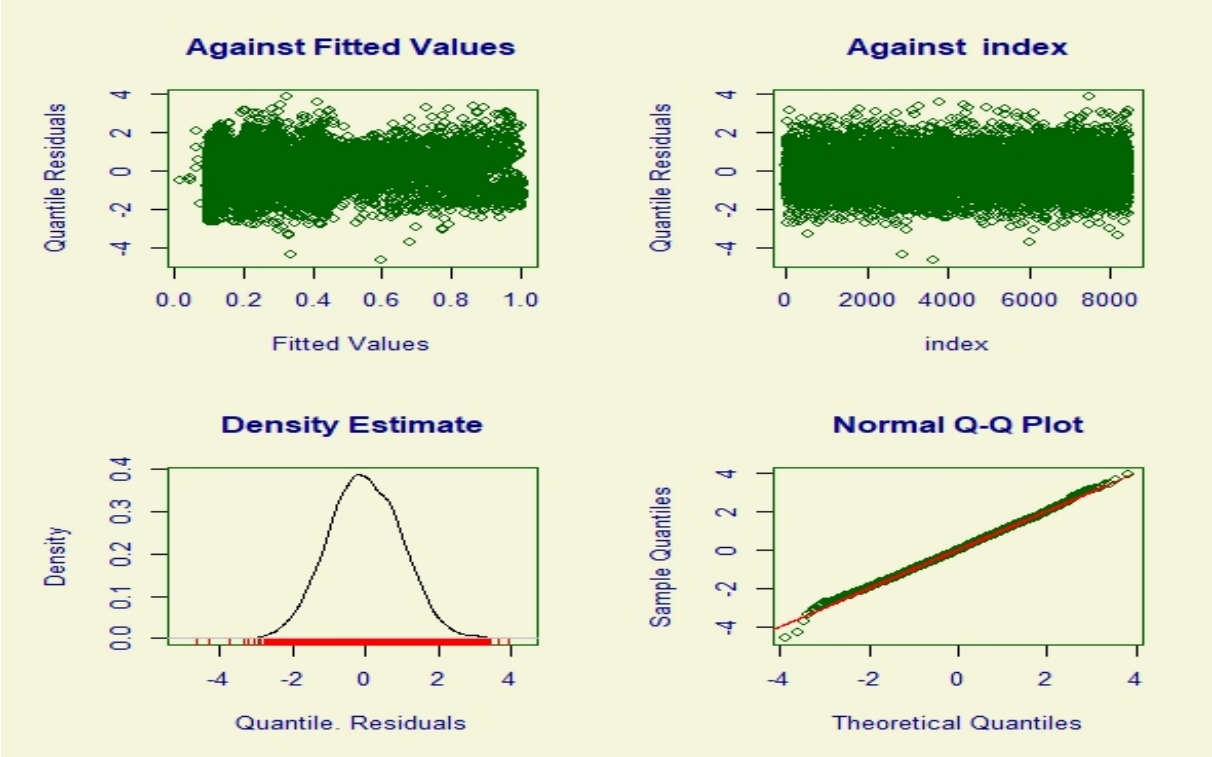
Considerando que uma das suposições fundamentais para um modelo estar bem ajustado, refere-se a necessidade de seus resíduos quantílicos apresentarem uma distribuição normal padronizada, independentemente da distribuição de probabilidade da variável resposta [63], a análise que segue visa interpretar o comportamento destes resíduos, elucidando suas características de maneira gráfica, bem como por algumas de suas métricas.

A análise de resíduos é iniciada por meio do gráfico gerado da função **plot**(·), pertencente ao pacote **gamlss**. As Figuras 4.7, 4.8 e 4.9 exibem os quatro gráficos obtidos nesta função para os três modelos GAMLSS ajustados, possibilitando construir conjecturas sobre a homocedasicidade dos seus resíduos, bem como sobre sua normalidade aproximada [34].

Considere os gráficos de resíduos quantílicos *versus* valores ajustados, situado na posição superior à esquerda das Figuras 4.7 e 4.8, referentes às distribuições BCTo e BCT, respectivamente. Pode-se observar que a nuvem de pontos existente apresenta uma variabilidade constante, com valores aproximados no intervalo de -4 até 4, ao redor da reta $y = 0$. Além disso, os ajustes GAMLSS BCTo e BCT exibem um comportamento aleatório, não apresentando um padrão sistemático.

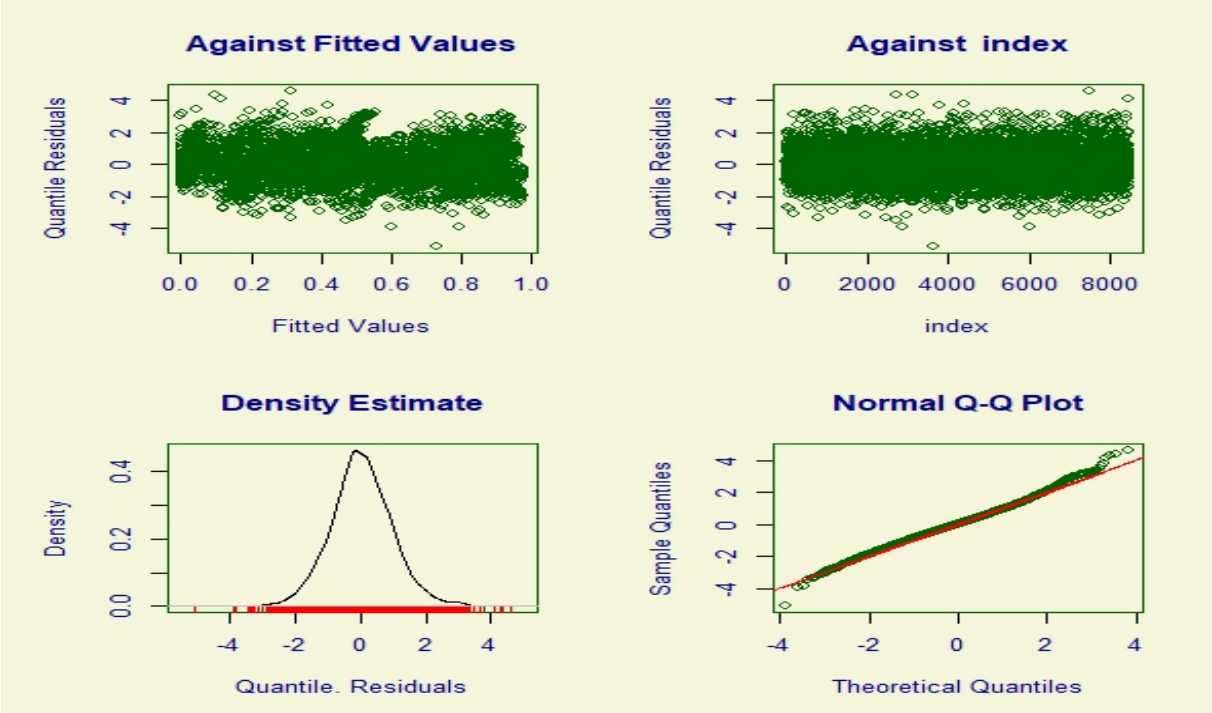
O mesmo não se verifica na Figura 4.9, referente ao ajuste com a distribuição GB2, uma vez que, além da variabilidade dos pontos ser maior (de -4 à 6), a partir do valor

Figura 4.7: Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição BCTo.



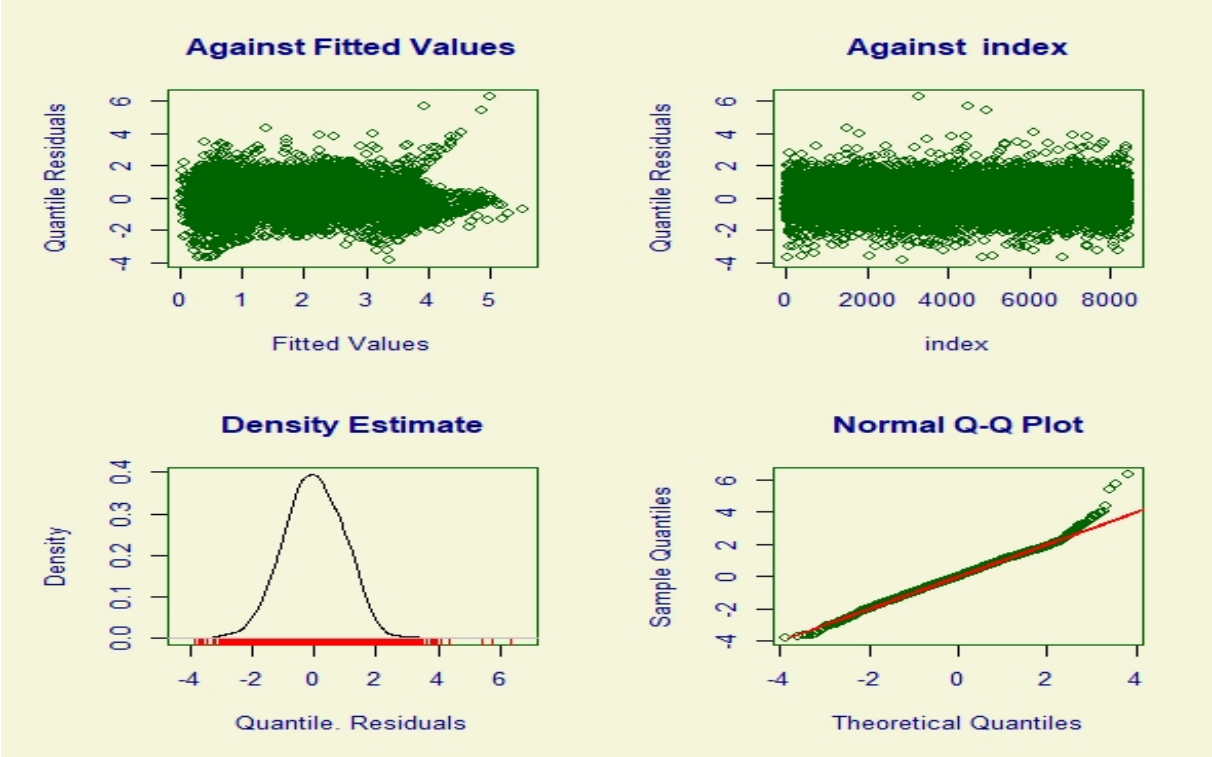
Fonte: Próprio autor.

Figura 4.8: Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição BCT.



Fonte: Próprio autor.

Figura 4.9: Gráficos dos resíduos do modelo GAMLSS ajustado com distribuição GB2.



Fonte: Próprio autor.

ajustado 4 (aproximadamente), se verifica uma tendência nestes pontos. Assim, observa-se que os ajustes BCT e BCTo apresentam características bastante satisfatórias, visto que tem-se fortes indícios da homocedasticidade dos resíduos, bem como de sua normalidade, diferentemente do ajuste GAMLSS GB2.

Outro gráfico que apresenta informações bastante relevantes, refere-se ao que ocupa a posição inferior à esquerda das Figuras 4.7, 4.8 e 4.9, isto é, o gráfico de resíduos quantílicos *versus* estimação da densidade dos resíduos. Note que as curvas exibidas possuem um formato semelhante a da distribuição normal padrão, indicando que possivelmente os resíduos são normalmente distribuídos [34].

Tal conjectura pode ser corroborada por meio do gráfico Normal Q-Q dos resíduos (posição inferior à direita das Figuras 4.7, 4.8 e 4.9). Observe que para os ajustes GAMLSS BCT e GAMLSS GB2, os pontos escapam à reta nas caudas à direita. Contudo, para o ajuste GAMLSS BCTo, os pontos posicionam-se aproximadamente sobre a reta diagonal, de maneira que dentre os três modelos ajustados, este é o que apresentou melhores resultados graficamente.

Assim, considerando em [19] que, para um modelo ajustar adequadamente às observações, seus resíduos quantílicos devem apresentar distribuição aproximadamente normal padrão (média aproximadamente zero, variância aproximadamente um, coeficiente de assimetria aproximadamente zero, e coeficiente de curtose aproximadamente três), é apresentada a

Tabela 4.9 com as medidas resumo dos resíduos quantílicos dos modelos propostos.

Tabela 4.9: Resumo das medidas dos resíduos quantílicos dos modelos GAMLSS ajustados.

Métrica	**Valor**		
	GAMLSS - BCTo	GAMLSS - BCT	GAMLSS - GB2
Média	-0,02475	0,03173	0,00359
Variância	0,99150	0,91579	0,99681
Coeficiente de assimetria	0,06077	0,13568	0,04052
Coeficiente de curtose	3,00246	3,77209	3,59428
Coeficiente de correlação de Filliben	0,99965	0,99738	0,99816

Fonte: Próprio autor.

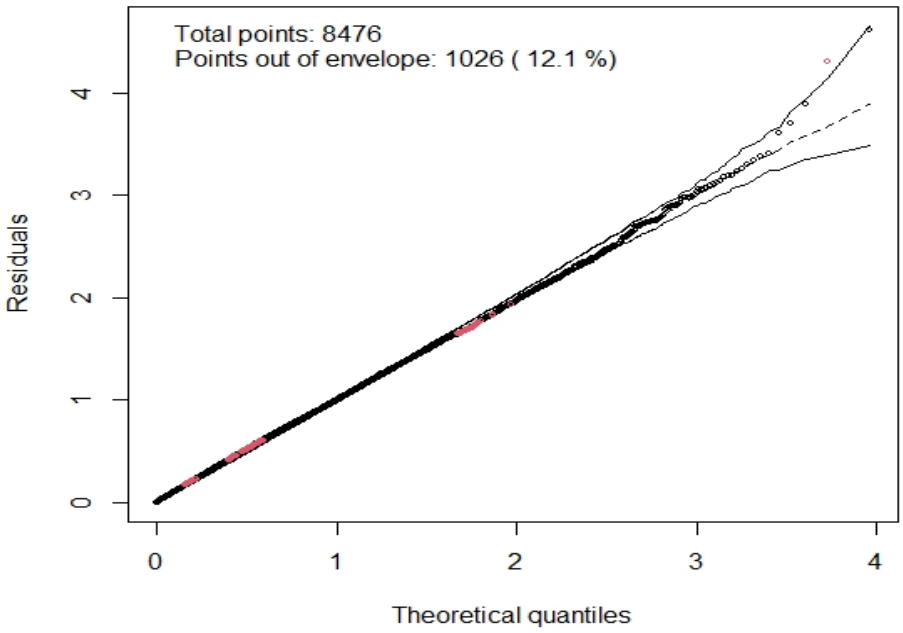
Embora todos os valores das medidas apresentadas na Tabela 4.9, se aproximem do que a literatura sugere para a normalidade dos resíduos, o ajuste realizado com a distribuição GB2 foi aquele que apresentou valores ligeiramente mais satisfatórios para a média e coeficiente de assimetria mais próximos de zero, bem como o coeficiente de assimetria mais próximo de um, com diferença extremamente baixa para os valores do ajuste BCTo (na ordem de 0,01 a 0,02). O mesmo não se verifica para o ajuste GAMLSS BCT, já que a medidas dos coeficientes de variância e assimetria valem $0,91579$ e $0,13568$, respectivamente (diferenças na ordem de 0,08 aproximadamente para os outros ajustes).

Contudo, ao analisar as medidas do coeficiente de curtose dos ajustes GAMLSS BCT e GAMLSS GB2, observa-se uma grande diferença dos valores para com o que a literatura trás luz (aproximadamente 3). Já para o ajuste GAMLSS BCTo, esse valor é muito próximo ao desejado e, portanto, é aquele que melhor capta a curtose dos dados para o processo de modelagem. Portanto, por meio dos padrões das distribuições exibidas nas Figuras 4.7, 4.8 e 4.9 e dos valores descritivos dos resíduos apresentados na Tabela 4.9, pode-se concluir que os resíduos quantílicos do modelo proposto com a distribuição BCTo é o que mais se aproxima da distribuição normal padrão, com relação aos demais modelos ajustados.

Por fim, buscando-se por mais uma alternativa que possibilite verificar a conjectura da normalidade dos resíduos quantílicos, além de indicativos que informe qual modelo GAMLSS melhor se ajusta aos dados [69], considerou-se o gráfico denominado *half Normal plot* com envelope simulado. Para os três modelos GAMLSS, as Figuras 4.10, 4.11 e 4.12 apresentam o gráfico supracitado.

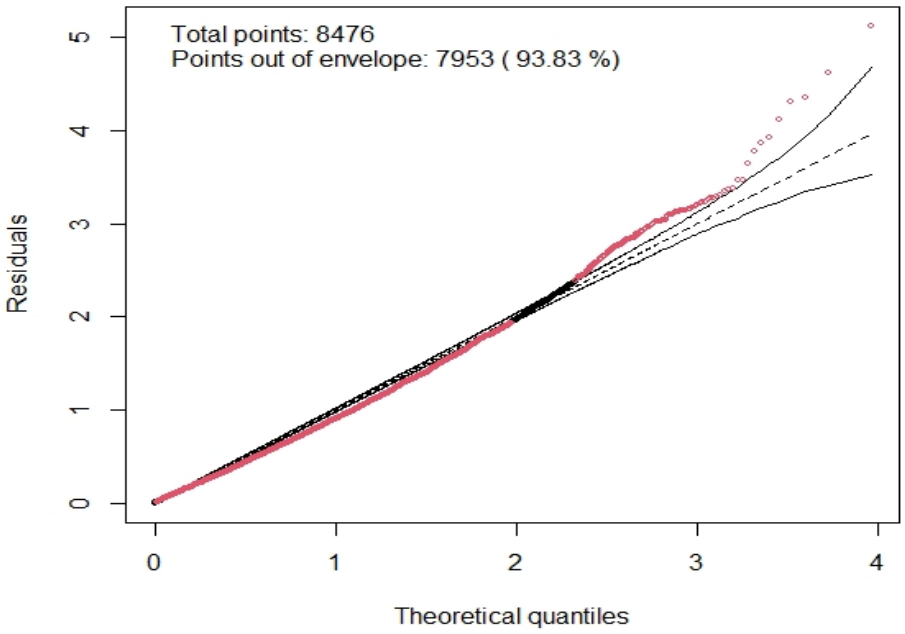
Tem-se em [69], que nesse tipo de gráfico, a estrutura esperada para um modelo adequado aos dados, corresponde aos resíduos quantílicos dispersos aleatoriamente internamente ao envelope simulado. Constata-se que o modelo GAMLSS BCTo atende razoavelmente ao que a literatura trás luz. O mesmo não é observado nas Figuras 4.11 e 4.12, em que os resíduos quantílicos dos modelos GAMLSS BCT e GAMLSS GB2, escapam em uma quantidade significativa do interior do envelope simulado, principalmente em suas caudas à direita

Figura 4.10: *Half Normal plot* dos resíduos do modelo GAMLSS BCTo.



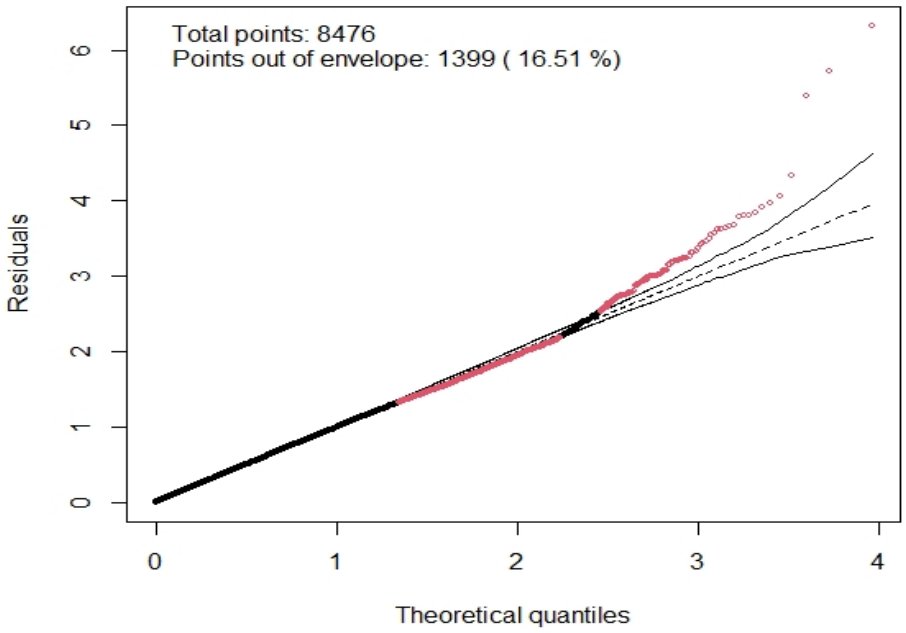
Fonte: Próprio autor.

Figura 4.11: *Half Normal plot* dos resíduos do modelo GAMLSS BCT.



Fonte: Próprio autor.

Figura 4.12: *Half Normal plot* dos resíduos do modelo GAMLSS GB2.



Fonte: Próprio autor.

(demarcados na cor vermelha).

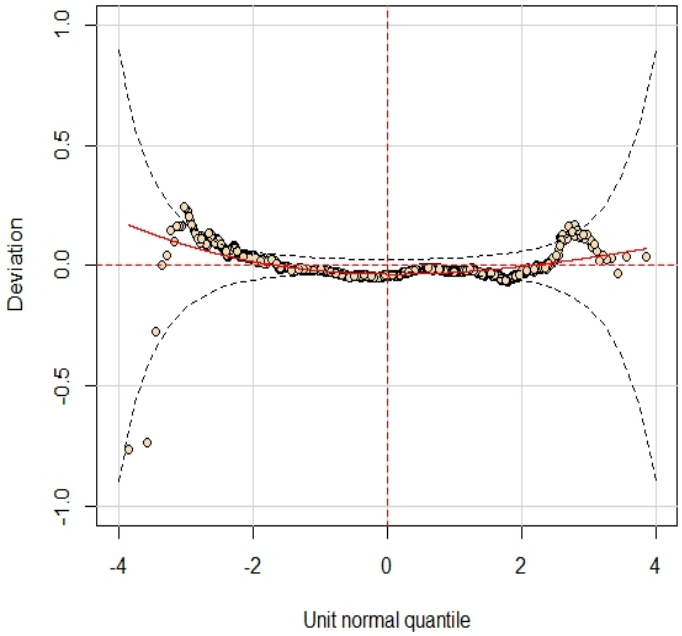
Tais pontos contabilizam um total de $12,1\%$ dos resíduos para o GAMLSS BCTo ajustado; para o modelo GAMLSS BCT e GB2 tem-se um total de $93,83\%$ e $16,51\%$, respectivamente.

Corroborando-se com as evidências de um bom ajuste do modelo GAMLSS BCTo, em que os resíduos quantílicos seguem aproximadamente a distribuição normal, além de apresentar um melhor ajuste frente aos modelos GAMLSS BCT e GAMLSS GB2, as Figuras 4.13, 4.14 e 4.15, ilustram os gráficos *worm plot* dos resíduos, permitindo identificar regiões (caso existam) em que o modelo não se ajusta adequadamente aos dados.

Tendo em vista que um modelo bem ajustado aos dados deve apresentar seus pontos dentro do intervalo de confiança, não apresentando uma forma de "minhoca", pode-se verificar no gráfico do *worm plot* do modelo GAMLSS BCTo que, exceto em uma pequena parcela de pontos próximos a cauda à esquerda, todos os pontos estão dispostos dentro da banda de confiança de 95% e próximos da reta $y = 0$, indicando um bom ajuste do modelo. O mesmo não é observado para os outros dois ajustes, uma vez que muitos dos resíduos estão contidos na região de violação, ou seja, fora da banda de confiança.

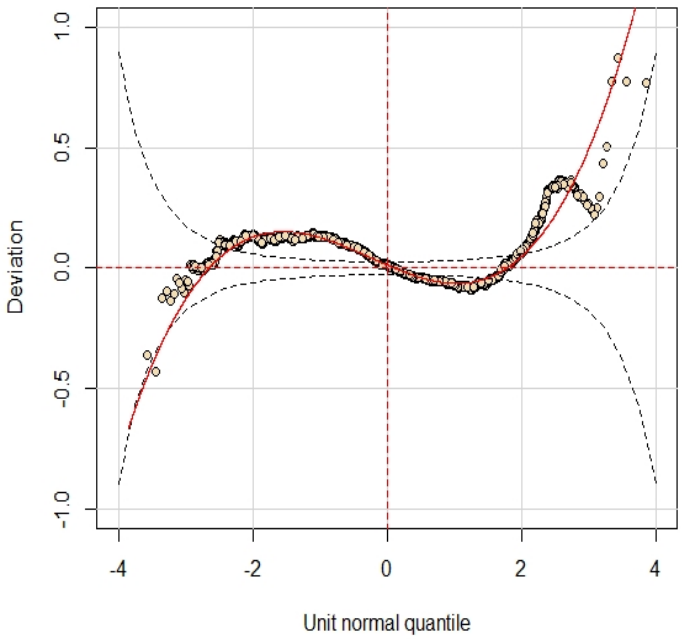
Assim sendo, considerando que o ajuste GAMLSS BCTo tenha apresentado melhores resultados frente aos ajustes GAMLSS BCT e GAMLSS GB2, buscou-se por mais um indicativo sobre a normalidade dos resíduos do modelo proposto com a distribuição BCTo. Para tanto, foi realizado o teste de kolmogorov-Smirnov ($ks.test$). Segundo [66], a estatística

Figura 4.13: *Worm plot* do modelo GAMLSS BCTo.



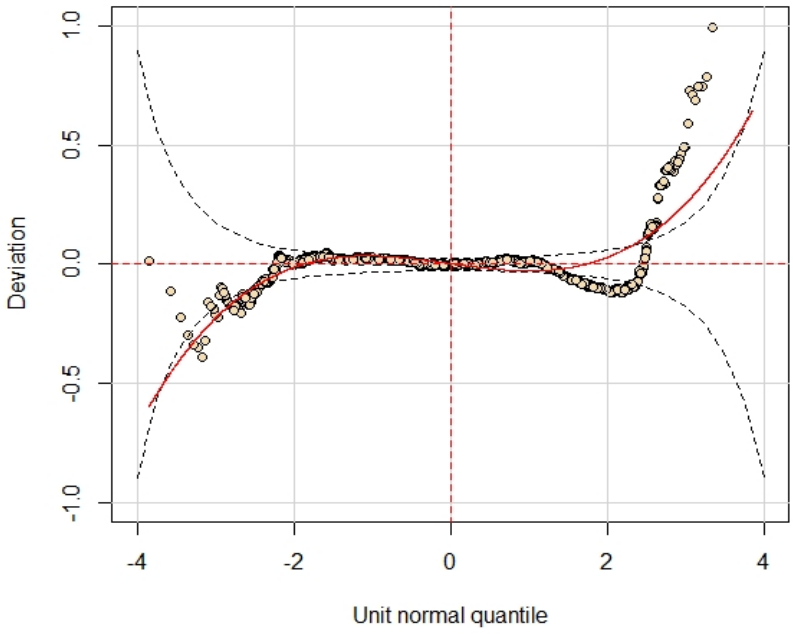
Fonte: Próprio autor.

Figura 4.14: *Worm plot* do modelo GAMLSS BCT.



Fonte: Próprio autor.

Figura 4.15: *Worm plot* do modelo GAMLSS GB2.



Fonte: Próprio autor.

Kolmogorov-Smirnov permite analisar a qualidade do ajuste entre uma distribuição teórica e um conjunto de observações, de maneira que por meio do seu $p-$valor pode-se ter evidências da aproximação, ou não, dos dados à distribuição verificada.

Nesse sentido, realizando-se o Teste de Kolmogorov-Smirnov para verificar se os resíduos do modelo GAMLSS BCTo se aproximam da distribuição Normal padrão, foi obtido um $p-$valor $= 0,276$, isto é, tem-se evidências de que tais resíduos são provenientes de uma distribuição normal padrão.

4.5 Efeito das funções de suavização nos preditores dos parâmetros distribucionais

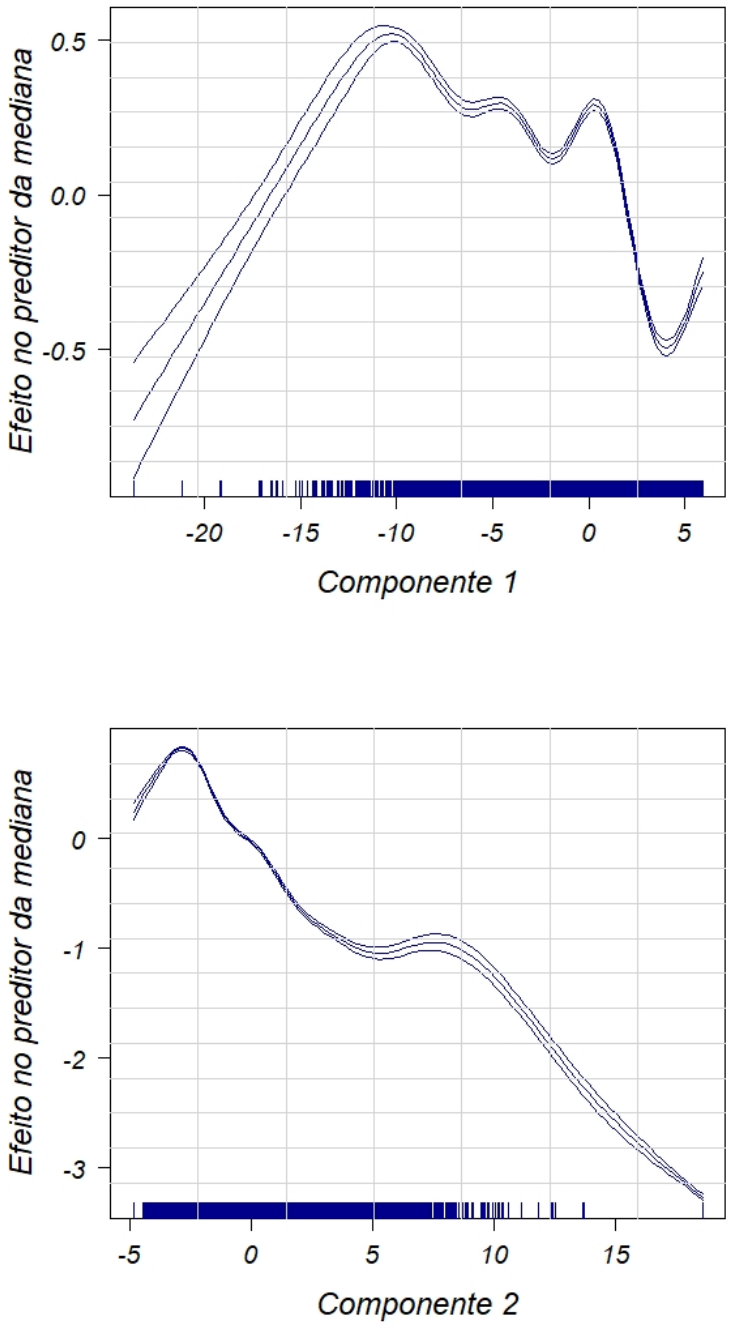
Embora seja bastante complexo realizar interpretações do relacionamento das onze bandas do espectro utilizadas para estimar o *redshift* fotométrico devido a técnica das componentes principais, pode-se observar o comportamento das funções de suavização utilizadas nestes objetos (que são combinações lineares das onze bandas), e que fazem parte dos preditores lineares dos quatro parâmetros distribucionais do modelo proposto.

A seguir, são apresentados os efeitos das funções de suavização nos quatro parâmetros distribucionais, exibindo como estes elementos interferem na variável resposta *redshift* fotométrico.

4.5.1 Parâmetro μ (mediana)

As Figuras $4.16(a)$ e $4.16(b)$ apresentam o efeito exercido pelas funções de suavização utilizadas nas duas componentes principais, para o preditor do parâmetro da mediana (μ).

Figura 4.16: Comportamento das funções de suavização para o parâmetro μ.



Fonte: Próprio autor.

O efeito exercido pela PC1, por meio da função de suavização estimada sobre o preditor linear da mediana distribucional, exibido na Figura $4.16(a)$, revela um comportamento bastante interessante ao longo dos valores assumidos por esta covariável.

Inicialmente, pode-se observar que o efeito exercido pelo suavizador $s_{11}(\text{PC1})$ sobre a mediana da variável *redshift* fotométrico apresenta um comportamento bastante oscilante. Como a função de ligação para este parâmetro é a logarítmica, o efeito de cada componente, por meio das funções de suavização é multiplicativo. Nesse sentido, para valores em que, $-17 < \text{PC1} < 2$ (aproximadamente), tem-se um efeito de acréscimo na mediana do *redshift* fotométrico, exibindo um efeito máximo quando PC1 ≈ -10. Já para, $-17 > \text{PC1} > 2$ (aproximadamente), observa-se um efeito de decréscimo na mediana do *redshift*, com efeito mínimo quando PC1 $\approx -23,6$.

Considerando que para esta primeira componente principal, as magnitudes que mais contribuem em sua composição são as provenientes dos filtros **IRAC1** e **IRAC2** (Quadro 4.2), pode-se inferir que estas bandas são as que mais interferem no efeito de incremento ou decréscimo sobre a mediana do *redshift* fotométrico.

Ao analisar o efeito exercido pela PC2, por meio da função de suavização estimada s_{12}, sobre o preditor linear da mediana distribucional exibido na Figura $4.16(b)$, tem-se que, em geral (PC2 > 0), o efeito da função de suavização estimada apresenta características que tendem a minorar o valor da mediana do *redshift* fotométrico. Somente para valores em que a PC2 seja menor do que zero, seu efeito é de incremento na mediana da variável resposta, apresentando máximo efeito quando PC2 ≈ -3. Ressalta-se que, em geral, o efeito de decréscimo exercido sobre a mediana do *redshift* é mais intenso quanto maior o valor assumido pela PC2, a partir de -3.

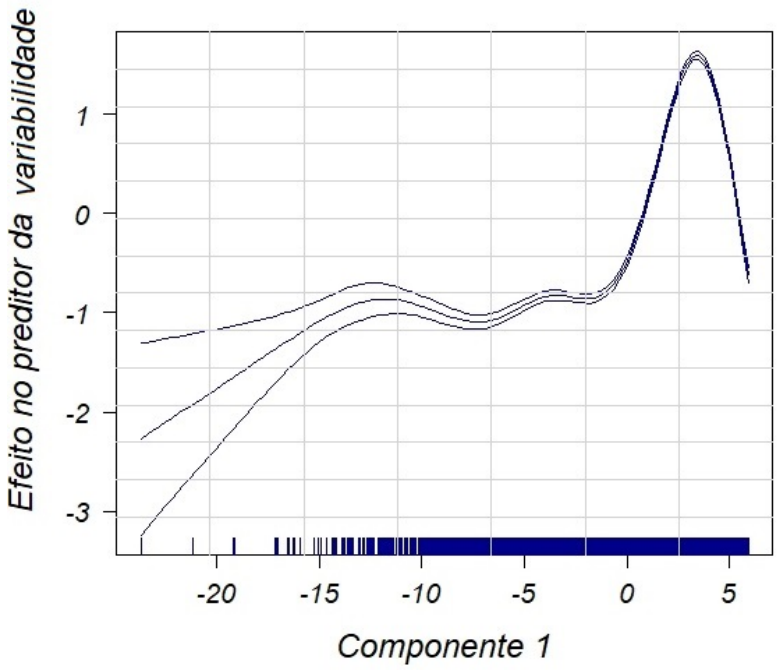
Além disso, por meio do Quadro 4.2, observa-se que considerando a segunda componente principal, os filtros que mais contribuem em sua estrutura são relacionadas as magnitudes de **IRAC1** e **up**. Nesse sentido, pode-se inferir que estas bandas são as que mais interferem no efeito de incremento ou decréscimo sobre a mediana do *redshift* fotométrico. Como o filtro **IRAC1** está presente como um dos principais elementos que contribuem na composição de ambas as componentes principais, pode-se inferir que suas magnitudes são as que mais interferem na mediana do *redshift* fotométrico.

4.5.2 Parâmetro σ (coeficiente de variação)

Para o parâmetro da variabilidade, ressalta-se que segundo a estrutura (4.4), a única componente significativa foi a PC1. Nesse sentido, a Figura 4.17 apresenta o efeito exercido pela PC1, por meio da função de suavização estimada s_{21}, sobre o preditor linear da variabilidade distribucional.

Por meio da Figura 4.17 pode-se observar que, em grande parte dos valores assumidos por esta componente principal, o efeito exercido sobre a variabilidade é quase que

Figura 4.17: Comportamento da função de suavização para o parâmetro σ.



Fonte: Próprio autor.

sempre minorador. Verifica-se que para valores desta componente menores do que -2, tem-se um efeito de decréscimo constante. Para valores entre -2 e 2 ($-2 <$ PC1 < 2, aproximadamente), ainda observa-se um efeito minorador sobre a variabilidade, porém de menor intensidade. No intervalo $(2, 5)$ assumidos por esta componente, o efeito deixa de ser minorador no coeficiente de variação, tornando-se de acréscimo na medida de variabilidade, apresentando máximo efeito quando PC1 ≈ 4. Para valores desta componente maiores do que 5 o efeito retorna a ser de decréscimo na variabilidade.

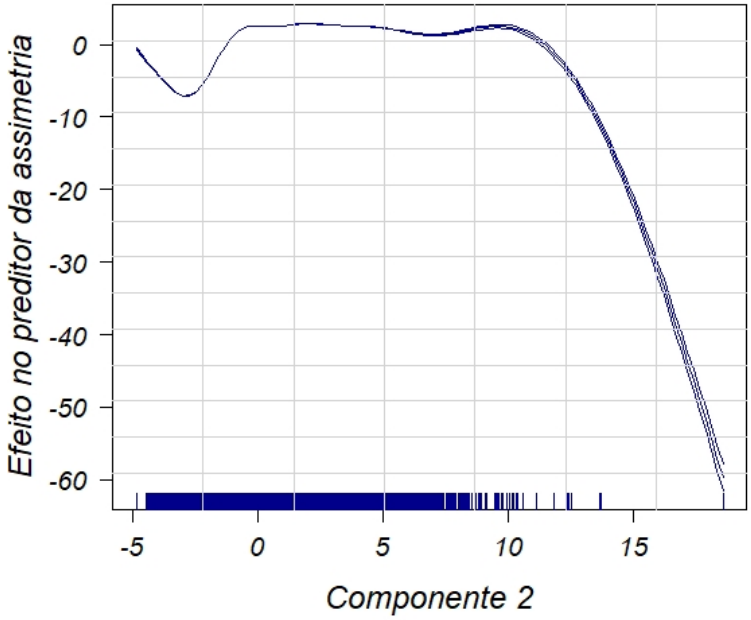
4.5.3 Parâmetro ν (coeficiente da assimetria)

Os efeitos realizados pela PC2, inserida no preditor do terceiro parâmetro distribucional (ν), sob a forma da função suavizadora s_{32}, pode ser observado por meio da Figura 4.18.

Pode-se verificar por meio da Figura 4.18 que, embora se tenha um intervalo significativo de incremento na medida do coeficiente de assimetria referente a PC2 $(-1, 11)$ aproximadamente, este acréscimo é de intensidade bastante pequena, próximo de zero. Efeito distinto é observado no decréscimo dos valores do coeficiente de assimetria. De fato, note que o intervalo de decréscimo nesta medida $(-5, -1) \cup (11, 18)$, aproximadamente, acarreta em um efeito de maior intensidade na diminuição do coeficiente de assimetria.

Ressalta-se que para valores maiores do que PC2 ≈ 11, quanto maior o va-

Figura 4.18: Comportamento da função de suavização para o parâmetro ν.



Fonte: Próprio autor.

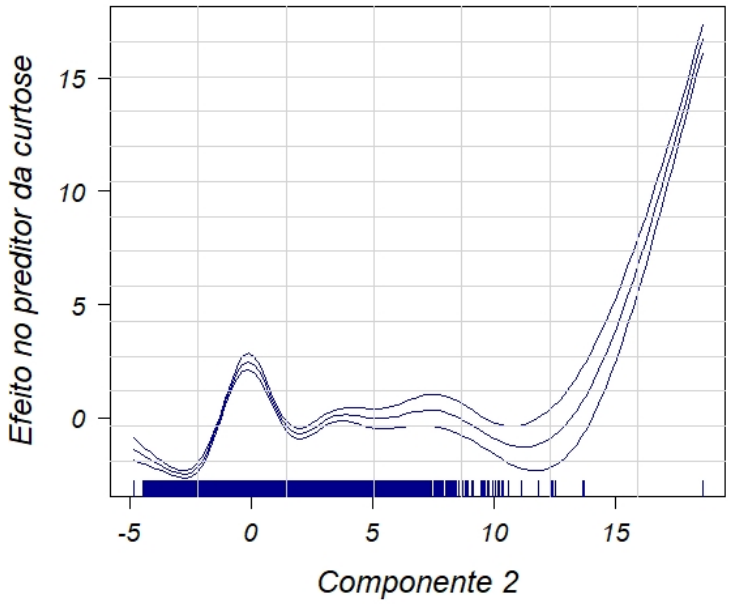
lor assumido por esta componente, mais intenso o efeito de decréscimo exercido pela função suavizadora no preditor da assimetria do modelo ajustado. Contudo, deve-se tomar bastante cautela nesta interpretação, fundamentalmente para valores em que PC2 seja maior que aproximadamente 11 (momento em que se verifica um intenso efeito de decréscimo no preditor da assimetria). Isso se justifica, uma vez que se observa um número bastante reduzido de informação proveniente da segunda componente principal.

4.5.4 Parâmetro τ (coeficiente da curtose)

Os efeitos realizados pela PC2, inserida no preditor do quarto parâmetro distribucional (τ), sob a forma de função suavizadora, pode ser observado por meio da Figura 4.19.

Assim como observado no efeito exercido pelo suavizador aplicado na PC1 sobre o parâmetro da mediana, é possível verificar na Figura 4.19, que o efeito exercido pela PC2 sobre o parâmetro da curtose apresenta um comportamento bastante oscilatório, com alternância entre momentos em que seu efeito é de acréscimo na medida deste parâmetro (quando PC2 $\in (-1, 9) \cup (12, 18)$, aproximadamente) e momentos em que seu efeito é de decréscimo, quando os valores assumidos pela PC2 sejam menores que -1 (PC2 < -1), e quando ($9 <$ PC2 < 12), aproximadamente. Assim como para o parâmetro da assimetria, deve-se apre-

Figura 4.19: Comportamento da função de suavização para o parâmetro τ.



Fonte: Próprio autor.

sentar um certo cuidado ao afirmar sobre o efeito da segunda componente principal no preditor da curtose, quando esta apresenta valores aproximadamente maiores do que 11.

Em síntese, tem-se que para os quatro parâmetros distribucionais, é possível compreender a necessidade das funções suavizadoras para cada componente principal, já que se verifica um relacionamento não linear dos mesmos para a explicação da variável resposta *redshift* fotométrico.

5 CONSIDERAÇÕES FINAIS

Por meio da pesquisa estabelecida, foi possível observar que a classe de Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS) se mostrou bastante eficaz na modelagem de *redshift* fotométrico. Embora se tenha utilizado a Análise de Componentes Principais, com intuito de superar o efeito de multicolinearidade no ajuste do modelo, tornando sua interpretabilidade mais complexa, pode-se obter uma estrutura que apresentou um bom ajuste dos dados, em que seus resíduos são normalmente distribuídos.

Nesse sentido, considerando o objetivo proposto neste trabalho, pode-se inferir que a estrutura GAMLSS atendeu aspectos relativos ao ajuste e a predição de modelos envolvendo *redshift* fotométrico. Contudo, pode-se destacar que uma limitação imposta nesta pesquisa, diz respeito ao elevado volume de dados contido no **CosmoPhotoz**. Foi observado que, quando estabelecido um conjunto de dados superiores a 5% da totalidade deste pacote, todos os aspectos de análise de distribuições marginais, estruturas dos preditores para cada parâmetro distribucional, processos de seleção das covariáveis e das componentes principais, se tornou extremamente moroso, com alto custo computacional.

Para tanto, foi tomada uma amostra deste banco de dados (8476 observações), para que assim, se tornasse viável o tratamento e modelagem dos dados considerados. Entende-se que para conjuntos de dados com um elevado número de observações (*Big Data*) outras alternativas podem ser analisadas, tais como, redes neurais ou modelagem para dados de alta dimensão.

Mesmo com essa dificuldade computacional, a estrutura de regressão apresentada trouxe resultados bastante satisfatórios, frente a trabalhos existentes na literatura. Isso se deve a flexibilidade dos GAMLSS, permitindo a modelagem de outros parâmetros além da média, como observado no trabalho de [20] que considerou um GLM com distribuição Gama para explicação do *redshift*. Ademais, o forte pressuposto da possibilidade de uso de termos não paramétricos (suavizadores) para os preditores lineares, trouxeram resultados robustos para o poder preditivo do modelo apresentado nessa pesquisa.

Em síntese, essa pesquisa mostrou algumas das potencialidades da estrutura GAMLSS, na modelagem de *redshift* fotométrico. Pesquisas que busquem por considerar *big data* de *redshift* fotométricos, para a estrutura GAMLSS é uma proposição que pode ser considerada para futuras investigações.

REFERÊNCIAS

[1] ABDALLA, F. B., BANERJI, M., LAHAV, O., AND RASHKOV, V. A comparison of six photometric redshift methods applied to 1.5 million luminous red galaxies. *Monthly Notices of the Royal Astronomical Society 417*, 3 (2011), 1891–1903.

[2] AKAIKE, H. Information measures and model selection. *Bulletin of the International Statistical Institute 44* (1982), 277–290.

[3] BARBOSA, J. C. Modelagem e modelos matemáticos na educação científica. *Alexandria: Revista de Educação em Ciência e Tecnologia 2*, 2 (2009), 69–85.

[4] BILICKI, M., HOEKSTRA, H., BROWN, M., AMARO, V., BLAKE, C., CAVUOTI, S., DE JONG, J., GEORGIOU, C., HILDEBRANDT, H., WOLF, C., ET AL. Photometric redshifts for the kilo-degree survey-machine-learning analysis with artificial neural networks. *Astronomy & Astrophysics 616* (2018), A69.

[5] BONNETT, C. Using neural networks to estimate redshift distributions. an application to cfhtlens. *Monthly Notices of the Royal Astronomical Society 449*, 1 (2015), 1043–1056.

[6] BOX, G. E., AND COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological) 26*, 2 (1964), 211–243.

[7] BUUREN, S. V., AND FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine 20*, 8 (2001), 1259–1277.

[8] CHEIN, F. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas.

[9] COLE, T. J., AND GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine 11*, 10 (1992), 1305–1319.

[10] COLLISTER, A. A., AND LAHAV, O. Annz: estimating photometric redshifts using artificial neural networks. *Publications of the Astronomical Society of the Pacific 116*, 818 (2004), 345.

[11] CORDEIRO, G. M., AND DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP* (2008).

[12] COSTA, F. D. C. *GAMLSS espaçotemporal para engenharia de avaliações.* PhD thesis, Universidade de Londrina, Brasil, 2022.

[13] CURRAN, S., MOSS, J., AND PERROTT, Y. Qso photometric redshifts using machine learning and neural networks. *Monthly Notices of the Royal Astronomical Society 503*, 2 (2021), 2639–2650.

[14] CYPRIANO, E. Análise de dados em astronomia i. *Notas de aula - IAG/USP 1*, 1 (2017).

[15] DA COSTA, S. C. *Modelos lineares generalizados mistos para dados longitudinais.* PhD thesis, University of São Paulo, Brazil, 2003.

[16] DA SERRA COSTA, J. F., CORREIA, M. G., AND DE SOUZA, L. T. T. Utilização do método de análise hierárquica na escolha de software estatístico para a demanda de uma universidade pública. *Produto & Produção 12*, 1 (2011).

[17] DE ARAÚJO FLORENCIO, L. Engenharia de avaliações com base em modelos gamlss. Master's thesis, Universidade Federal de Pernambuco, 2010.

[18] DE SOUZA SAMPAIO, N. A., AND DE MORAES DANELON, M. C. T. Aplicações da estatística nas ciências.

[19] DUNN, P. K., AND SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and graphical statistics 5*, 3 (1996), 236–244.

[20] ELLIOTT, J., DE SOUZA, R. S., KRONE-MARTINS, A., CAMERON, E., ISHIDA, E. E., HILBE, J., COLLABORATION, C., ET AL. The overlooked potential of generalized linear models in astronomy-ii: Gamma regression and photometric redshifts. *Astronomy and computing 10* (2015), 61–72.

[21] FIRTH, A. E., LAHAV, O., AND SOMERVILLE, R. S. Estimating photometric redshifts with artificial neural networks. *Monthly Notices of the Royal Astronomical Society 339*, 4 (2003), 1195–1202.

[22] FREUND, R., WILSON, W., AND SA, P. Statistical modeling of a response variable. *Regression Analysis; Academic Press: St. Louis, MI, USA* (1998).

[23] FRIAÇA, A. Cronos e cosmos. *Ciência e Cultura 54*, 2 (2002), 37–40.

[24] GERDES, D. W., SYPNIEWSKI, A. J., MCKAY, T. A., HAO, J., WEIS, M. R., WECHSLER, R. H., AND BUSHA, M. T. Arborz: Photometric redshifts using boosted decision trees. *The Astrophysical Journal 715*, 2 (2010), 823.

[25] GIAVALISCO, M., FERGUSON, H., KOEKEMOER, A., DICKINSON, M., ALEXANDER, D., BAUER, F., BERGERON, J., BIAGETTI, C., BRANDT, W., CASERTANO, S., ET AL. The great observatories origins deep survey: initial results from optical and near-infrared imaging. *The Astrophysical Journal 600*, 2 (2004), L93.

[26] HARDIN, J. W., HARDIN, J. W., HILBE, J. M., AND HILBE, J. *Generalized linear models and extensions*. Stata press, 2007.

[27] HASTIE, R., AND TIBSHIRANI, T. Generalized additive models. *CRC Monographs on Statistics & Applied Probability. New York: Chapman & Hall* (1990).

[28] HENGHES, B., PETTITT, C., THIYAGALINGAM, J., HEY, T., AND LAHAV, O. Benchmarking and scalability of machine-learning methods for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society 505*, 4 (2021), 4847–4856.

[29] HILDEBRANDT, H., ARNOUTS, S., CAPAK, P., MOUSTAKAS, L., WOLF, C., ABDALLA, F. B., ASSEF, R., BANERJI, M., BENÍTEZ, N., BRAMMER, G., ET AL. Phat: Photo-z accuracy testing. *Astronomy & Astrophysics 523* (2010), A31.

[30] HOFFMANN, R. Análise de regressão: uma introdução à econometria.

[31] HONGYU, K., SANDANIELO, V. L. M., AND DE OLIVEIRA JUNIOR, G. J. Análise de componentes principais: resumo teórico, aplicação e interpretação. *ES Engineering and Science 5*, 1 (2016), 83–90.

[32] HOTELLING, H. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology 10*, 2 (1957), 69–79.

[33] IHAKA, R., AND GENTLEMAN, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics 5*, 3 (1996), 299–314.

[34] JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

[35] JOHNSON, R. A., WICHERN, D. W., ET AL. Applied multivariate statistical analysis. *New Jersey 405* (1992).

[36] JOHNSON, R. A., WICHERN, D. W., ET AL. *Applied multivariate statistical analysis*, vol. 5. Prentice hall Upper Saddle River, NJ, 2002.

[37] JOLLIFFE, I. Principal component analysis. *Encyclopedia of statistics in behavioral science* (2005).

[38] JOLLIFFE, I. T. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 21*, 2 (1972), 160–173.

[39] JUNIOR, W. A. D. S. *Redshifts Fotométricos com Regressão Ponderada Localmente*. PhD thesis, Universidade de São Paulo, 2007.

[40] KAISER, H. F. The application of electronic computers to factor analysis. *Educational and psychological measurement 20*, 1 (1960), 141–151.

[41] KENDALL, M. *A Course in Multivariate Analysis*, vol. 620. London: Grin, 1957.

[42] MISHRA, S. P., SARKAR, U., TARAPHDER, S., DATTA, S., SWAIN, D., SAIKHOM, R., PANDA, S., AND LAISHRAM, M. Multivariate statistical data analysis-principal component analysis (pca). *International Journal of Livestock Research 7*, 5 (2017), 60–78.

[43] MONTGOMERY, D. C., PECK, E. A., AND VINING, G. G. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[44] MORETTIN, PEDRO ALBERTO E SINGER, J. D. M. *Estatística e ciência de dados*, 1 ed. LTC, 2022.

[45] MUSETTI, M., AND IZBICKI, R. Combinando métodos de aprendizado supervisionado para a melhoria da previsão do redshift de galáxias. *TEMA (São Carlos) 21* (2020), 117–131.

[46] NAGELKERKE, N. J., ET AL. A note on a general definition of the coefficient of determination. *Biometrika 78*, 3 (1991), 691–692.

[47] NASCIMENTO, CLÁUDIO AUGUSTO OLLER; GUARDANI, R. Análise estatística multivariada aplicada a processos químicos. *notas de aula* (2007).

[48] NELDER, J. A., AND WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General) 135*, 3 (1972), 370–384.

[49] NETO, G. B. L. Cosmologia. *Notas de Aula. Universidade de São Paulo, Instituto de Astronomia, Geofísica e Ciências Atmosféricas* (2022).

[50] OLIVEIRA FILHO, K. D. S., AND SARAIVA, M. D. F. O. Astronomia e astrofísica. *São Paulo: Editora Livraria da Física 780*, 2004 (2004), 183.

[51] PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science 2*, 11 (1901), 559–572.

[52] PROJECT, R. "what is r?". *Available at http://www.r-project. org/about.html/.* (2004).

[53] REGAZZI, A. J. Análise multivariada, notas de aula inf 766. *Departamento de Informática da Universidade Federal de Viçosa 2* (2000).

[54] RIGBY, R. A., AND STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing 6*, 1 (1996), 57–65.

[55] RIGBY, R. A., AND STASINOPOULOS, D. A flexible regression approach using gamlss in r. *London Metropolitan University, London* (2009), 47.

[56] RIGBY, R. A., AND STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*, 3 (2005), 507–554.

[57] RIGBY, R. A., AND STASINOPOULOS, D. M. Using the box-cox t distribution in gamlss to model skewness and kurtosis. *Statistical Modelling 6*, 3 (2006), 209–229.

[58] RIGBY, R. A., STASINOPOULOS, D. M., AND AKANTZILIOTOU, C. A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse gaussian distribution. *Computational Statistics & Data Analysis 53*, 2 (2008), 381–393.

[59] RIGBY, R. A., STASINOPOULOS, M. D., HELLER, G. Z., AND DE BASTIANI, F. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press, 2019.

[60] SCHNEIDER, P. *Extragalactic astronomy and cosmology: an introduction*, vol. 146. Springer, 2006.

[61] STASINOPOULOS, M., RIGBY, B., AND STASINOPOULOS, M. M. The gamlss package. *R help files* (2006).

[62] STASINOPOULOS, M., RIGBY, B., VOUDOURIS, V., AKANTZILIOTOU, C., ENEA, M., AND KIOSE, D. Package 'gamlss', 2022.

[63] STASINOPOULOS, M., RIGBY, B., VOUDOURIS, V., HELLER, G., AND DE BASTIANI, F. Flexible regression and smoothing. the gamlss packages in r. *GAMLSS for Statistical Modelling. GAMLSS for Statistical Modeling* (2015).

[64] THOMAS, G. Gamlss with applications to zero inflated and hierarchical data. Master's thesis, DIssertação. Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP., 2017.

[65] VARELLA, C. A. A. Análise multivariada aplicada as ciências agrárias: análise de componentes principais. *Seropédica, RJ: Pós-Graduação em Agronomia Ciência do Solo: CPGA-CS, Universidade Federal Rural do Rio de Janeiro* (2008).

[66] WEISS, M. S. Modification of the kolmogorov-smirnov statistic for use with correlated data. *Journal of the American Statistical Association 73*, 364 (1978), 872–875.

[67] WOOD, S. N. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006.

[68] WOOD, S. N. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 3 (2008), 495–518.

[69] YANG, Z., AND SUN, X. Generating half-normal plot for zero-inflated binomial regression. *Proceedings of PharmaSUG* (2006).

[70] ZELTERMAN, D. *Applied multivariate statistics with R*. Springer, 2015.

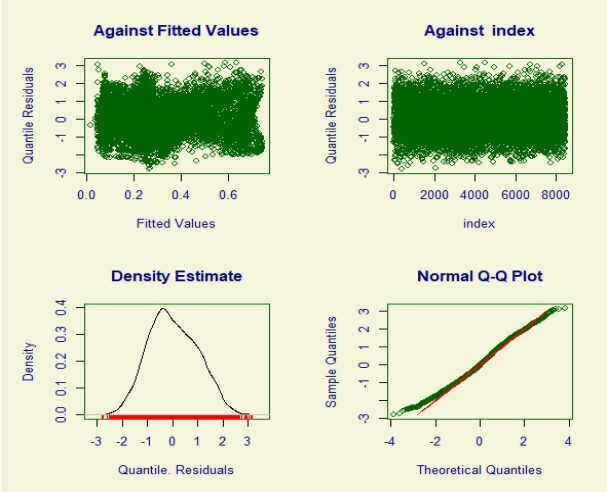
[71] ZHOU, X., GONG, Y., MENG, X.-M., ZHANG, X., CAO, Y., CHEN, X., AMARO, V., FAN, Z., AND FU, L. Spectroscopic and photometric redshift estimation by neural networks for the china space station optical survey (css-os). *The Astrophysical Journal 909*, 1 (2021), 53.

A APÊNDICE A: RESULTADOS OBTIDOS COM OS PREDITORES INDICADOS PELAS FUNÇÕES ADDTERM(·) E STEPGAICALL.A(·)

Os ajustes realizados para as distribuições BCTo, BCT e GB2, apresentam o mesmo número de observações (8476) e os mesmos preditores para os parâmetros distribucionais, indicados pela estrutura (4.2).

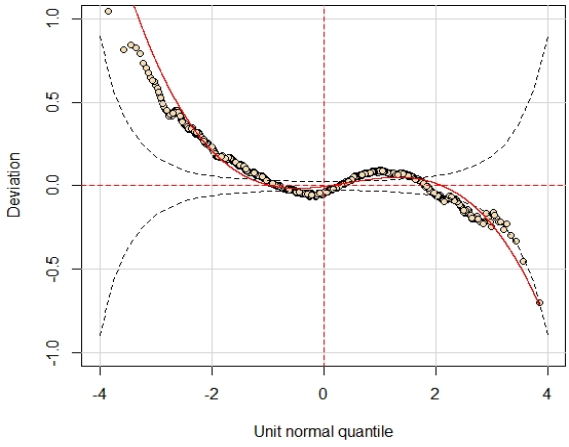
Ajuste realizado com a distribuição BCTo

Figura A.1: Gráficos dos resíduos do modelo ajustado para BCTo.



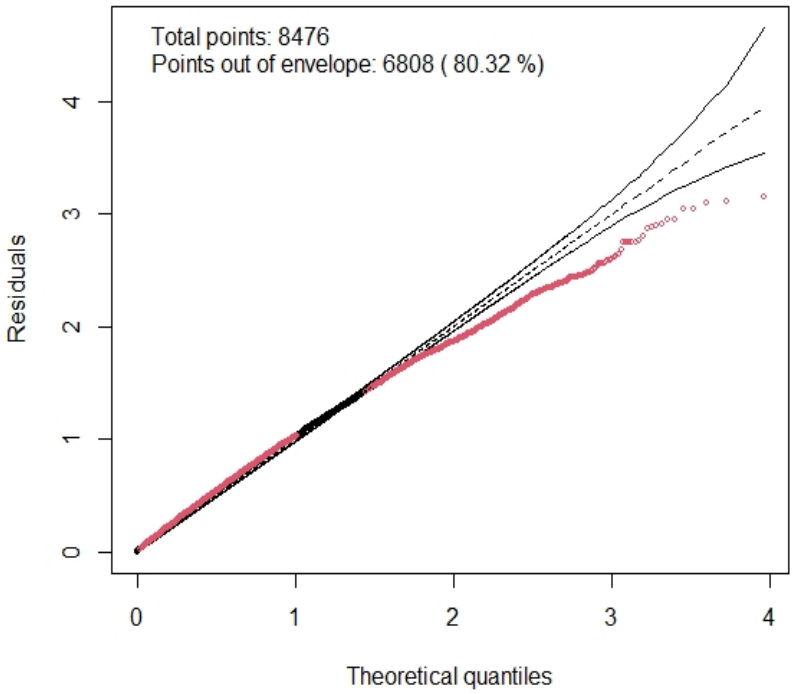
Fonte: Próprio autor.

Figura A.2: *Worm plot* do modelo ajustado para a distribuição BCTo.



Fonte: Próprio autor.

Figura A.3: *Half Normal Plot* dos resíduos do modelo.



Fonte: Próprio autor.

Tabela A.1: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,02507
Variância	0,95440
Assimetria	0,16557
Curtose	2,52776

Fonte: Próprio autor.

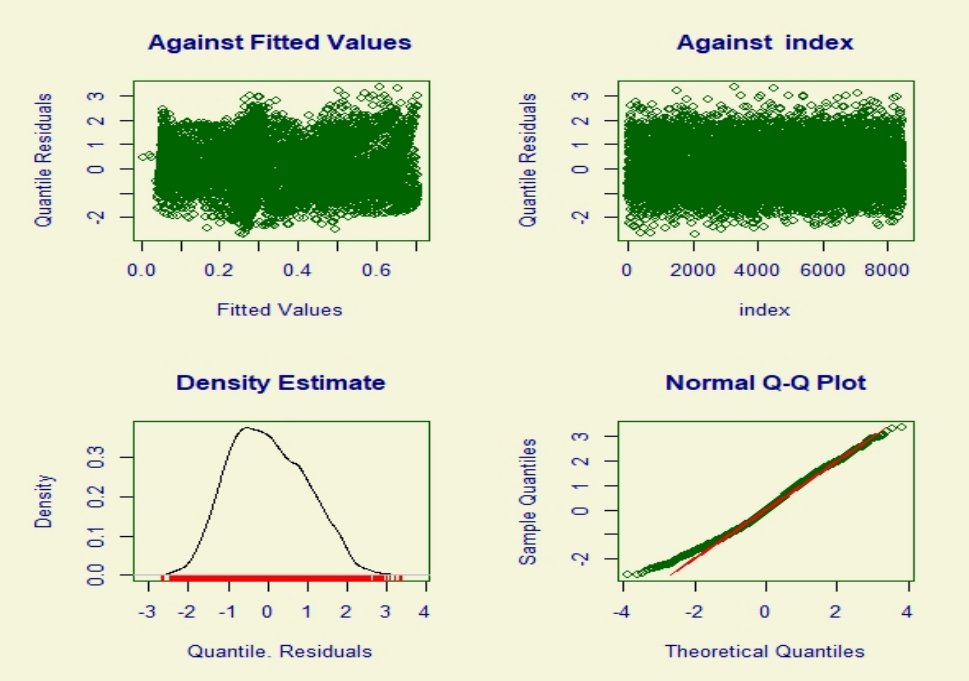
Tabela A.2: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-6900,849
BIC	-6558,402
Deviance Global	-6998,066
Graus de Liberdade do Ajuste	48,60858
Graus de Liberdade dos Resíduos	8427,391

Fonte: Próprio autor.

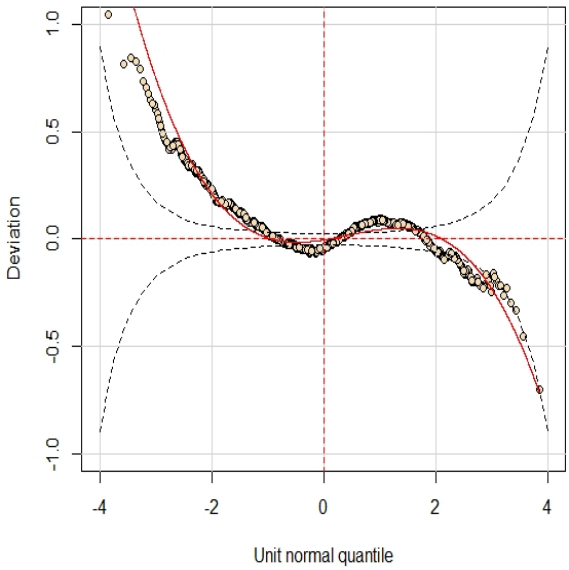
Ajuste realizado com a distribuição BCT

Figura A.4: Gráficos dos resíduos do modelo ajustado para BCT.



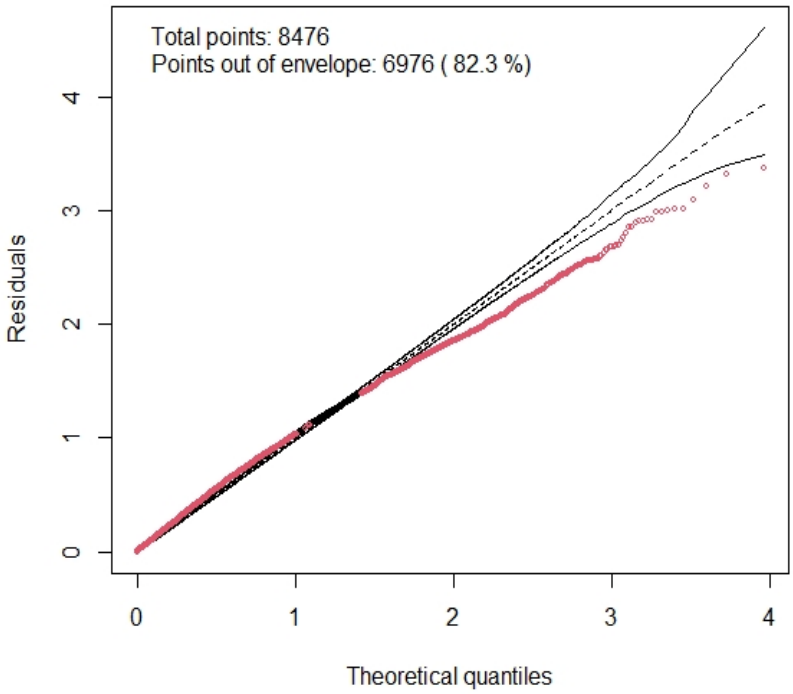
Fonte: Próprio autor.

Figura A.5: *Worm plot* do modelo ajustado para a distribuição BCT.



Fonte: Próprio autor.

Figura A.6: *Half Normal Plot* dos resíduos do modelo.



Fonte: Próprio autor.

Tabela A.3: Resumo das medidas dos resíduos quantílicos.

Medidas	**Valor**
Média	0,00876
Variância	0,95598
Assimetria	0,23935
Curtose	2,52549

Fonte: Próprio autor.

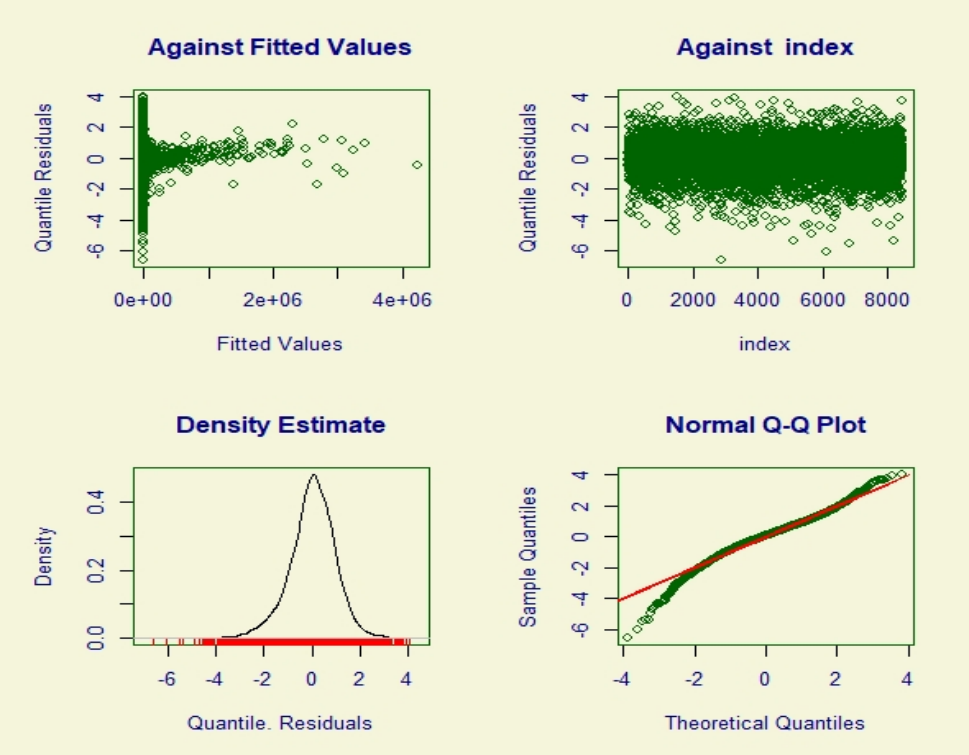
Tabela A.4: Resumo das medidas do modelo ajustado.

Métrica	**Valor**
AIC	-6546,505
BIC	-6203,202
Deviance Global	-6643,965
Graus de Liberdade do Ajuste	48,72997
Graus de Liberdade dos Resíduos	8427,27

Fonte: Próprio autor.

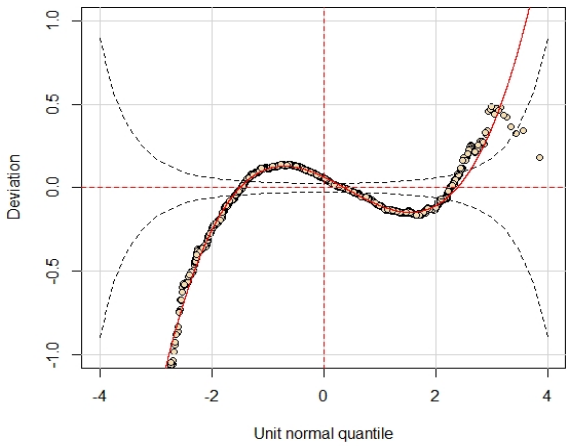
Ajuste realizado com a distribuição GB2

Figura A.7: Gráficos dos resíduos do modelo ajustado para GB2.



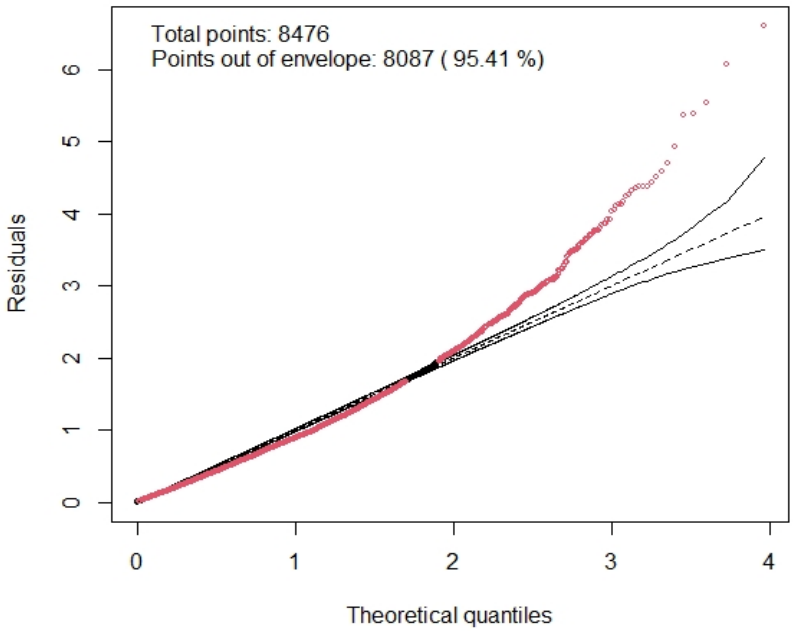
Fonte: Próprio autor.

Figura A.8: *Worm plot* do modelo ajustado para a distribuição GB2.



Fonte: Próprio autor.

Figura A.9: *Half Normal Plot* dos resíduos do modelo.



Fonte: Próprio autor.

Tabela A.5: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,00302
Variância	0,99784
Assimetria	-0,49178
Curtose	5,01525

Fonte: Próprio autor.

Tabela A.6: Resumo das medidas do modelo ajustado.

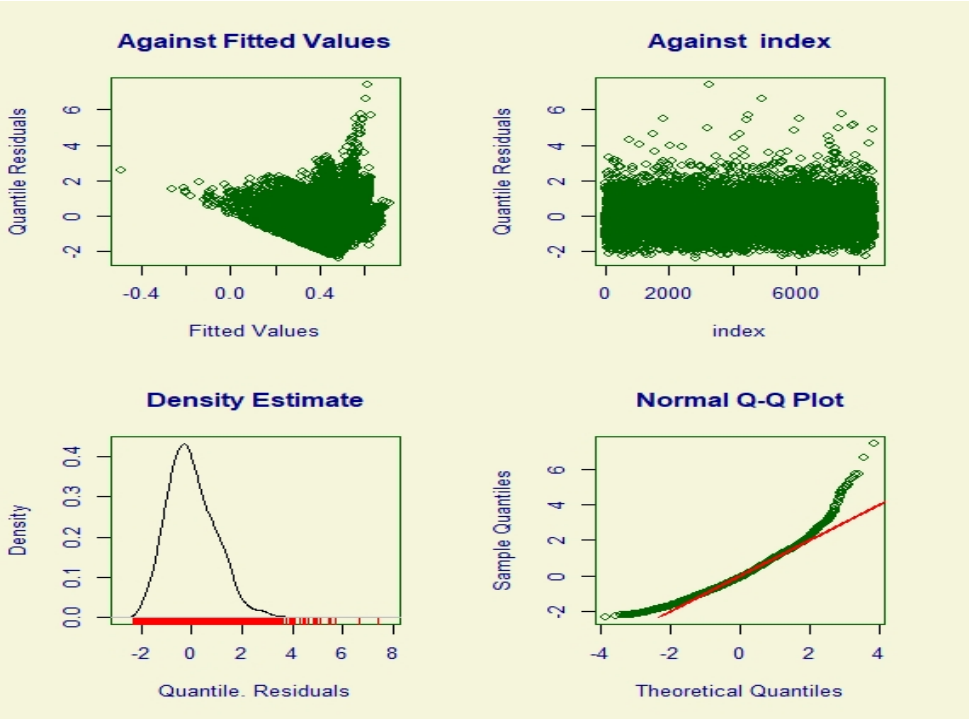
Métrica	Valor
AIC	-7923,085
BIC	-7578,53
Deviance Global	-8020,901
Graus de Liberdade do Ajuste	48,90781
Graus de Liberdade dos Resíduos	8427,092

Fonte: Próprio autor.

B APÊNDICE B: RESULTADOS OBTIDOS COM OS AJUSTES DOS 5 MODELOS DE REGRESSÃO: 1 MODELO LM, 2 MODELOS GLM E 2 MODELOS GAM.

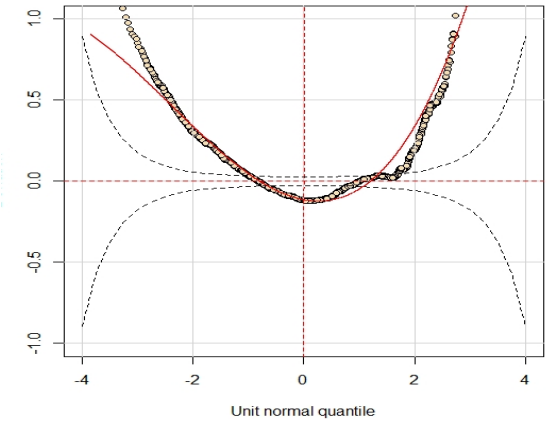
Modelo LM ajustado

Figura B.1: Gráficos dos resíduos do modelo LM ajustado.



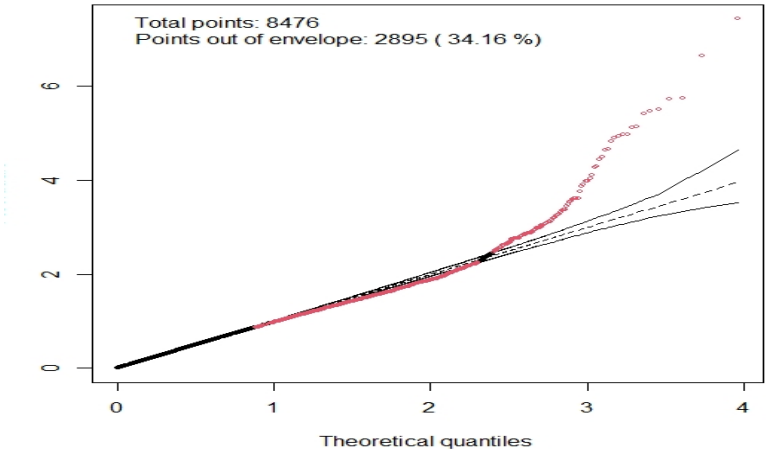
Fonte: Próprio autor.

Figura B.2: *Worm plot* do modelo LM ajustado.



Fonte: Próprio autor.

Figura B.3: *Half Normal Plot* dos resíduos do modelo LM.



Fonte: Próprio autor.

Tabela B.1: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,00
Variância	1,00
Assimetria	0,792
Curtose	4,801

Fonte: Próprio autor.

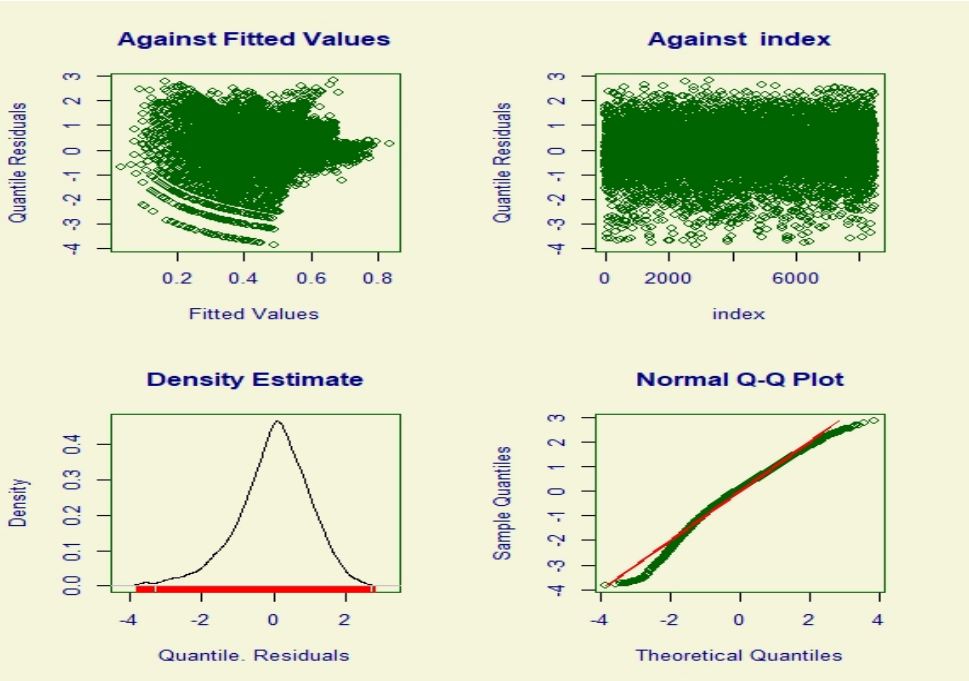
Tabela B.2: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-3447,407
BIC	-3419,227
Deviance Global	-3455,407
Graus de Liberdade do Ajuste	4
Graus de Liberdade dos Resíduos	8472

Fonte: Próprio autor.

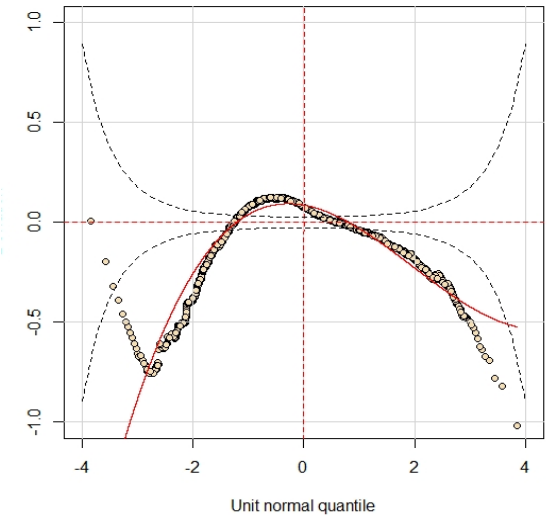
Modelo GLM GAMA

Figura B.4: Gráficos dos resíduos do modelo GLM GAMA.



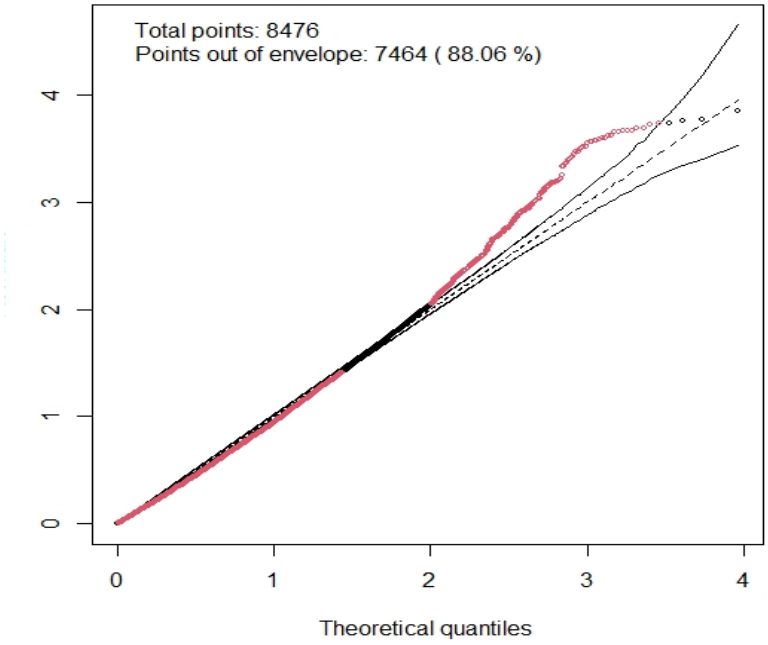
Fonte: Próprio autor.

Figura B.5: *Worm plot* do modelo GLM GAMA.



Fonte: Próprio autor.

Figura B.6: *Half Normal Plot* dos resíduos do modelo GLM GAMA.



Fonte: Próprio autor.

Tabela B.3: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,005
Variância	1,002
Assimetria	-0,539
Curtose	3,680

Fonte: Próprio autor.

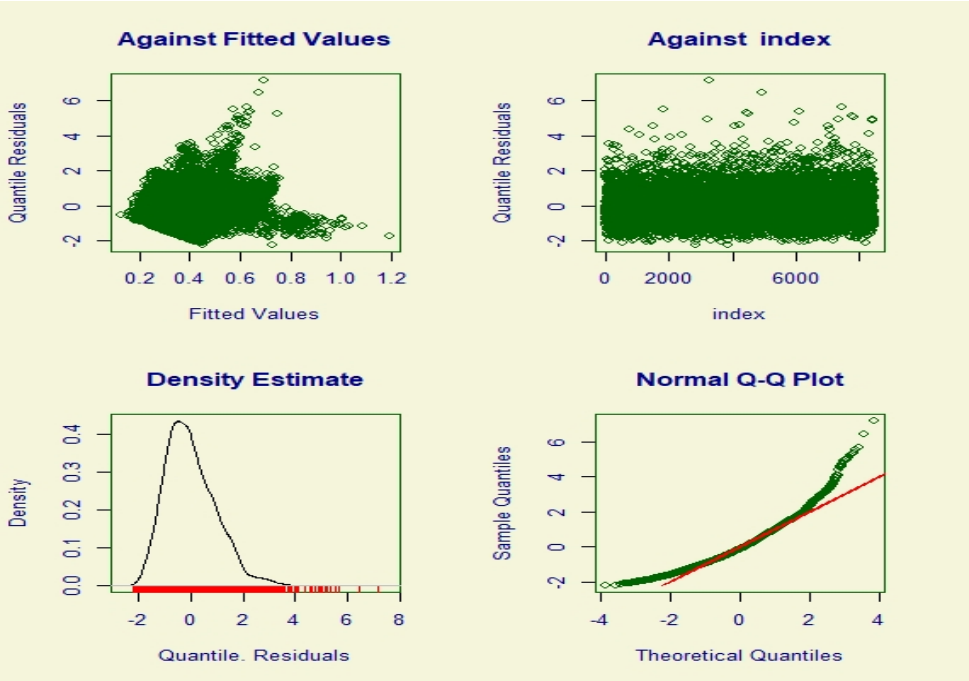
Tabela B.4: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-4063,298
BIC	-4035,118
Deviance Global	-4071,298
Graus de Liberdade do Ajuste	4
Graus de Liberdade dos Resíduos	8472

Fonte: Próprio autor.

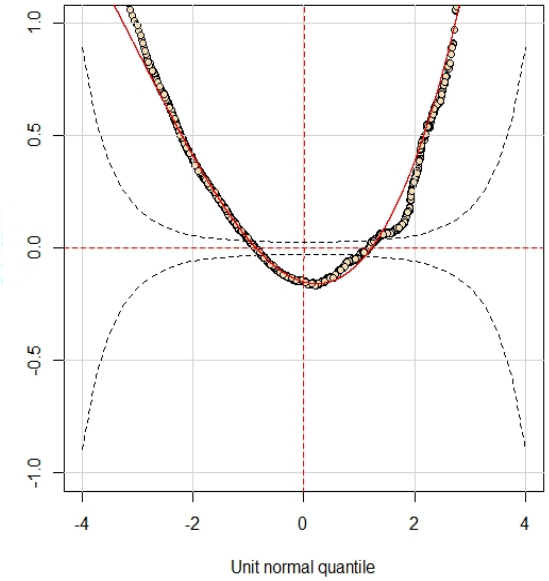
Modelo GLM NORMAL INVERSA

Figura B.7: Gráficos dos resíduos do modelo GLM NORMAL INVERSA.



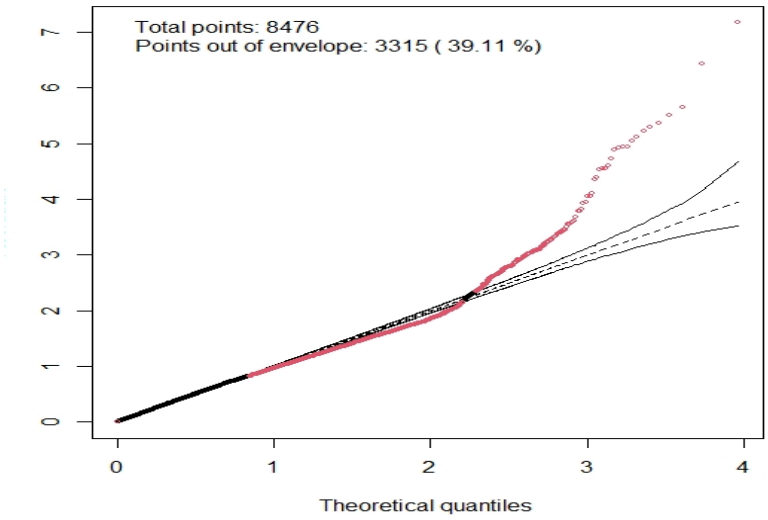
Fonte: Próprio autor.

Figura B.8: *Worm plot* do modelo GLM NORMAL INVERSA.



Fonte: Próprio autor.

Figura B.9: *Half Normal Plot* dos resíduos do modelo GLM NORMAL INVERSA.



Fonte: Próprio autor.

Tabela B.5: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	-0,012
Variância	1,000
Assimetria	0,920
Curtose	4,857

Fonte: Próprio autor.

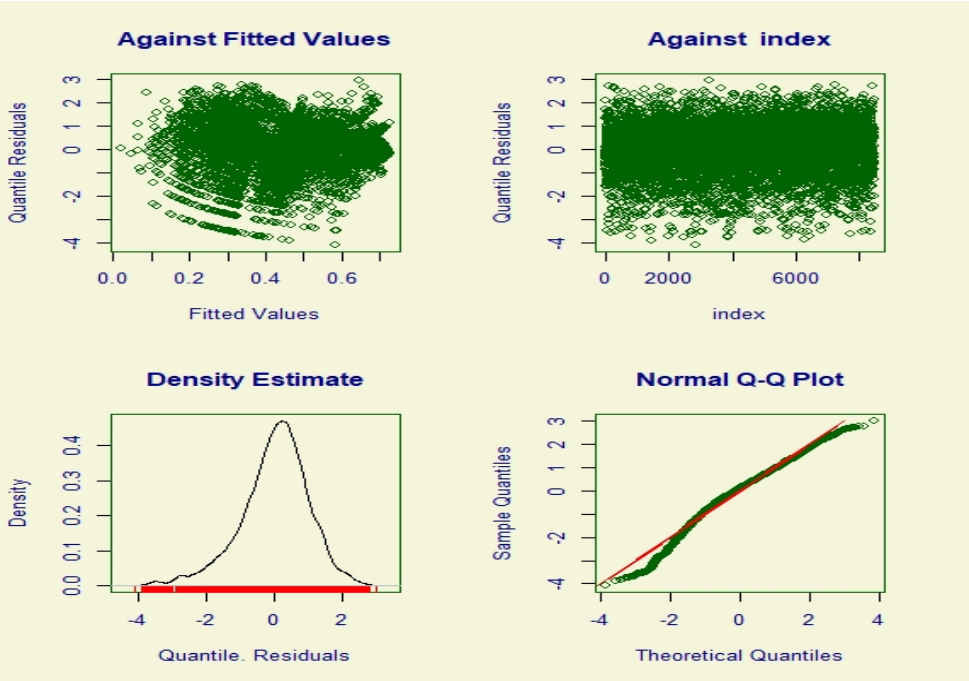
Tabela B.6: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-3775,635
BIC	-3747,455
Deviance Global	-3783,635
Graus de Liberdade do Ajuste	4
Graus de Liberdade dos Resíduos	8472

Fonte: Próprio autor.

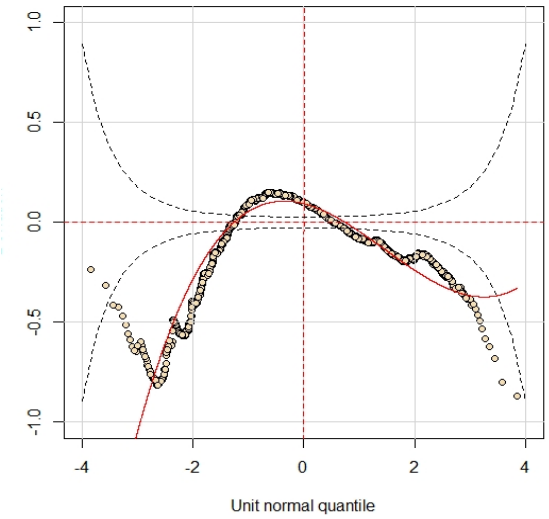
Modelo GAM GAMA

Figura B.10: Gráficos dos resíduos do modelo GAM GAMA.



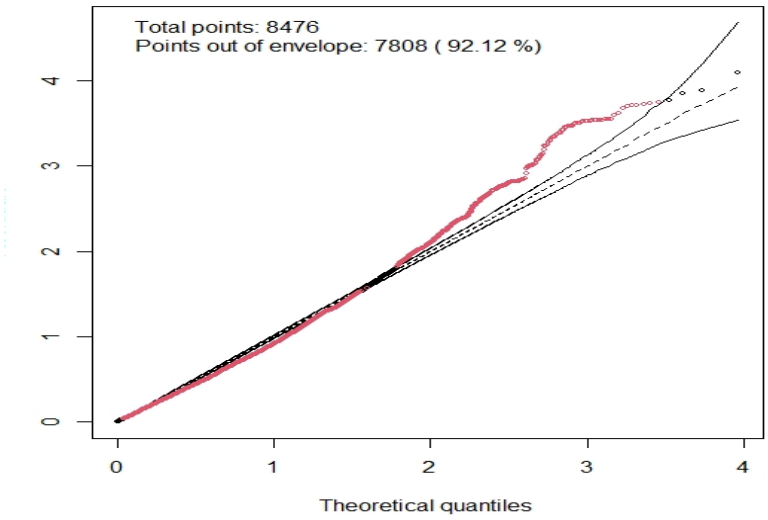
Fonte: Próprio autor.

Figura B.11: *Worm plot* do modelo GAM GAMA.



Fonte: Próprio autor.

Figura B.12: *Half Normal Plot* dos resíduos do modelo GAM GAMA.



Fonte: Próprio autor.

Tabela B.7: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,005
Variância	1,002
Assimetria	-0,597
Curtose	3,866

Fonte: Próprio autor.

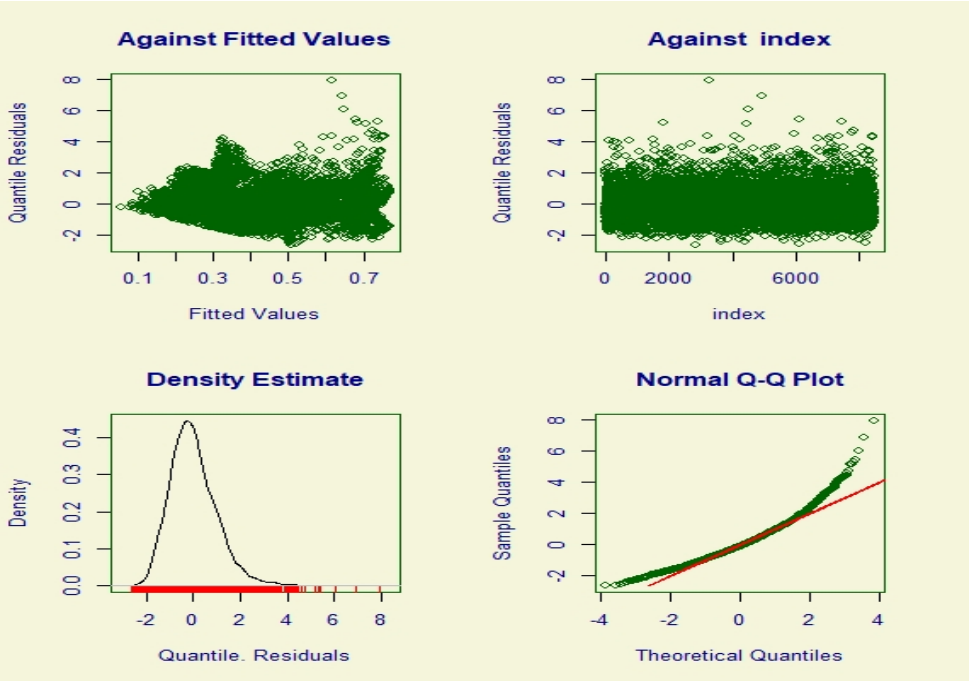
Tabela B.8: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-4425,473
BIC	-4285,306
Deviance Global	-4465,264
Graus de Liberdade do Ajuste	19,896
Graus de Liberdade dos Resíduos	8456,104

Fonte: Próprio autor.

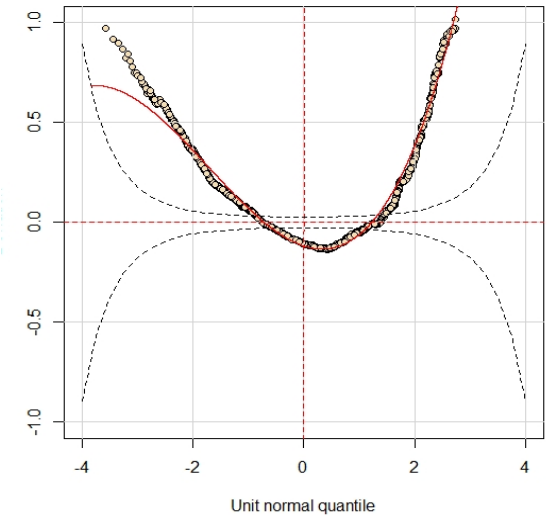
Modelo GAM NORMAL INVERSA

Figura B.13: Gráficos dos resíduos do modelo GAM NORMAL INVERSA.



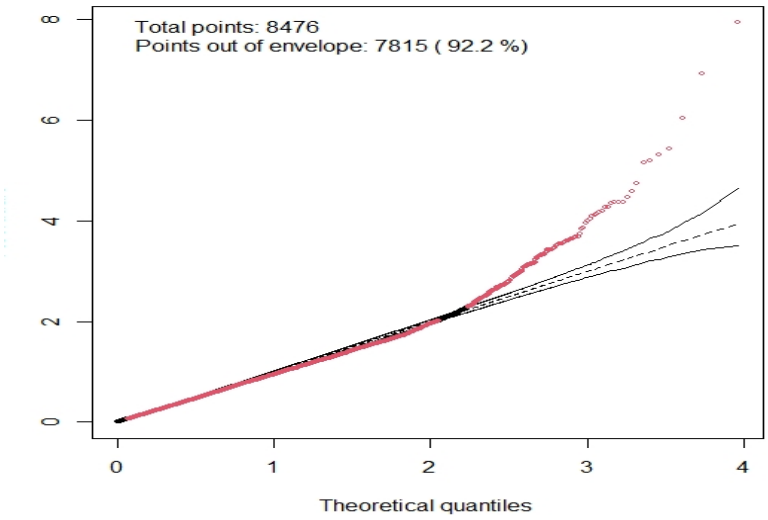
Fonte: Próprio autor.

Figura B.14: *Worm plot* do modelo GAM NORMAL INVERSA.



Fonte: Próprio autor.

Figura B.15: *Half Normal Plot* dos resíduos do modelo GAM NORMAL INVERSA.



Fonte: Próprio autor.

Tabela B.9: Resumo das medidas dos resíduos quantílicos.

Medidas	Valor
Média	0,005
Variância	1,000
Assimetria	0,871
Curtose	5,037

Fonte: Próprio autor.

Tabela B.10: Resumo das medidas do modelo ajustado.

Métrica	Valor
AIC	-4564,592
BIC	-4441,006
Deviance Global	-4599,676
Graus de Liberdade do Ajuste	17,542
Graus de Liberdade dos Resíduos	8458,458

Fonte: Próprio autor.