Modelos preditivos para LGD

João Flávio Andrade Silva

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

JOÃO FLÁVIO ANDRADE SILVA

MODELOS PREDITIVOS PARA LGD

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

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

São Carlos
Junho de 2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAM INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

JOÃO FLÁVIO ANDRADE SILVA

PREDICTIVE MODELS FOR LGD

Master dissertation submitted to the Department of Statistics – DEs-UFSCar and to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics UFSCar-USP.

Advisor: Prof. Dr. Carlos Alberto Ribeiro Diniz

São Carlos
June 2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato João Flávio Andrade Silva, realizada em 04/05/2018:

Prof. Dr. Carlos Alberto Ribeiro Diniz
UFSCar

Prof. Dr. Filidor Edilfonso Vilca Labra
UNICAMP

Prof. Dr. Marcio Luis Lanfredi Viola
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Filidor Edilfonso Vilca Labra e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Carlos Alberto Ribeiro Diniz

Dedico este trabalho à minha família,
especialmente à minha mãe Risomar e ao meu pai João.

AGRADECIMENTOS

Ao Prof. Dr. Carlos Alberto Ribeiro Diniz, pela orientação, compreensão e apoio, necessários ao longo do mestrado.

Às professoras Dra. Juliana Cobre e Dra. Daiane Aparecida Zuanetti, que fizeram parte da banca de qualificação, e aos professores Dr. Filidor Edilfonso Vilca Labra e Dr. Márcio Luis Lanfredi Viola, que participaram da banca de defesa, pelas contribuições que proporcionaram ao trabalho.

“Se as pessoas não acreditam que a matemática é simples,
é apenas porque não percebem como a vida é complicada.”
(John von Neumann)

RESUMO

SILVA, J. F. A. **Modelos preditivos para LGD**. 2018. 113 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

As instituições financeiras que pretendem utilizar a IRB (*Internal Ratings Based*) avançada precisam desenvolver métodos para estimar a componente de risco LGD (*Loss Given Default*). Desde a década de 1950 são apresentadas propostas para modelagem da PD (*Probability of default*), em contrapartida, a previsão da LGD somente recebeu maior atenção após a publicação do Acordo Basileia II. A LGD possui ainda uma literatura pequena, se comparada a PD, e não há um método eficiente em termos de acurácia e interpretação como é a regressão logística para a PD. Modelos de regressão para LGD desempenham um papel fundamental na gestão de risco das instituições financeiras. Devido sua importância este trabalho propõe uma metodologia para quantificar a componente de risco LGD. Considerando as características relatadas sobre a distribuição da LGD e na forma flexível que a distribuição beta pode assumir, propomos uma metodologia de estimação da LGD por meio do modelo de regressão beta bimodal inflacionado em zero. Desenvolvemos a distribuição beta bimodal inflacionada em zero, apresentamos algumas propriedades, incluindo momentos, definimos estimadores via máxima verossimilhança e construímos o modelo de regressão para este modelo probabilístico, apresentamos intervalos de confiança assintóticos e teste de hipóteses para este modelo, bem como critérios para seleção de modelos, realizamos um estudo de simulação para avaliar o desempenho dos estimadores de máxima verossimilhança para os parâmetros da distribuição beta bimodal inflacionada em zero. Para comparação com nossa proposta selecionamos os modelos de regressão beta e regressão beta inflacionada, que são abordagens mais usuais, e o algoritmo SVR , devido a significativa superioridade relatada em outros trabalhos.

Palavras-chave: Loss Given Default, regressão, distribuição beta bimodal inflacionada em zero, modelo de regressão beta bimodal inflacionado em zero.

ABSTRACT

SILVA, J. F. A. **Predictive models for LGD**. 2018. 113 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Financial institutions willing to use the advanced Internal Ratings Based (IRB) need to develop methods to estimate the LGD (Loss Given Default) risk component. Proposals for PD (Probability of default) modeling have been presented since the 1950s, in contrast, LGD's forecast has received more attention only after the publication of the Basel II Accord. LGD also has a small literature, compared to PD, and there is no efficient method in terms of accuracy and interpretation such as logistic regression for PD. Regression models for LGD play a key role in the risk management of financial institutions, due to their importance this work proposes a methodology to quantify the LGD risk component. Considering the characteristics reported on the distribution of LGD and in the flexible form that the beta distribution may assume, we propose a methodology for estimation of LGD using the zero inflated bimodal beta regression model. We developed the zero inflated bimodal beta distribution, presented some properties, including moments, defined estimators via maximum likelihood and constructed the regression model for this probabilistic model, presented asymptotic confidence intervals and hypothesis test for this model, as well as selection criteria of models, we performed a simulation study to evaluate the performance of the maximum likelihood estimators for the parameters of the zero inflated bimodal beta distribution. For comparison with our proposal we selected the beta regression models and inflated beta regression, which are more usual approaches, and the SVR algorithm, due to the significant superiority reported in other studies.

Keywords: Loss Given Default, regression, zero inflated bimodal beta distribution, zero inflated bimodal beta regression model.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE CÓDIGOS-FONTE

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O ato de emprestar provavelmente surgiu no momento em que os seres humanos começaram a se comunicar. Há um documento datado por volta de 2000 a.C. que registra um acordo de empréstimo com cobrança de juros. De fato, o desenvolvimento do crédito somente ocorreu após 1350 d.C., na Europa, com as lojas de penhores. Estas lojas cobravam juros sobre os empréstimos e aceitavam como garantia praticamente qualquer coisa, desde que pudesse ser guardado. Em 1920, com a comercialização do veículo automotor surgiu uma demanda de empréstimos para aquisição deste bem, porém devido a sua mobilidade não poderia configurar como penhora, nem como garantia fixa - a qual o credor conhece sua localização. Desse modo, empresas financeiras foram criadas para atender tal demanda. E no decorrer da última metade do século passado, o crédito pessoal apresentou vultuoso crescimento, marcado pela criação do cartão de crédito (Anderson, 2007).

Atualmente as instituições financeiras exercem diversas atividades, dentre as quais a intermediação financeira[1] possui um dos papéis mais relevantes para o mercado. Na intermediação financeira as instituições financeiras estão sujeitas a diversos riscos e incertezas. Um deles é o risco de crédito, definido pelo BACEN como:

> [...] a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação (BACEN, 2009).

O risco de crédito divide-se em subcategorias: risco de contraparte, risco de concentração, risco país, risco *commitment*, risco de intermediadoras ou convenentes, entre outras. Dentre

[1] A intermediação financeira é definida como a atividade de captar de agentes superavitários e emprestar a agentes deficitários.

estas subcategorias destaca-se o risco de contraparte, que está associada à possibilidade de descumprimento das obrigações financeiras pela contraparte (BACEN, 2009).

A preocupação com os riscos que as instituições financeiras estão sujeitas não é exclusivo da atualidade, mas ela ganhou força com a expansão do crédito, que provocou aumento da exposição ao risco de crédito das instituições.

No ano seguinte após o fim, ocorrido no ano de 1973, do Sistema Monetário Internacional baseado em taxas de câmbio fixas, o mercado financeiro apresentava graves sinais de fragilidade. Com o objetivo de fortalecer a estabilidade do sistema bancário internacional os responsáveis pela supervisão bancária nos países do G-10 criaram o Comitê de Basileia. E em 1988, com o objetivo de reforçar a estabilidade do sistema bancário internacional e minimizar desigualdades competitivas entre os bancos internacionais foi celebrado o primeiro Acordo de Basileia, que padronizou a aplicação de Fatores de Ponderação de Risco (FPR) aos ativos e a exigência de capital mínimo. Em 1996, foi realizado uma Emenda de Risco de Mercado, que apresentou alguns ajustes no Acordo de Basileia I, os mais relevantes foram: alocação de capital para cobertura de Riscos de Mercado, ampliação dos controles sobre riscos incorridos pelos bancos, possibilidade de utilização de modelos internos de mensuração de riscos, desde que aprovados pelo regulador local (BANCO DO BRASIL, 2008).

Ao longo do tempo foi observado que as medidas tomadas não haviam sido suficientes para impedir que os bancos ficassem expostos a determinados riscos, assim o Comitê, em 2004, divulgou o Acordo de Basileia II, em que estabeleceu novos critérios de requerimento de capital regulamentar, considerando os riscos associados às exposições, governança e transparência das instituições financeiras (BANCO DO BRASIL, 2008).

O acordo de Basileia II permitiu às instituições financeiras calcular o capital exigido pelo risco de crédito com uma abordagem baseada em classificações internas – IRB (*Internal Ratings Based*). Na apuração do requerimento de capital as instituições, que fazem uso da IRB, deverão considerar os seguintes componentes de risco (BACEN, 2013):

- Probabilidade de Descumprimento - PD (*Probability of Default*) – medida que representa a expectativa de longo prazo das taxas de descumprimento;

- Exposição no Momento do Descumprimento – EAD (*Exposure at Default*) – valor de exposição da instituição no momento de ocorrência do descumprimento;

- Perda Dado o Descumprimento - LGD (*Loss Given Default*) – corresponde ao percentual, em relação a EAD observada, da perda econômica decorrente do descumprimento;

- Prazo Efetivo de Vencimento - M (*Effective Maturity*) - é o prazo até o vencimento da operação.

Existem duas abordagens para avaliar IRB, a básica e a avançada. A diferença entre elas está na estimação dos componentes de risco. Na abordagem básica, deve ser calculado o valor do componente M (Prazo Efetivo de Vencimento) e apenas a PD estimada internamente, os demais componentes são atribuídos pela autoridade de supervisão, já na abordagem avançada, a instituição financeira deve calcular o parâmetro M e estimar todos os componentes de risco internamente (BANCO DO BRASIL, 2008).

Portanto, as instituições financeiras que pretendem utilizar a IRB avançada precisam desenvolver métodos para estimar os componentes de risco LGD e EAD para cada segmento de sua carteira, além da PD. Desde a década de 1950 são apresentadas propostas para modelagem da PD, em contrapartida, a previsão da LGD somente recebeu maior atenção após a publicação do acordo Basileia II. A LGD possui ainda uma literatura pequena, se comparada a PD, e não há um método eficiente em termos de acurácia e interpretação como é a regressão logística para a PD. Neste sentido esta pesquisa aborda a modelagem do componente de risco LGD.

A LGD é calculada quando o empréstimo está inadimplente. No contexto de risco de crédito, conseguir uma definição prática para a inadimplência, ou descumprimento[2], não é tarefa simples, apesar da inadimplência ser definida de forma única como a falta de cumprimento de um contrato ou de qualquer de suas condições (Buarque, 1993). Em partes isto é devido a existência de conflito de interesses por parte dos analistas de crédito (Annibal, 2009).

Porém, no âmbito das instituições financeiras que utilizam a abordagem IRB, o BACEN (2013) define o descumprimento pela ocorrência de pelo menos um dos seguintes eventos:

- a instituição considera que o devedor não irá honrar integralmente suas obrigações sem que a instituição recorra a ações tais como acionamento de garantias;

- a obrigação esteja em atraso superior a uma certa quantidade de dias, que dependendo do tipo de exposição ao risco pode ser de 90 ou 180 dias.

Segundo Thomas, Edelman e Crook (2002), a inadimplência pode ocorrer devido a diversos fatores, e a ocorrência deste evento não implica necessariamente que o cliente é ruim. A maioria dos casos de inadimplência podem ser separados em duas categorias: motivos relacionados a eventos inesperados que impedem o pagamento como está no contrato e razões relacionadas com a vontade do cliente em realizar o pagamento, provenientes do caráter.

A LGD ou perda dado o descumprimento é equivalente a $1 - R$, em que R representa a taxa de recuperação. De forma simplista, a LGD pode ser compreendida como o percentual, em relação ao valor em exposição, da perda dado a ocorrência de inadimplência. Schuermann (2004) acrescenta que, uma vez ocorrido o descumprimento, a LGD inclui três tipos de perdas:

[2] Descumprimento será considerado sinônimo de inadimplência, embora exista a possibilidade dos conceitos diferir em algum aspecto.

perda do principal; perda decorrente de custos de empréstimos não pagos, incluindo custo de oportunidade; e despesas relacionadas ao processo de cobrança e recuperação do crédito.

O BACEN pronuncia a respeito da definição de LGD:

> Perda Dado o Descumprimento (LGD), que corresponde ao percentual, em relação ao parâmetro EAD observado, da perda econômica decorrente do descumprimento, considerados todos os fatores relevantes, inclusive descontos concedidos para recuperação do crédito e todos os custos diretos e indiretos associados à cobrança da obrigação (BACEN, 2013).

De acordo com Schuermann (2004), existem três formas para calcular a LGD de um contrato:

- *Market* LGD: baseada nos preços de mercado de títulos inadimplentes ou empréstimos negociáveis logo após a inadimplência;
- *Workout* LGD: com base no conjunto de fluxos de caixa estimados resultantes do processo de cobrança, propriamente descontado, e a exposição estimada;
- *Implied market* LGD: derivado dos preços de títulos adimplentes com risco utilizando um modelo teórico de precificação de ativos.

Silva, Marins e Neves (2009) relatam uma quarta forma de calcular a LGD, chamada *Implied historical* LGD, a qual utiliza dados históricos de recuperação e estimativas de probabilidade de inadimplência. Também mencionam que *Workout* LGD é a forma mais utilizada pela indústria, dentre as quatro apresentadas.

Segundo Zohova (2015), o cálculo utilizado para obter a LGD individual sob a abordagem *Workout* é, geralmente, da seguinte forma

$$\mathrm{LGD}_i(t) = 1 - \frac{\sum_{j=1}^{t} \mathrm{PV}(\mathrm{CF}_{ij})}{\mathrm{EAD}_i},$$

em que $\mathrm{LGD}_i(t)$: LGD de uma exposição inadimplente i no tempo t após a inadimplência; EAD_i: saldo da exposição i no momento da inadimplência; PV: função de desconto para trazer a valor presente; CF_{ij}: fluxo de caixa resultante do processo de workout da exposição i registrada no tempo t.

De acordo com Fernandes e Tomazella (2013), o parâmetro de risco LGD, geralmente, possui domínio no intervalo [0,1], entretanto eventuais valores fora deste intervalo poderão ser observados. Porém, as instituições costumam tomar uma posição conservadora e atribuir LGD nula para as observações negativas. Desta forma, esta LGD conservadora pode ser expressa por

$$\mathrm{LGD}_i^{+} = \max\{0, \mathrm{LGD}_i\}.$$

Com a utilização da abordagem IRB avançada, existem duas perspectivas de interpretação da LGD, cíclico e acíclico. Sob ponto de vista cíclico, a LGD estará sincronizada com as mudanças de ciclos econômicos, neste caso, a metodologia utilizada para estimação da LGD é conhecida como point-in-time (PIT), em contrapartida, sob a perspectiva acíclica, a LGD permaneceria constante ao longo do tempo, neste caso, a metodologia utilizada é a through-the-cicle (TTC) (Silva, Marins e Neves, 2009).

Diversos estudos empíricos foram realizados com a intenção de caracterizar a distribuição da perda dada inadimplência e obter relações com determinados fatores. Nestes estudos, algumas características são frequentemente documentadas.

Yao, Crook e Andreeva (2015) informam que, geralmente, a LGD apresenta distribuição bimodal e variação no intervalo fechado [0,1]. Schuermann (2004) concorda e acrescenta que as observações costumam se concentrar nos extremos do intervalo, indicando que ocorrem muitas perdas altas e baixas, e devido a este fato, utilizar o conceito de LGD média pode causar interpretações errôneas.

Entretanto, Fernandes e Tomazella (2013) relatam que a LGD pode apresentar distribuição unimodal e até trimodal, e que também podem ser observadas perdas negativas e acima de 1, e que, em geral, a observação de perda maior que 1 está ligada a um elevado custo do processo de cobrança. Contam que o processo de cobrança e as características da operação influenciam a distribuição da LGD, e sintetizam as características das distribuições, geralmente, observadas. Na Figura 1 são apresentadas possíveis distribuições de proporções para observações de LGD. A forma “Bimodal"apresentando concentrações em zero e um é a mais comum. A distribuição das perdas de operações que possuem garantia forte, como por exemplo os financiamentos imobiliários, geralmente, apresentam forma semelhante a “Unimodal", com perdas concentradas ao redor de zero. A forma “Trimodal", representa a distribuição da LGD de uma operação com garantia de rápida depreciação, por exemplo os financiamentos de automóveis. A distribuição “Acima de 1"representa observações de contratos que o processo de cobrança é caro, resultando em LGD acima de 1 (Fernandes e Tomazella, 2013).

Qi e Yang (2009) mostram que a perda dada a inadimplência, em seu estudo, foi amplamente explicada por várias características associadas ao empréstimo, e a covariável que apresentou maior impacto foi a razão entre o valor do empréstimo e o valor da garantia.

Castle e Keisman (1999) aplicaram análise de regressão em um banco de dados da Standard and Poor’s Credit Loss, e verificaram que o tipo de garantia é uma característica importante para a previsão da LGD.

Alguns estudos citam correlação entre LGD e PD. Para Altman e Kalotay (2014), as recuperações tendem a ser mais baixas em períodos econômicos que apresentam altos índices de inadimplência. Frye (2000) utilizou uma base de dados da Moody’s Default Risk Service e observou, que numa recessão econômica hipotética com uma taxa de inadimplência de 10%,

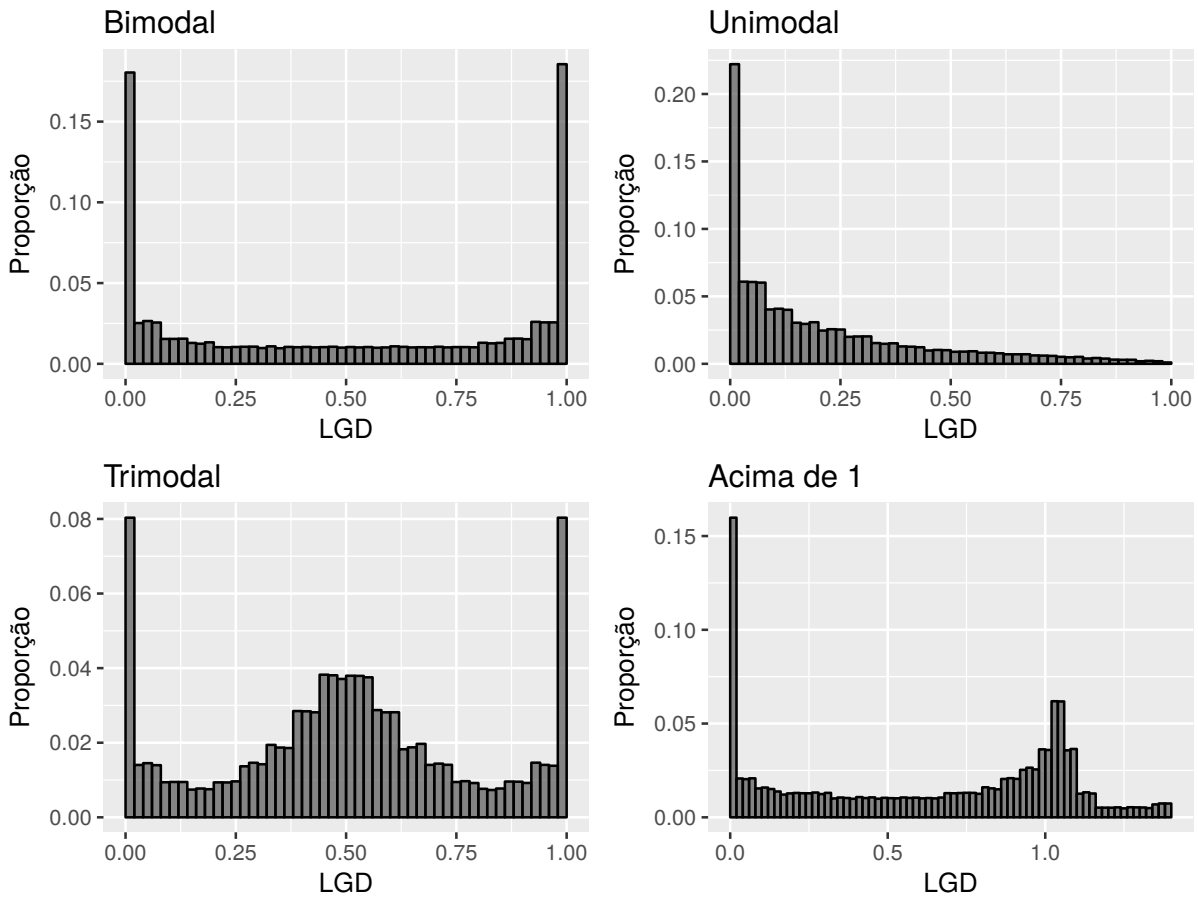

Figura 1 – Possíveis histogramas para dados de LGD.

as recuperações de obrigações podem cair de 20 a 25 pontos percentuais com relação à média normal do ano.

A respeito das abordagens utilizadas para modelar a LGD, Altman e Kalotay (2014) recomendam cautela ao utilizar a regressão linear e a regressão beta, que apesar de fornecerem interpretações simples, fazem fortes suposições sobre os dados. Yao, Crook e Andreeva (2015) comentam que devido a LGD apresentar frequentemente distribuição empírica bimodal e delimitada no intervalo [0,1], um modelo de regressão linear pode se encaixar mal. Servigny e Renault (2004) observam que, apesar da distribuição beta apresentar forma flexível, esta não comporta bimodalidade e também não está definida para os pontos 0 e 1.

Para representar a alta concentração nos pontos 0 e 1, Calabrese (2014) aplicou a regressão beta inflacionada[3]. Seu estudo mostrou que a predição do modelo de regressão beta inflacionada superou o modelo de regressão linear e o modelo de regressão de resposta fracionada, em termos de erro quadrático médio e desvio médio absoluto.

[3] Modelagem de dados em que a variável resposta é medida no intervalo [0,1], este modelo assume que a distribuição dos dados é uma mistura entre uma distribuição contínua definida no intervalo (0,1), distribuição beta, e uma distribuição que atribui massa de probabilidade aos pontos {0,1}, distribuição Bernoulli (Martinez, 2008).

Trabalhos recentes têm comparado o desempenho de técnicas paramétricas com não paramétricas para modelagem da LGD. Segundo Yao, Crook e Andreeva (2015), a flexibilidade é uma das principais vantagens dos métodos não paramétricos, uma vez que não assumem distribuição específica.

Com base em um banco de dados de perda de empréstimos bancários portugueses, Bastos (2010) comparou o desempenho de árvores de regressão, método não paramétrico, com a regressão de reposta fracionada e concluiu que as árvores de regressão produzem melhores resultados em períodos de tempo mais curtos, 12 a 24 meses. Qi e Zhao (2011) comparam seis diferentes métodos para modelar a LGD e afirmam que os métodos não paramétricos (redes neurais e árvores de decisão) superam os métodos paramétricos (regressão de resposta fracionada, regressão linear - transformação normal inversa, transformação beta e sem transformação da variável resposta) em termos de ajuste do modelo e poder preditivo, e concluem que esta vantagem decorre da capacidade dos métodos não paramétricos de acomodar relações não lineares entre variável resposta e covariáveis contínuas. Bastos (2010) e Qi e Zhao (2011) concordam que, apesar de apresentarem desempenho preditivo superior, a interpretação dos resultados dos métodos não paramétricos é mais complexa.

Yao, Crook e Andreeva (2015) propuseram algumas modificações na regressão de vetores suporte mínimos quadrados (LS-SVR, do inglês: *Least Square Support Vector Regression*) com a intenção de aumentar a precisão de predição da LGD, compararam treze técnicas, em que 9 delas correspondem a 3 modelos LS-SVR combinados com 2 transformações (beta e *logit*) e sem transformação na variável resposta, as demais foram: regressão linear, regressão linear com transformação beta, regressão de resposta fracionada e um modelo de dois estágios. Relatam que os modelos LS-SVR demonstraram superioridade comparados as demais técnicas e que as transformações na variável resposta não apresentaram melhorias.

Altman e Kalotay (2014), com objetivo de obter um modelo de fácil interpretação e quase tão flexível quanto um não paramétrico, apresentaram uma abordagem semi-paramétrica, utilizando misturas de distribuições gaussianas. Relatam que este modelo é flexível o suficiente para acomodar as características da distribuição da LGD e ao mesmo tempo permite estabelecer relação com as covariáveis, e que, no estudo, superou os modelos paramétricos e também um modelo não paramétrico, árvores de regressão.

Considerando as características relatadas da distribuição da LGD, na forma flexível que a distribuição beta pode assumir e na abordagem de Altman e Kalotay (2014), que utiliza mistura de distribuições normais, propomos um modelo para estimação da LGD por mistura de duas distribuições beta e uma degenerada em zero. Para comparação com nossa proposta selecionamos os modelos de regressão beta e regressão beta inflacionada, que são abordagens mais usuais, e a SVR (*Support Vector Regression Machines*), devido a significativa superioridade relatada por Yao, Crook e Andreeva (2015).

Este trabalho está organizado em 5 capítulos, incluindo o presente Capítulo 1 com o

objetivo de introduzir, contextualizar e justificar o trabalho. No Capítulo 2 são apresentados os modelos de regressão beta, beta inflacionado em zero e o *support vector regression machines*, bem como algumas vantagens e desvantagens no uso de cada modelo. No Capítulo 3 desenvolvemos a distribuição beta bimodal inflacionada em zero, apresentamos algumas propriedades, incluindo momentos, definimos estimadores via máxima verossimilhança e construímos o modelo de regressão para este modelo probabilístico. No Capítulo 4 realizamos estudo de simulação para avaliar o desempenho dos estimadores de máxima verossimilhança para os parâmetros da distribuição beta bimodal inflacionada em zero e do modelo de regressão para este modelo probabilístico, realizamos também comparação entre os modelos descritos no Capítulo 2 e o modelo proposto no Capítulo 3, avaliados em uma base de dados de LGD simulada. No Capítulo 5 apresentamos as considerações finais para este trabalho.

CAPÍTULO

2

MODELOS DE REGRESSÃO E ALGORITMOS DE APRENDIZADO DE MÁQUINA

Segundo Breiman (2001), existem dois objetivos na análise de dados: predição e informação. A predição diz respeito à capacidade de predizer quais serão os valores das variáveis respostas dado futuras covariáveis. Já informação, ligado ao conceito de inferência, está relacionado a extrair informações a partir de uma associação entre variável resposta e covariáveis. Enquanto o objetivo preditivista se importa apenas com o quão bom é o poder preditivo, o inferencial procura responder questões como: Quais covariáveis são relevantes? Qual a relação entre a covariável e a variável resposta? Ao alterar o valor de uma covariável qual o impacto observado na variável resposta? Em alguns problemas práticos estes objetivos são bem definidos, isto é, é exclusivamente inferencial ou preditivo, mas em outros casos há a necessidade de unir predição e inferência. No problema de estimação da LGD, o objetivo é tanto inferencial quanto preditivo, inferencial devido às instituições financeiras desejarem conhecer quais covariáveis são importantes, como elas se relacionam com a LGD, e assim desenvolver ou alterar políticas com a intenção de reduzir sua exposição ao risco de perdas, e preditivo consequente do anseio das instituições em avaliar o seu risco esperado em um dado tempo a frente.

Os itens deste capítulo tratam de 3 diferentes abordagens, as duas primeiras - regressão beta e regressão beta inflacionado - possuem motivação inferencial e a última possui objetivo principal de predição, o *support vector regression machines*. Estas técnicas serão comparadas no Capítulo 4 com o modelo beta bimodal inflacionado em zero, proposto no Capítulo 3.

2.1 Regressão beta

Dentre os modelos de regressão, o modelo linear é, provavelmente, o mais explorado para estabelecer uma relação entre variável resposta e covariáveis. Entretanto, quando a variável resposta assume valores no intervalo $(0, 1)$, este modelo se torna inadequado devido a possibilitar

predições fora do intervalo. Uma alternativa para este problema é a transformação da variável resposta de modo que a nova variável assuma valores no conjunto dos reais. Em um relatório que descreve e documenta o modelo LossCalc, modelo da Moody's para prever a LGD, Gupton *et al.* (2002) observaram que as taxas de recuperações eram aproximadamente beta distribuídas, assim, com a intenção de obter uma variável Y aproximadamente normal propuseram a seguinte transformação na variável taxa de recuperação (RR), $Y = \Phi^{-1}(F_{\hat{p},\hat{q}}(RR))$, onde Φ é a função distribuição normal padrão e $F_{p,q}$ é a função distribuição $Beta(p,q)$.

A respeito da transformação dos dados Cribari-Neto e Zeileis (2010) relatam que esta abordagem apresenta algumas deficiências, uma delas está relacionada com as interpretações dos parâmetros que são obtidas em termos da média da variável transformada ao invés da variável original, e geralmente regressões envolvendo variável resposta restrita em um intervalo apresentam heterocedasticidade e assimetria, ambas condições não acomodadas pelo modelo de regressão linear.

Wong e Lai (2008) afirmam que uma melhor abordagem, em termos de predição, para estimar a LGD é utilizar o modelo de regressão beta. Proposto por Ferrari e Cribari-Neto (2004), o modelo de regressão beta representa uma alternativa para estabelecer uma relação entre uma variável contínua no intervalo restrito $(0,1)$ e uma ou mais covariáveis. Tal proposta é fundamentada na aplicação da distribuição beta, a qual é bastante flexível quanto a sua forma. A modelagem da variável de interesse é realizada por meio de uma estrutura de regressão com uma função de ligação, parâmetros e covariáveis.

A função de densidade de uma variável aleatória Y com distribuição de probabilidade contínua beta, com parâmetros p e q, é dada por

$$f_Y(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \tag{2.1}$$

em que $p,q > 0$, $y \in (0,1)$ e $\Gamma(\cdot)$[1] é a função gama. A média e variância de Y são dadas, respectivamente, por $E[Y] = p/(p+q)$ e $Var[Y] = pq/[(p+q)^2(p+q+1)]$. Os parâmetros p e q são responsáveis por definir a forma que a distribuição (2.1) assume.

No modelo de regressão proposto por Ferrari e Cribari-Neto (2004) é utilizado uma reparametrização da densidade beta, com $\mu = p/(p+q)$ e $\phi = p+q$, portanto, temos que $p = \mu\phi$ e $q = (1-\mu)\phi$. Segue que a densidade de Y dada por (2.1) após a reparametrização é expressa da seguinte forma:

$$f_Y(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{2.2}$$

com $y,\mu \in (0,1)$ e $\phi > 0$. Desta forma as expressões da média e da variância de Y em termos de

[1] A função gama é uma extensão da função fatorial para o conjunto dos números reais e complexos e definida por $\Gamma(p) = \int_0^\infty y^{p-1}e^{-y}dy$, $p > 0$.

μ e ϕ são dadas por:

$$E[Y] = \mu,$$
$$Var[Y] = \frac{\mu(1-\mu)}{1+\phi} = \frac{V(\mu)}{1+\phi},$$

em que $V(\mu)$ é denominada função de variância e ϕ pode ser interpretado como um parâmetro de dispersão da distribuição ou um parâmetro de precisão, isto quer dizer que ao aumentar o valor de ϕ menor será a variância de Y considerando μ fixo.

Considere $Y_1, ..., Y_n$ variáveis aleatórias independentes, em que cada Y_i, $i = 1, ..., n$ possui função densidade da forma (2.2) com média μ_i e precisão ϕ, desconhecidos. O modelo de regressão beta pode ser expresso como

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i,$$

em que, η_i é denominado preditor linear, $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^T$ é o vetor de parâmetros (desconhecidos) de interesse do modelo e $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})^T$ é o i-ésimo vetor das observações de d $(d < n)$ covariáveis de valores fixos e conhecidos e $g(\cdot)$, denominada função de ligação, é estritamente monótona e duas vezes diferenciável, com domínio em $(0, 1)$ e imagem em $\mathbb{R}$.

O método da máxima verossimilhança é utilizado para obter as estimações dos parâmetros, $\boldsymbol{\beta}$ e ϕ. A função de verossimilhança para $\boldsymbol{\beta}$ e ϕ, considerando uma amostra aleatória da distribuição beta, é

$$L(\boldsymbol{\beta}, \phi; \boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \mu_i, \phi), \tag{2.3}$$

em que $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ é o vetor com as observações. Observe que $L(\boldsymbol{\beta}, \phi; \boldsymbol{y})$ é função dos parâmetros, e os dados são quantidades fixas, posto que foram observados, sendo assim para cada valor dos parâmetros a verossimilhança representa uma medida de plausibilidade com a amostra observada. As estimativas de máxima verossimilhança para os parâmetros são obtidas através da maximização da função (2.3).

2.2 Regressão beta inflacionado em zero

Na prática, frequentemente observamos LGD com presença de zeros e/ou uns, isto é, com probabilidade nestes pontos. Nestes casos aplicar um modelo de regressão com variável resposta variando no intervalo aberto (0,1), como o modelo de regressão beta, pode não ser uma boa escolha.

Para representar as altas concentrações das taxas de recuperações nos extremos 0 e 1, Calabrese (2014) propôs considerar as recuperações uma variável aleatória proveniente de uma mistura entre as distribuições Bernoulli e beta.

Martinez (2008) propôs alternativas para modelagem de dados em que a variável resposta pode variar nos intervalos [0,1], [0,1) e (0,1], cujos modelos assumem que os dados são

originários de uma distribuição de mistura entre uma beta e uma Bernoulli, degenerada em zero e degenerada em 1, respectivamente. Ele observa que estas inflações podem estar relacionados com alguma intervenção, truncamento ou censura nos dados.

Conforme a proposta de Martinez (2008) para variável resposta no intervalo $[0,1)$, a função densidade de uma variável aleatória Y com distribuição de probabilidade beta inflacionada em zero (BIZ), com parâmetros α, μ e ϕ, é definida por

$$biz_Y(y;\alpha,\mu,\phi) = \begin{cases} \alpha, & \text{se } y = 0, \\ (1-\alpha) f_Y(y;\mu,\phi), & \text{se } y \in (0,1), \end{cases} \tag{2.4}$$

com $y \in \{0\} \cup (0,1)$, parâmetros $\alpha, \mu \in (0,1)$, $\phi > 0$ e $f_Y(y;\mu,\phi)$ é a função de densidade da distribuição beta (2.2). Desta forma, α é o parâmetro de mistura, e neste caso, $Y = 0$ com probabilidade α e com probabilidade $(1-\alpha)$, Y possui distribuição beta.

A esperança e a variância de uma variável aleatória Y com distribuição BIZ de parâmetros α, μ e ϕ são, respectivamente, expressas por:

$$E[Y] = (1-\alpha)\mu,$$
$$Var[Y] = (1-\alpha)\frac{\mu(1-\mu)}{\phi+1} + \alpha(1-\alpha)\mu^2.$$

O modelo de regressão beta inflacionado em zero (RBIZ) considerando $Y_1, ..., Y_n$ variáveis aleatórias independentes, em que cada Y_i, $i = 1, ..., n$, possui função densidade da forma (2.4) com parâmetros α, μ e ϕ, pode ser definido pelos seguintes componentes sistemáticos

$$g_0(\alpha_i) = \eta_{0i},$$
$$g_1(\mu_i) = \eta_{1i},$$

em que, $\eta_{0i} = \boldsymbol{x}_{0i}^T \boldsymbol{\beta}_0$, e $\eta_{1i} = \boldsymbol{x}_{1i}^T \boldsymbol{\beta}_1$ são os preditores lineares, $\boldsymbol{\beta}_0 = (\beta_{01}, ..., \beta_{0d_0})^T$ e $\boldsymbol{\beta}_1 = (\beta_{11}, ..., \beta_{1d_1})^T$ são vetores de parâmetros (desconhecidos) de interesse do modelo, $\boldsymbol{\beta}_0 \in \mathbb{R}^{d_0}$ e $\boldsymbol{\beta}_1 \in \mathbb{R}^{d_1}$, $\boldsymbol{x}_{0i} = (x_{0i1}, ..., x_{0id_0})^T$ e $\boldsymbol{x}_{1i} = (x_{1i1}, ..., x_{1id_1})^T$ são constantes conhecidas, i.e., vetores das observações de d_0 e d_1 covariáveis, $d_0 + d_1 < n$. As funções de ligação $g_0(\cdot)$ e $g_1(\cdot)$ são estritamente monótonas e duas vezes diferenciáveis, com domínio em $(0,1)$ e imagem em $\mathbb{R}$.

A estimação do vetor de parâmetros $(\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \phi)^T$ é realizada através do método de máxima verossimilhança. A função de verossimilhança para uma amostra aleatória da distribuição BIZ (2.4) de tamanho n, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ vetor das observações, é

$$\begin{aligned} L(\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \phi; \boldsymbol{y}) &= \prod_{i=1}^{n} biz_{Y_i} f(y_i;\alpha_i,\mu_i,\phi) \\ &= \prod_{i=1}^{n} \alpha_i^{\mathbb{1}_{\{0\}}(y_i)} (1-\alpha_i)^{1-\mathbb{1}_{\{0\}}(y_i)} \prod_{i:y_i \in (0,1)} f_{Y_i}(y_i;\mu_i,\phi), \end{aligned} \tag{2.5}$$

com $\mathbb{1}_{\{0\}}(y)$ função indicadora de $y \in \{0\}$, i.e., possui valor 1 se $y = 0$ e assume valor 0 se $y \neq 0$. As estimativas de máxima verossimilhança para os parâmetros $\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T$ e ϕ são obtidas através da maximização da função (2.5) com relação a estes parâmetros.

2.3 Support Vector Machine

Support Vector Machine (SVM) é uma técnica de aprendizado de máquina proposta por Cortes e Vapnik (1995), inicialmente apresentada em sua formulação para classificação de dois grupos linearmente não separáveis, generalizando uma ideia anterior utilizada para encontrar o hiperplano ótimo nos casos com dois grupos linearmente separáveis. O classificador SVM implementa a ideia de permitir a ampliação da dimensão do espaço para um espaço de alta dimensão, que em alguns casos pode ser infinita. Em geral, os limites lineares no espaço ampliado conseguem uma melhor separação da classe de treinamento e se traduzem em limites não lineares no espaço original. *Support Vector Regression Machines* (SVR) é uma adaptação do SVM para o tratamento de problemas de regressão, proposta por Drucker *et al.* (1996).

Yao, Crook e Andreeva (2015) realizaram estudo comparativo de técnicas de regressão para estimação da LGD e relataram que os modelos SVR demonstraram superioridade, em termos de precisão preditiva do ajuste, comparados às demais técnicas utilizadas no estudo. Observam também que modelos SVR possuem como limitação a característica de "caixa preta", no que concerne a dificuldade de distinguir o papel de cada variável no modelo.

Diferente de Yao, Crook e Andreeva (2015) que utilizaram LS-SVR [2], neste estudo utilizamos regressão SVR proposta por Drucker *et al.* (1996), o qual é apresentado a seguir.

2.3.1 Support vector regression machines

Admita um conjunto de dados composto por n pares, $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)\}$, em que $y_i \in \mathbb{R}$ é a variável dependente e $\boldsymbol{x}_i \in \mathbb{R}^d$ é um vetor de covariáveis de dimensão d. A regressão SVM proposta por Drucker *et al.* (1996), também conhecida como SVR ε-insensível, possui como objetivo encontrar uma função $f(\boldsymbol{x})$ que tenha no máximo ε desvio da observação y_i para todos os dados de treinamento com menor $\|\boldsymbol{\beta}\|^2$ possível, $\boldsymbol{\beta}$ vetor de parâmetros, para a seguinte função linear

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta} + \beta_0,$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros das covariáveis $\boldsymbol{x}$ e β_0 é o intercepto. Para encontrar estimativas para o vetor $\boldsymbol{\beta}$, a SVR formula um problema de otimização que consiste em minimizar uma função,

$$\arg\min_{\beta,\beta_0} \sum_{i=1}^{n} C(y_i, f(\boldsymbol{x}_i)) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2, \tag{2.6}$$

que é chamada de problema primal, em que

$$C(y_i, f(\boldsymbol{x}_i)) = \begin{cases} 0, & \text{se } |y_i - f(\boldsymbol{x}_i)| \leq \varepsilon, \\ |y_i - f(\boldsymbol{x}_i)| - \varepsilon, & \text{caso contrário.} \end{cases} \tag{2.7}$$

[2] *Least Squares Support Vector Regression* (LS-SVR) proposto por Suykens e Vandewalle (1999).

$C(y_i, f(\boldsymbol{x}_i))$ corresponde a função de perda (conhecida também como função de custo) ε-insensível, a qual não considera erros de tamanho inferior a $\varepsilon > 0$, e caso o erro seja maior que ε a perda é dada por $|y_i - f(\boldsymbol{x}_i)| - \varepsilon$. A Figura 2 ilustra este caso, em qua a função de perda ε-insensível linear ignora erros que estão dentro da distância ε, região cinza na Figura 2, tratando-os como iguais a zero, e a perda para as observações situadas fora da região cinza é medida de forma linear com base na distância entre o valor observado y_i e o limite ε (Smola e Schölkopf, 2004).

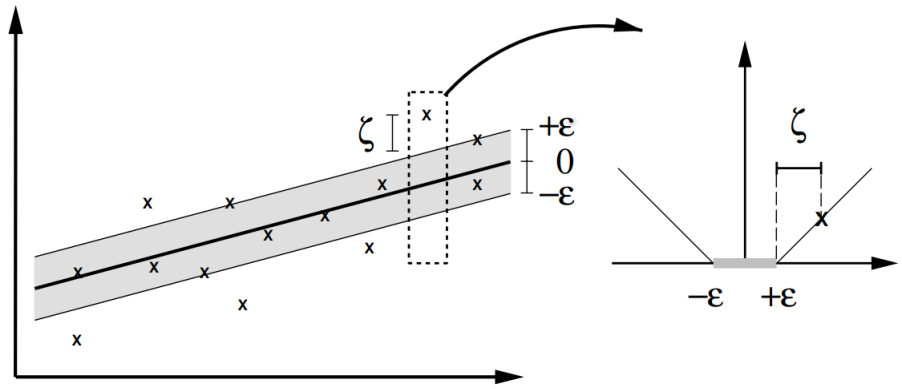

Figura 2 – Representação da perda ε-insensível para SVR linear. Fonte: Smola e Schölkopf (2004).

A constante $\lambda > 0$, em (2.6), controla a penalidade imposta nas observações fora da margem ε. Este valor equilibra entre redução de $\|\boldsymbol{\beta}\|^2$ e intolerância dos desvios maiores que ε.

De forma geral, o problema de otimização primal (2.6) pode ser mais facilmente solucionado em sua formulação dual. Conforme Friedman, Hastie e Tibshirani (2001), a função dual é uma Lagrangiana obtida a partir do problema primal introduzindo multiplicadores não negativos δ_i e δ_i^* para cada observação $\boldsymbol{x}_i$, levando a minimização

$$\arg\min_{\delta_i, \delta_i^*} \varepsilon \sum_{i=1}^{n} (\delta_i^* + \delta_i) - \sum_{i=1}^{n} y_i(\delta_i^* - \delta_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_i^* - \delta_i)(\delta_j^* - \delta_j) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle, \tag{2.8}$$

sujeito às restrições

$$\begin{cases} 0 \leq \delta_i, \delta_i^* \leq 1/\lambda, \\ \sum_{i=1}^{n} (\delta_i^* - \delta_i) = 0, \\ \delta_i^* \delta_i = 0. \end{cases}$$

Assim, de posse de $\hat{\delta}_i$ e $\hat{\delta}_i^*$, soluções do problema dual, temos que o estimador para $\boldsymbol{\beta}$ que minimiza o problema primal é da forma (Friedman, Hastie e Tibshirani, 2001)

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} (\hat{\delta}_i - \hat{\delta}_i^*) \boldsymbol{x}_i.$$

Desta forma, o estimador de $f(\boldsymbol{x})$ pode ser expresso por

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} (\hat{\delta}_i^* - \hat{\delta}_i)\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + \beta_0. \tag{2.9}$$

Devido à natureza dessas restrições, tipicamente apenas um subconjunto dos valores da solução $(\hat{\delta}_i^* - \hat{\delta}_i)$ são diferentes de zero, e a forma (2.9) é chamada "expansão do vector de suporte". Observe que a solução para função $\hat{f}(\boldsymbol{x})$ depende somente dos valores de entrada através dos produtos internos $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, ou seja, $\hat{\boldsymbol{\beta}}$ pode ser completamente descrito como uma combinação linear dos padrões de treinamento $\boldsymbol{x}_i$. Isto permite generalizar o método para uma abordagem não linear, isto é, a formulação dual Lagrangiana permite a utilização de *kernel*, e assim a proposta pode ser estendida para funções não lineares.

Para obter SVR não linear substituímos o produto escalar $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ em (2.8) por uma função kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, assim temos

$$\arg\min_{\delta_i, \delta_i^*} \varepsilon \sum_{i=1}^{n} (\delta_i^* + \delta_i) - \sum_{i=1}^{n} y_i(\delta_i^* - \delta_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_i^* - \delta_i)(\delta_j^* - \delta_j) K(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

sujeito as restrições

$$\begin{cases} 0 \leq \delta_i, \delta_i^* \leq 1/\lambda, \\ \sum_{i=1}^{n} (\delta_i^* - \delta_i) = 0, \\ \delta_i^* \delta_i = 0. \end{cases}$$

E assim obtemos a função $f(\boldsymbol{x})$, sem a necessidade de calcular $\boldsymbol{\beta}$ diretamente, por

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} (\hat{\delta}_i^* - \hat{\delta}_i) K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{\beta}_0, \tag{2.10}$$

com $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{M} h_m(\boldsymbol{x}_i) h_m(\boldsymbol{x}_j)$. É como se buscássemos a função

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{x}) + \beta_0$$

em que consideramos um conjunto de funções $\{h_m(\boldsymbol{x})\}$, $m = 1, 2, ..., M$, para o problema primal (2.6). Desta forma, com (2.10) não é necessário especificar ou avaliar todo o conjunto de funções $h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), ..., h_M(\boldsymbol{x})$, mas apenas o kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (Friedman, Hastie e Tibshirani, 2001).

Neste contexto, um kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ pode ser interpretado como uma medida de similaridade entre os vetores $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$. Um exemplo de kernel é o gaussiano, da forma

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2} \right).$$

Considerando que a distribuição da LGD, em alguns casos, pode apresentar trimodalidade (uma moda relacionada a inflação em zero e as outras duas referentes a beta bimodal) e que os

modelos de regressão beta e beta inflacionado em zero não são adequados para estes casos, e embora o modelo de regressão ε-SVM possa se ajustar a tais tipos de dados possui a característica de "caixa preta", neste sentido propomos um modelo de regressão baseado na mistura de duas distribuições beta e uma degenerada em zero, este modelo não padece da característica de "caixa preta"e permite ajuste a dados com distribuição que apresenta trimodalidade, este modelo é tema do próximo capítulo.

CAPÍTULO

3

MODELO DE REGRESSÃO BETA BIMODAL INFLACIONADO EM ZERO

Apoiado nas pesquisas empíricas e nos resultados de estudos recentes documentando os méritos relativos às abordagens de regressão paramétricas e não paramétricas, apresentamos uma proposta para estimar a LGD *Workout* em um único valor de t (fixo) e *point-in-time* utilizando mistura de duas distribuições beta com inflação em zero. Neste sentido, seguem os pressupostos que fundamentam a proposta.

- Existe massa de probabilidade no ponto 0: devido a posição conservadora das instituições de atribuir LGD $= 0$ às observações negativas. Desta forma teremos que o ponto 0 concentra toda a região de LGD $\leqslant 0$. Assim temos que

$$\text{LGD} \geqslant 0.$$

- A LGD é limitada superiormente por um k finito: devido a LGD ser proveniente de valores monetários, e considerando que os recursos monetários são finitos, então a LGD é limitada. Portanto

$$\text{LGD} \in [0,k].$$

  Aplicando a seguinte transformação

$$y_i = \frac{\text{LGD}_i}{k+\varepsilon},$$

  em que $\varepsilon > 0$ pequeno (exemplo $\varepsilon = 10^{-4}$) temos então que $y \in [0,1)$. Na prática, atribuímos a k o maior valor de LGD observada.

- A mistura de duas distribuições beta é adequada para ajustar $y > 0$: com base no formato extremamente flexível que a mistura de duas distribuições beta pode proporcionar. Não

existe prova teórica para esta suposição, porém de acordo com os relatos, provenientes de estudos conforme indicado no Capítulo 1, das características da LGD torna a mistura de duas distribuições beta uma suposição razoável para acomodar a distribuição da LGD.

Desta forma, consideramos uma variável aleatória definida em $[0,1)$, a qual é proveniente de uma mistura de duas distribuições beta e uma degenerada em zero.

3.1 Distribuição beta bimodal inflacionada em zero

Considere, inicialmente, uma variável aleatória W proveniente de uma mistura de duas distribuições beta, a distribuição desta variável aleatória possui função densidade da forma

$$bb_W(w;\pi,\mu_1,\phi_1,\mu_2,\phi_2) = \pi f_W(w;\mu_1,\phi_1) + (1-\pi) f_W(w;\mu_2,\phi_2) \tag{3.1}$$

em que $0 < w, \pi < 1$, $f_W(w;\mu_1,\phi_1)$ e $f_W(w;\mu_2,\phi_2)$ (densidades componentes) são funções densidade da distribuição beta da forma (2.2) referentes às duas subpopulações misturadas aleatoriamente com proporções π e $(1-\pi)$, respectivamente, os quais são chamados pesos das componentes (Frühwirth-Schnatter, 2006). Assim dizemos que W segue distribuição beta bimodal com parâmetros $\pi, \mu_1, \phi_1, \mu_2$ e ϕ_2. No Apêndice, Seção A.1, apresentamos a distribuição de mistura de duas beta com médias iguais, bem como algumas propriedades, incluindo esperança e variância.

A partir deste momento considere Y uma variável aleatória, que assume valores no intervalo $[0,1)$, originária da mistura de uma distribuição degenerada em zero e uma distribuição beta bimodal (3.1), cuja função densidade é dada por

$$bbz_Y(y;\alpha,\pi,\mu_1,\phi_1,\mu_2,\phi_2) = \begin{cases} \alpha, & \text{se } y = 0, \\ (1-\alpha) bb_Y(y;\pi,\mu_1,\phi_1,\mu_2,\phi_2), & \text{se } y \in (0,1), \end{cases} \tag{3.2}$$

com $\alpha, \pi, \mu_1, \mu_2 \in (0,1)$ e $\phi_1, \phi_2 > 0$. A função $bb_Y(y;\pi,\mu_1,\phi_1,\mu_2,\phi_2)$ refere-se a densidade beta bimodal (3.1). Observe que $\alpha = P(Y=0)$, configura a probabilidade de se observar o valor zero, e com probabilidade $(1-\alpha)$ a variável aleatória tem origem beta bimodal. Então $Y \sim BBZ(\alpha,\pi,\mu_1,\phi_1,\mu_2$ e $\phi_2)$, isto é, Y é uma variável aleatória com distribuição beta bimodal inflacionada em zero (BBZ) com parâmetros $\alpha, \pi, \mu_1, \phi_1, \mu_2$ e ϕ_2. Esta distribuição foi abordada inicialmente por Evandro Luiz de Souza Jackson em sua dissertação de mestrado, não finalizada, intitulada "Modelos de regressão beta-bimodal inflacionados de zeros". A distribuição BBZ é uma extensão da distribuição beta inflacionada em zero proposta por Martinez (2008). Empregamos neste trabalho notação semelhante a utilizada por este autor.

Considere $\boldsymbol{\vartheta} = (\alpha, \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_1 = (\mu_1, \phi_1)$ e $\boldsymbol{\theta}_2 = (\mu_2, \phi_2)$, utilizando $\mathbb{1}_{\{0\}}(y)$, função

indicadora de $y \in \{0\}$. Podemos expressar (3.2) da seguinte maneira

$$\begin{aligned} bbz_Y(y;\boldsymbol{\vartheta}) &= \alpha \mathbb{1}_{\{0\}}(y) + (1-\alpha) bb_Y(y;\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)(1-\mathbb{1}_{\{0\}}(y)) \\ &= \alpha \mathbb{1}_{\{0\}}(y) + (1-\alpha)\pi f_Y(y;\boldsymbol{\theta}_1)(1-\mathbb{1}_{\{0\}}(y)) + \\ &\quad (1-\alpha)(1-\pi) f_Y(y;\boldsymbol{\theta}_2)(1-\mathbb{1}_{\{0\}}(y)). \end{aligned} \tag{3.3}$$

De forma análoga a densidade beta inflacionada (Martinez, 2008), a função densidade (3.3) pode ser reescrita de maneira que fique fatorada em dois termos, um dependendo apenas de α e o outro termo dependendo dos parâmetros $\pi, \boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$. Desta forma,

$$\begin{aligned} bbz_Y(y;\boldsymbol{\vartheta}) &= \alpha^{\mathbb{1}_{\{0\}}(y)}[(1-\alpha)(\pi f_Y(y;\boldsymbol{\theta}_1) + (1-\pi) f_Y(y;\boldsymbol{\theta}_2))]^{(1-\mathbb{1}_{\{0\}}(y))} \\ &= [\alpha^{\mathbb{1}_{\{0\}}(y)}(1-\alpha)^{(1-\mathbb{1}_{\{0\}}(y))}][\pi f_Y(y;\boldsymbol{\theta}_1) + (1-\pi) f_Y(y;\boldsymbol{\theta}_2)]^{(1-\mathbb{1}_{\{0\}}(y))} \end{aligned}$$

Para determinar os momentos da distribuição beta bimodal inflacionada em zero, consideramos a esperança de uma função $m(Y)$ com respeito a densidade (3.3), $E[m(Y)]$, e utilizamos de propriedades da esperança condicional, semelhante ao realizado por Frühwirth-Schnatter (2006).

Teorema 1. Seja $E[m(Y)]$ a esperança de uma função $m(Y)$ com respeito a densidade (3.3). Então,

$$E[m(Y)] = m(0)p_0 + \sum_{k=1}^{2} E[m(Y)|\boldsymbol{\theta}_k]p_k,$$

em que $p_0 = \alpha$, $p_1 = (1-\alpha)\pi$ e $p_2 = (1-\alpha)(1-\pi)$.

Demonstração. Temos que

$$\begin{aligned} E[m(Y)] &= E[E[m(Y)|\mathbb{1}_{\{0\}}(y)]] \\ &= E[m(Y)|\mathbb{1}_{\{0\}}(y)=1]P(\mathbb{1}_{\{0\}}(y)=1) + E[m(Y)|\mathbb{1}_{\{0\}}(y)=0]P(\mathbb{1}_{\{0\}}(y)=0) \\ &= m(0)\alpha + E[m(Y)|\mathbb{1}_{\{0\}}(y)=0](1-\alpha). \end{aligned} \tag{3.4}$$

Resolvemos $E[m(Y)|\mathbb{1}_{\{0\}}(y)=0]$ separadamente

$$\begin{aligned} E[m(Y)|\mathbb{1}_{\{0\}}(y)=0] &= \int_{y:y\in(0,1)} m(y) f(y;\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)dy \\ &= \int_{y:y\in(0,1)} m(y)[\pi f(y;\boldsymbol{\theta}_1) + (1-\pi) f(y;\boldsymbol{\theta}_2)]dy \\ &= \pi \int_{y:y\in(0,1)} m(y) f(y;\boldsymbol{\theta}_1) + (1-\pi) \int_{y:y\in(0,1)} m(y) f(y;\boldsymbol{\theta}_2)dy \\ &= \pi E[m(Y)|\boldsymbol{\theta}_1] + (1-\pi)E[m(Y)|\boldsymbol{\theta}_2] \end{aligned} \tag{3.5}$$

em que $E[m(Y)|\boldsymbol{\theta}_k]$ é a esperança de $m(Y)$ dado que a variável aleatória Y é originária da distribuição beta com parâmetros μ_k e ϕ_k, $k = 1,2$. Se $m(0)$ e as esperanças condicionais $E[m(Y)|\boldsymbol{\theta}_k]$ existem, então de (3.4) e (3.5) temos que a esperança de $m(Y)$ é

$$E[m(Y)] = m(0)\alpha + E[m(Y)|\boldsymbol{\theta}_1]\pi(1-\alpha) + E[m(Y)|\boldsymbol{\theta}_2](1-\pi)(1-\alpha).$$

Assuma $p_0 = \alpha$, $p_1 = (1-\alpha)\pi$ e $p_2 = (1-\alpha)(1-\pi)$. Assim $E[m(Y)]$ pode ser expressa por

$$\begin{aligned} E[m(Y)] &= m(0)p_0 + E[m(Y)|\boldsymbol{\theta}_1]p_1 + E[m(Y)|\boldsymbol{\theta}_2]p_2 \\ &= m(0)p_0 + \sum_{k=1}^{2} E[m(Y)|\boldsymbol{\theta}_k]p_k \end{aligned}$$

□

Proposição 1. A esperança de Y é dada por

$$E[Y] = \mu_1\pi(1-\alpha) + \mu_2(1-\pi)(1-\alpha).$$

Demonstração. A partir do Teorema 1, tomando $m(Y) = Y$ obtemos a esperança de Y, como segue

$$\begin{aligned} E[Y] &= 0p_0 + \sum_{k=1}^{2} E[Y|\boldsymbol{\theta}_k]p_k \\ &= E[Y|\boldsymbol{\theta}_1]p_1 + E[Y|\boldsymbol{\theta}_2]p_2 \\ &= \mu_1\pi(1-\alpha) + \mu_2(1-\pi)(1-\alpha). \end{aligned}$$

□

Observe que a esperança de Y é uma média ponderada, pelos pesos das componentes, apenas das esperanças referentes às distribuições beta, uma vez que a densidade componente remanescente é uma distribuição degenerada em zero.

Proposição 2. A variância de Y é dada por

$$Var[Y] = \sum_{k=1}^{2} E[Y^2|\boldsymbol{\theta}_k]p_k - E[Y]^2.$$

Demonstração. Para obter a variância de Y empregamos o Teorema 1 tomando a função $m(Y) = (Y - E[Y])^2$, assim

$$\begin{aligned} Var[Y] &= E[(Y - E[Y])^2] \\ &= E[Y]^2 p_0 + \sum_{k=1}^{2} E[Y^2 - 2YE[Y] + E[Y]^2|\boldsymbol{\theta}_k]p_k \\ &= E[Y]^2 p_0 + \sum_{k=1}^{2} (E[Y^2|\boldsymbol{\theta}_k] - 2E[Y|\boldsymbol{\theta}_k]E[Y] + E[Y]^2)p_k \\ &= \sum_{k=1}^{2} E[Y^2|\boldsymbol{\theta}_k]p_k - 2E[Y]^2 + E[Y]^2 \\ &= \sum_{k=1}^{2} E[Y^2|\boldsymbol{\theta}_k]p_k - E[Y]^2. \end{aligned}$$

□

Da Proposição 2 temos que a variância de Y é a diferença entre a média ponderada, pelos pesos das componentes, do momento de segunda ordem das densidades componentes beta pelo quadrado da esperança de Y.

Corolário 1. A variância pode ser expressa em termos da variância das densidades componentes

$$Var[Y] = \left(\frac{\mu_1(1-\mu_1)}{\phi_1+1} + \mu_1^2\right)\pi(1-\alpha) + \left(\frac{\mu_2(1-\mu_2)}{\phi_2+1} + \mu_2^2\right)(1-\pi)(1-\alpha) - \mu^2, \quad (3.6)$$

em que $\mu = E[Y]$, $\frac{\mu_k(1-\mu_k)}{\phi_k+1} = Var[Y|\boldsymbol{\theta}_k]$ e $\mu_k = E[Y|\boldsymbol{\theta}_k]$, $k = 1,2$.

Demonstração. Da Proposição 2, temos que

$$\begin{aligned}
Var[Y] &= \sum_{k=1}^{2} E[Y^2|\boldsymbol{\theta}_k]p_k - E[Y]^2 \\
&= \sum_{k=1}^{2} (Var[Y|\boldsymbol{\theta}_k] + E[Y|\boldsymbol{\theta}_k]^2)p_k - E[Y]^2 \\
&= \left(\frac{\mu_1(1-\mu_1)}{\phi_1+1} + \mu_1^2\right)\pi(1-\alpha) + \left(\frac{\mu_2(1-\mu_2)}{\phi_2+1} + \mu_2^2\right)(1-\pi)(1-\alpha) - \mu^2.
\end{aligned}$$

□

Corolário 2. A variância de Y pode ser expressa da seguinte forma

$$\begin{aligned}
Var[Y] = (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + (1-\alpha)\pi(1-\pi)(\mu_1-\mu_2)^2 + \\
\alpha(1-\alpha)[\mu_1\pi + \mu_2(1-\pi)]^2. \quad (3.7)
\end{aligned}$$

Demonstração. De (3.6), temos que

$$\begin{aligned}
Var[Y] &= \left(\frac{\mu_1(1-\mu_1)}{\phi_1+1} + \mu_1^2\right)\pi(1-\alpha) + \left(\frac{\mu_2(1-\mu_2)}{\phi_2+1} + \mu_2^2\right)(1-\pi)(1-\alpha) - \mu^2 \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + (1-\alpha)\left[\mu_1^2\pi + \mu_2^2(1-\pi)\right] - \mu^2 \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + (1-\alpha)\left[\mu_1^2\pi + \mu_2^2(1-\pi)\right] \\
&\quad - \{[\mu_1\pi + \mu_2(1-\pi)](1-\alpha)\}^2 \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi + \mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\left[\mu_1^2\pi + \mu_2^2(1-\pi)\right] - (1-\alpha)[\mu_1\pi + \mu_2(1-\pi)]^2 \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi + \mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\left[\mu_1^2\pi + \mu_2^2(1-\pi) - \mu_1^2\pi^2 - \mu_2^2(1-\pi)^2 - 2\mu_1\mu_2\pi(1-\pi)\right]
\end{aligned}$$

$$
\begin{aligned}
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi+\mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\left[\mu_1^2(\pi-\pi^2) + \mu_2^2(1-\pi-(1-\pi)^2) - 2\mu_1\mu_2\pi(1-\pi)\right] \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi+\mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\left[\mu_1^2\pi(1-\pi) + \mu_2^2\pi(1-\pi) - 2\mu_1\mu_2\pi(1-\pi)\right] \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi+\mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\pi(1-\pi)\left[\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2\right] \\
&= (1-\alpha)\left[\frac{\mu_1(1-\mu_1)}{\phi_1+1}\pi + \frac{\mu_2(1-\mu_2)}{\phi_2+1}(1-\pi)\right] + \alpha(1-\alpha)[\mu_1\pi+\mu_2(1-\pi)]^2 \\
&\quad + (1-\alpha)\pi(1-\pi)(\mu_1-\mu_2)^2.
\end{aligned}
$$

□

Assim em (3.7) a variância de Y é a soma de três termos, em que um corresponde a variância dos componentes beta, outro relativo a diferença entre a média dos componentes beta e o termo restante diz respeito ao quadrado da esperança de Y.

Os parâmetros ϕ_1 e ϕ_2 podem ser interpretados como parâmetros de precisão, uma vez que, fixados os demais parâmetros, quanto maior forem os valores para ϕ_1 e ϕ_2 menor serão as variâncias destas distribuições. Quando $\phi_1, \phi_2 \to \infty$ temos que

$$
\begin{aligned}
Var[Y] \longrightarrow (1-\alpha)\pi(1-\pi)(\mu_1-\mu_2)^2 + \alpha(1-\alpha)[\mu_1\pi+\mu_2(1-\pi)]^2 \\
= (1-\alpha)\pi(1-\pi)(\mu_1-\mu_2)^2 + \frac{\alpha}{1-\alpha}\mu^2.
\end{aligned}
$$

Proposição 3. O r-ésimo momento em torno de zero da distribuição beta bimodal inflacionada em zero é dado por

$$
E[Y^r] = \frac{(\mu_1\phi_1)_{(r)}}{\phi_{1(r)}}(1-\alpha)\pi + \frac{(\mu_2\phi_2)_{(r)}}{\phi_{2(r)}}(1-\alpha)(1-\pi).
$$

Demonstração. A expressão para os momentos de ordem superior em torno de zero pode ser obtida tomando $m(Y) = Y^r$ no resultado do Teorema 1, assim o r-ésimo momento em torno de zero é dado por

$$
E[Y^r] = 0^r p_0 + \sum_{k=1}^{2} E[Y^r|\boldsymbol{\theta}_k]p_k
$$

temos que $E[Y^r|\boldsymbol{\theta}_1]$ e $E[Y^r|\boldsymbol{\theta}_2]$ são os momentos em torno de zero de ordem r da distribuição beta (2.2) e como observado por Martinez (2008) é dado por

$$
E[Y^r|\boldsymbol{\theta}_k] = \frac{\Gamma(\phi_k)\Gamma(\mu_k\phi_k+1)}{\Gamma(\phi_k+r)\Gamma(\mu_k\phi_k)} = \prod_{j=0}^{r-1}\frac{(\mu_k\phi_k+j)}{(\phi_k+j)} = \frac{(\mu_k\phi_k)_{(r)}}{\phi_{k(r)}}
$$

para $k = 1$ e 2, $j = 0,1,2,...,r-1$, e $a_{(r)} = a(a+1)(a+2)...(a+r-1)$. Desta forma, para $r > 0$, temos que

$$\begin{aligned} E[Y^r] &= 0^r p_0 + E[Y^r|\boldsymbol{\theta}_k]p_k \\ &= \sum_{k=1}^{2} E[Y^r|\boldsymbol{\theta}_k]p_k \\ &= \frac{(\mu_1\phi_1)_{(r)}}{\phi_{1(r)}}(1-\alpha)\pi + \frac{(\mu_2\phi_2)_{(r)}}{\phi_{2(r)}}(1-\alpha)(1-\pi). \end{aligned}$$

□

Assim o momento de ordem r da distribuição BBZ é uma média ponderada dos momentos de ordem r das distribuições beta pertencentes a mistura.

Proposição 4. O r-ésimo momento em torno da média para uma variável aleatória com distribuição beta bimodal inflacionada em zero pode ser expressa por

$$E[(Y-\mu)^r] = (-\mu)^r + \sum_{k=1}^{2}\sum_{j=0}^{r}\binom{r}{j}(\mu_k-\mu)^{r-j}\left[\frac{(\mu_k\phi_k)_{(j)}}{\phi_{k(j)}} + \sum_{i=0}^{j-1}\binom{j}{i}(-\mu_k)^{j-i}\frac{(\mu_k\phi_k)_{(i)}}{\phi_{k(i)}}\right]p_k.$$

Demonstração. A expressão para os momentos de ordem superior em torno da média pode ser obtida tomando a função $m(Y) = (Y-\mu)^r$, usada no resultado do Teorema 1, e o auxílio da fórmula binomial. Assim,

$$\begin{aligned} E[(Y-\mu)^r] &= (0-\mu)^r + \sum_{k=1}^{2} E[(Y-\mu)^r|\boldsymbol{\theta}_k]p_k \\ &= (-\mu)^r + \sum_{k=1}^{2} E[(Y-\mu_k+\mu_k-\mu)^r|\boldsymbol{\theta}_k]p_k \\ &= (-\mu)^r + \sum_{k=1}^{2} E\left[\sum_{j=0}^{r}\binom{r}{j}(Y-\mu_k)^j(\mu_k-\mu)^{r-j}|\boldsymbol{\theta}_k\right]p_k \\ &= (-\mu)^r + \sum_{k=1}^{2}\sum_{j=0}^{r}\binom{r}{j}(\mu_k-\mu)^{r-j}E[(Y-\mu_k)^j|\boldsymbol{\theta}_k]p_k, \end{aligned} \tag{3.8}$$

em que $E[(Y-\mu_k)^j|\boldsymbol{\theta}_k]$ é o momento de ordem j em torno da média da distribuição beta com parâmetros μ_k e ϕ_k, a qual pode ser obtida por

$$\begin{aligned} E[(Y-\mu_k)^j|\boldsymbol{\theta}_k] &= \int_{(0,1)}(y-\mu_k)^j f(y;\mu_k,\phi_k)dy \\ &= \int_{(0,1)}\sum_{i=0}^{j}\binom{j}{i}(y)^i(-\mu_k)^{j-i}f(y;\mu_k,\phi_k)dy \\ &= \sum_{i=0}^{j}\binom{j}{i}(-\mu_k)^{j-i}\int_{(0,1)}(y)^i f(y;\mu_k,\phi_k)dy \\ &= \sum_{i=0}^{j}\binom{j}{i}(-\mu_k)^{j-i}\mu_{k(i)} \end{aligned}$$

$$= \mu_{k(j)} + \sum_{i=0}^{j-1} \binom{j}{i} (-\mu_k)^{j-i} \mu_{k(i)}$$
$$= \frac{(\mu_k \phi_k)_{(j)}}{\phi_{k(j)}} + \sum_{i=0}^{j-1} \binom{j}{i} (-\mu_k)^{j-i} \frac{(\mu_k \phi_k)_{(i)}}{\phi_{k(i)}}. \tag{3.9}$$

Portanto, de (3.8) e (3.9) o r-ésimo momento em torno da média para uma variável aleatória com distribuição beta bimodal inflacionada em zero pode ser expressa por

$$E[(Y-\mu)^r] = (-\mu)^r + \sum_{k=1}^{2} \sum_{j=0}^{r} \binom{r}{j} (\mu_k - \mu)^{r-j} \left[\frac{(\mu_k \phi_k)_{(j)}}{\phi_{k(j)}} + \sum_{i=0}^{j-1} \binom{j}{i} (-\mu_k)^{j-i} \frac{(\mu_k \phi_k)_{(i)}}{\phi_{k(i)}} \right] p_k.$$

□

Corolário 3. A assimetria para uma variável aleatória com distribuição beta bimodal inflacionada é obtida com o terceiro momento em torno da média, expresso por

$$E[(Y-\mu)^3] = (-\mu)^3 + \sum_{k=1}^{2} \sum_{j=0}^{3} \binom{3}{j} (\mu_k - \mu)^{3-j} E[(Y-\mu_k)^j | S=k] p_k$$
$$= (-\mu)^3 + \sum_{k=1}^{2} \sum_{j=0}^{3} \binom{3}{j} (\mu_k - \mu)^{3-j} \left[\frac{(\mu_k \phi_k)_{(j)}}{\phi_{k(j)}} + \sum_{i=0}^{j-1} \binom{j}{i} (-\mu_k)^{j-i} \frac{(\mu_k \phi_k)_{(i)}}{\phi_{k(i)}} \right] p_k,$$

resultado obtido ao tomar $r = 3$ na expressão do r-ésimo momento em torno da média, Proposição 4.

Corolário 4. A curtose para uma variável aleatória com distribuição beta bimodal inflacionada é obtida com o quarto momento em torno da média, expreso por

$$E[(Y-\mu)^4] = (-\mu)^4 + \sum_{k=1}^{2} \sum_{j=0}^{4} \binom{4}{j} (\mu_k - \mu)^{4-j} E[(Y-\mu_k)^j | S=k] p_k$$
$$= (-\mu)^4 + \sum_{k=1}^{2} \sum_{j=0}^{4} \binom{4}{j} (\mu_k - \mu)^{4-j} \left[\frac{(\mu_k \phi_k)_{(j)}}{\phi_{k(j)}} + \sum_{i=0}^{j-1} \binom{j}{i} (-\mu_k)^{j-i} \frac{(\mu_k \phi_k)_{(i)}}{\phi_{k(i)}} \right] p_k,$$

resultado obtido quando tomamos $r = 4$ na expressão do r-ésimo momento em torno da média, Proposição 4.

Na Figura 3 alguns gráficos são apresentados para ilustrar a flexibilidade da distribuição beta bimodal inflacionada em zero para diferentes vetores de parâmetros. Nestes gráficos a probabilidade de ocorrer zero, isto é, $\alpha = P(Y=0)$ é representada por uma barra vertical e um ponto para demarcar tal probabilidade. Iniciamos com $\boldsymbol{\vartheta}_1 = (\alpha = 0.25, \pi = 0.4, \mu_1 = 0.3, \phi_1 = 10, \mu_2 = 0.9, \phi_2 = 12)$. Em $\boldsymbol{\vartheta}_2$ apenas π é alterado para 0.1, na sequência, em $\boldsymbol{\vartheta}_3$ alteramos o valor de π para $0,8$. Em $\boldsymbol{\vartheta}_4$ alteramos μ_1 e μ_2 para 0.4 e 0.8, respectivamente, com relação à $\boldsymbol{\vartheta}_1$. Aproximamos as médias das distribuições beta em $\boldsymbol{\vartheta}_5$, com $\mu_1 = 0.5$ e $\mu_2 = 0.7$, observando uma aparência unimodal no intervalo $(0,1)$. Em $\boldsymbol{\vartheta}_6$, ambos parâmetros de precisão são reduzidos, ϕ_1

para 3 e ϕ_2 para 5, juntamente com alterações nas médias μ_1 e μ_2 para 0.3 e 0.8, respectivamente, e assim uma curva em forma de U é observada no intervalo $(0,1)$. No vetor $\boldsymbol{\vartheta}_7$ aumentamos as precisões ϕ_1 para 13 e ϕ_2 para 15, e voltamos a observar bimodalidade em $(0,1)$. Por fim em $\boldsymbol{\vartheta}_8$ temos $\alpha = 0.2, \pi = 0.3, \mu_1 = 0.4, \phi_1 = 30, \mu_2 = 0.9$ e $\phi_2 = 45$. Observe que todas as distribuições apresentadas possuem formas assimétricas, uma característica da distribuição BBZ, fato decorrente da presença de probabilidade não nula no ponto zero.

O termo bimodal, da distribuição beta bimodal inflacionada em zero, refere-se a possibilidade da distribuição em apresentar duas modas no intervalo $(0,1)$, mas nem sempre isto ocorre, como visto na Figura 3.

3.2 Estimação dos parâmetros

A estimação dos parâmetros para $\alpha, \pi, \mu_1, \phi_1, \mu_2$ e ϕ_2 é realizada através do método de máxima verossimilhança. A função de verossimilhança para $\boldsymbol{\vartheta}$ dada uma amostra aleatória, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, da distribuição beta bimodal inflacionada em zero, é

$$\begin{aligned} L(\boldsymbol{\vartheta}; \boldsymbol{y}) &= \prod_{i=1}^{n} [\alpha^{\mathbb{1}_{\{0\}}(y_i)} (1-\alpha)^{(1-\mathbb{1}_{\{0\}}(y_i))}][\pi f(y_i; \boldsymbol{\theta}_1) + (1-\pi) f(y_i; \boldsymbol{\theta}_2)]^{(1-\mathbb{1}_{\{0\}}(y_i))} \\ &= L_1(\alpha; \boldsymbol{y}) L_2(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}), \end{aligned} \tag{3.10}$$

em que

$$\begin{aligned} L_1(\alpha; \boldsymbol{y}) &= \prod_{i=1}^{n} \alpha^{\mathbb{1}_{\{0\}}(y_i)} (1-\alpha)^{(1-\mathbb{1}_{\{0\}}(y_i))}, \\ L_2(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}) &= \prod_{i=1}^{n} [\pi f(y_i; \boldsymbol{\theta}_1) + (1-\pi) f(y_i; \boldsymbol{\theta}_2)]^{(1-\mathbb{1}_{\{0\}}(y_i))}. \end{aligned}$$

Observe que (3.10) é fatorada em dois termos $L_1(\alpha; \boldsymbol{y})$ e $L_2(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y})$. Obtemos o logaritmo da função de verossimilhança, a qual, geralmente, é de mais fácil maximização que a original e produz os mesmos estimadores, aplicando a função logarítmica em (3.10). Assim,

$$\ell(\boldsymbol{\vartheta}; \boldsymbol{y}) = \ell_1(\alpha; \boldsymbol{y}) + \ell_2(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}), \tag{3.11}$$

em que

$$\ell_1(\alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(y_i) \log(\alpha) + \left(n - \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(y_i) \right) \log(1-\alpha), \tag{3.12}$$

$$\ell_2(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}) = \sum_{i; y_i \in (0,1)} \log[\pi f(y_i; \mu_1, \phi_1) + (1-\pi) f(y_i; \mu_2, \phi_2)]. \tag{3.13}$$

Os estimadores de máxima verossimilhança $\hat{\boldsymbol{\vartheta}}$ para os parâmetros $\boldsymbol{\vartheta}$ podem ser obtidas através da maximização da função (3.11), isto é,

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} \{\ell_1(\alpha; \boldsymbol{y}) + \ell_2(\pi, \mu_1, \phi_1, \mu_2, \phi_2; \boldsymbol{y})\} \tag{3.14}$$

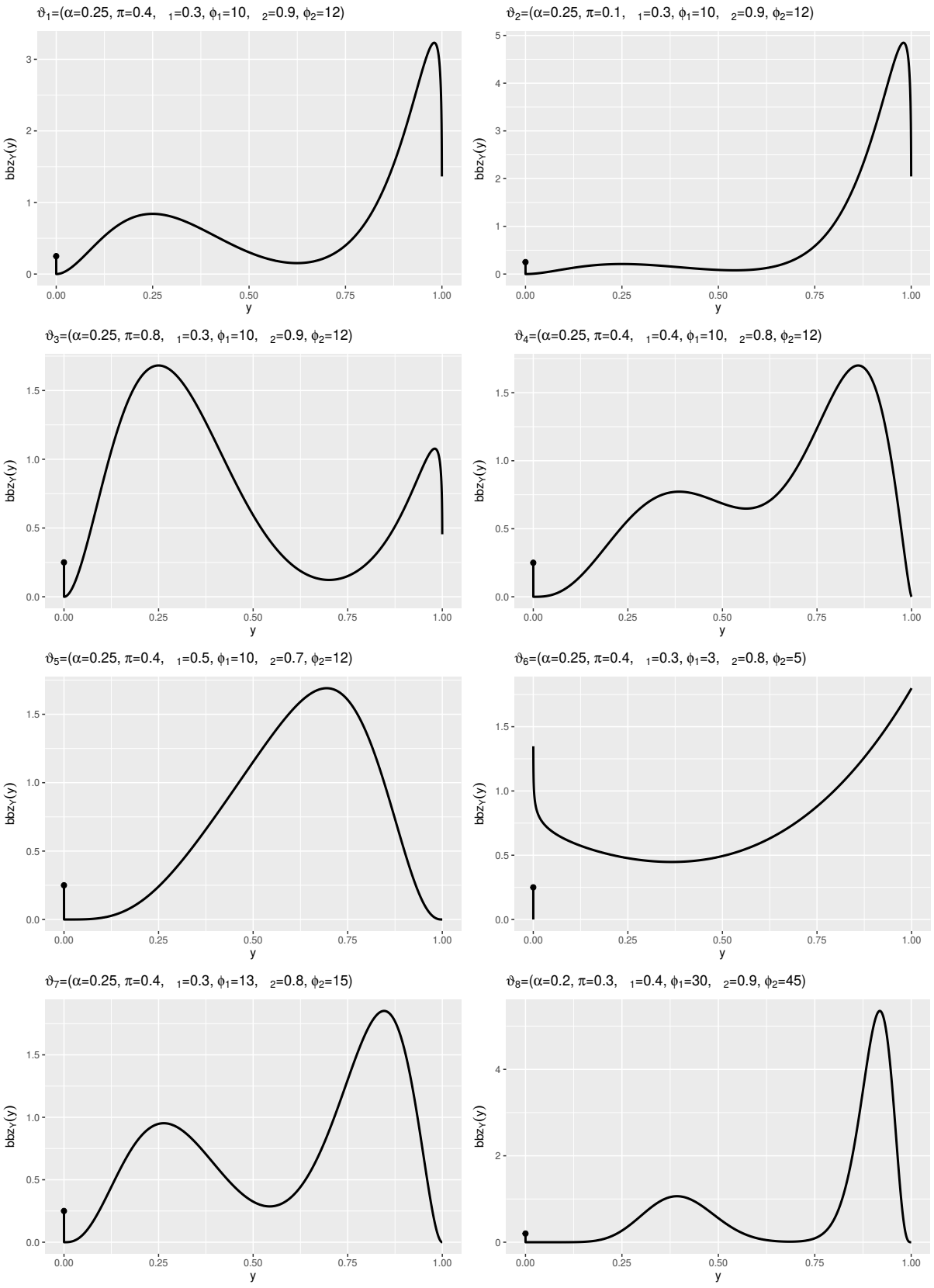

Figura 3 – Gráficos de densidades da distribuição beta bimodal inflacionada em zero para diferentes vetores de parâmetros.

Devido a função (3.10) ser fatorável obtivemos função log-verossimilhança (3.11) determinada pela soma de duas funções, $\ell_1(\alpha;\boldsymbol{y})$ e $\ell_2(\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{y})$, e assim a solução para o problema de maximização (3.14) pode ser obtida separadamente, uma vez que $\ell_1(\alpha;\boldsymbol{y})$ depende apenas de α, e $\ell_2(\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{y})$ de π,μ_1,ϕ_1,μ_2 e ϕ_2. Desta forma, os estimadores são dados por

$$\hat{\alpha} = \arg\max_{\alpha} \ell_1(\alpha;\boldsymbol{y}),$$
$$(\hat{\pi},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2) = \arg\max_{(\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)} \ell_2(\pi,\mu_1,\phi_1,\mu_2,\phi_2;\boldsymbol{y}).$$

O estimador de máxima verossimilhança para α possui expressão em forma fechada, tomando a derivada de $\ell_1(\alpha;\boldsymbol{y})$, Equação 3.12, em relação a α,

$$\frac{\partial \ell(\alpha;\boldsymbol{y})}{\alpha} = \frac{\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)}{\alpha} - \frac{n-\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)}{1-\alpha},$$

e igualando a zero,

$$\begin{aligned}
\frac{\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)}{\hat{\alpha}} - \frac{n-\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)}{1-\hat{\alpha}} &= 0 \\
\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)(1-\hat{\alpha}) - \hat{\alpha}\left(n - \sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)\right) &= 0 \\
\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i) - \sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)\hat{\alpha} - \hat{\alpha}n + \hat{\alpha}\sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i) &= 0 \\
\hat{\alpha} &= \sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)/n.
\end{aligned}$$

Assim, o estimador de máxima verossimilhança para α é $\hat{\alpha} = \sum_{i=1}^n \mathbb{1}_{\{0\}}(y_i)/n$. No entanto, o mesmo não ocorre com os estimadores $(\hat{\pi},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)$, que devido a soma de termos dentro do logaritmo em $\ell_2(\pi,\mu_1,\phi_1,\mu_2,\phi_2;\boldsymbol{y})$ torna sua maximização direta muito difícil, porém existe uma abordagem mais simples (Friedman, Hastie e Tibshirani, 2001).

A função (3.13) envolve a função densidade da mistura de duas betas da forma (3.1) para as observações y_i, $i = 1,2,...,n$, que pertencem ao intervalo $(0,1)$, ou seja, podemos considerar apenas as observações $y_i \in (0,1)$, $i = 1,2,...,n$, e, desta maneira, tratarmos (3.13) como a função log-verossimilhança da densidade (3.1) para $\pi,\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$. Assim sendo, considere uma variável latente não observada $S_i \in \{0,1\}$ tal que

$$S_i = \begin{cases} 1, & \text{se a observação } i \text{ vem do componente } f(y;\boldsymbol{\theta}_1), \\ 0, & \text{se a observação } i \text{ vem do componente } f(y;\boldsymbol{\theta}_2). \end{cases}$$

Suponha conhecido os valores para S_i. Então a função de verossimilhança para a mistura de duas betas pode ser escrita como

$$L(\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\boldsymbol{y},\boldsymbol{S}) = \prod_{i;y_i\in(0,1)} [\pi f(y_i;\mu_1,\phi_1)]^{S_i}[(1-\pi)f(y_i;\mu_2,\phi_2)]^{(1-S_i)},$$

em que $\boldsymbol{S} = (S_1, S_2, ..., S_n)$ e, assim, a função log-verossimilhança é dada por

$$\ell(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}, \boldsymbol{S}) = \sum_{i; y_i \in (0,1)} [S_i \log f(y_i; \mu_1, \phi_1) + (1 - S_i) \log f(y_i; \mu_2, \phi_2)] + \sum_{i; y_i \in (0,1)} [S_i \log \pi + (1 - S_i) \log(1 - \pi)]. \quad (3.15)$$

O estimador de máxima verossimilhança (MV) de μ_1 e ϕ_1 podem ser obtidos maximizando (3.15) com relação a μ_1 e ϕ_1 para as observações em que $S_i = 1$, e da mesma maneira pode ser feito para obter os estimadores MV de μ_2 e ϕ_2 para os dados que $S_i = 0$. Mesmo assim, a maximização destas funções não permitem solução em forma fechada, sendo necessário uso de algum método numérico.

Como os valores de S_i são desconhecidos, utilizamos então o algoritmo EM (*Expectation-Maximization*) (Dempster, Laird e Rubin, 1977), e assim continuamos de forma iterativa substituindo cada S_i em (3.15) por seu valor esperado

$$\begin{aligned}
\gamma_i &= E[S_i | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{y}] \\
&= P(S_i = 1 | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{y}) \\
&= \frac{P(S_i = 1, y_i | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{P(y_i | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\
&= \frac{bb_Y(y_i | S_i = 1, \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) P(S_i = 1 | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{bb_Y(y_i | \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \\
&= \frac{f(y_i | \mu_1, \phi_1)\pi}{\pi f(y_i | \mu_1, \phi_1) + (1 - \pi) f(y_i | \mu_2, \phi_2)}
\end{aligned}$$

para a i-ésima observação, γ_i é a probabilidade condicional da observação y_i vir do componente $f(y; \mu_1, \phi_1)$, dado y_i e $(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. O algoritmo EM começa com estimativas iniciais e alterna entre passo E e passo M até obter a convergência. No passo E (Esperança), a esperança condicional do log-verossimilhança dos dados completos (3.15) é calculada sobre os dados faltantes $\boldsymbol{S}$, dado $\boldsymbol{y}$ e as estimativas atuais de $(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. No passo M (Maximização), as estimativas para γ_i, obtidas no passo E, são utilizadas para maximizar $E[\ell(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2; \boldsymbol{y}, \boldsymbol{S}) | \boldsymbol{y}, \hat{\pi}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2]$ com relação a $(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ e assim atualizar as estimativas desses parâmetros, ou seja,

$$(\hat{\pi}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \arg \max_{(\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \sum_{i; y_i \in (0,1)} [\hat{\gamma}_i \log f(y_i; \mu_1, \phi_1) + (1 - \hat{\gamma}_i) \log f(y_i; \mu_2, \phi_2)] + \sum_{i; y_i \in (0,1)} [\hat{\gamma}_i \log \pi + (1 - \hat{\gamma}_i) \log(1 - \pi)].$$

Este problema de maximização pode ser separado em 3 problemas de maximização menores dependendo cada um deles de (μ_1, ϕ_1), (μ_2, ϕ_2) e π, sendo que para π a solução fornece uma expressão de forma fechada, o que não ocorre para os demais parâmetros, veja Algoritmo 1.

Existem diversas variações propostas para o algoritmo EM, tais como as versões SEM (*Stochastic* EM) e CEM (*Classification* EM), para ambas um passo é adicionado entre os passos

Algoritmo 1 – Algoritmo EM para mistura de dois componentes beta

Entrada: Valores iniciais para as estimativas dos parâmetros $\pi, \mu_1, \phi_1, \mu_2$ e ϕ_2, isto é, estabeleça valores para $\hat{\pi}^{(0)}, \hat{\mu}_1^{(0)}, \hat{\phi}_1^{(0)}, \hat{\mu}_2^{(0)}$ e $\hat{\phi}_2^{(0)}$
Saída: Estimativas MV para os parâmetros $\pi, \mu_1, \phi_1, \mu_2$ e ϕ_2

Início
$\hat{\pi}^{(0)}, \hat{\mu}_1^{(0)}, \hat{\phi}_1^{(0)}, \hat{\mu}_2^{(0)}$ e $\hat{\phi}_2^{(0)}$
repita
procedimento PASSO E, NA ITERAÇÃO $k+1$.
para i **faça**1n

$$\hat{\gamma}_i^{(k+1)} = \frac{\hat{\pi}^{(k)} f(y_i|\hat{\mu}_1^{(k)}, \hat{\phi}_1^{(k)})}{\hat{\pi}^{(k)} f(y_i|\hat{\mu}_1^{(k)}, \hat{\phi}_1^{(k)}) + (1-\hat{\pi}^{(k)}) f(y_i|\hat{\mu}_2^{(k)}, \hat{\phi}_2^{(k)})}, \ i \text{ tal que } y_i \in (0,1).$$

fim para
fim procedimento
procedimento PASSO M, CALCULA $\hat{\mu}_1^{(k+1)}, \hat{\phi}_1^{(k+1)}, \hat{\mu}_2^{(k+1)}, \hat{\phi}_2^{(k+1)}$ E $\hat{\pi}^{(k+1)}$, EM QUE

$$(\hat{\mu}_1^{(k+1)}, \hat{\phi}_1^{(k+1)}) = \arg \max_{(\mu_1^{(k)}, \phi_1^{(k)})} \sum_{i; y_i \in (0,1)} [\hat{\gamma}_i^{(k+1)} \log f(y_i; \mu_1^{(k)}, \phi_1^{(k)})$$

$$(\hat{\mu}_2^{(k+1)}, \hat{\phi}_2^{(k+1)}) = \arg \max_{(\mu_2^{(k)}, \phi_2^{(k)})} \sum_{i; y_i \in (0,1)} (1-\hat{\gamma}_i^{(k+1)}) \log f(y_i; \mu_2^{(k)}, \phi_2^{(k)})]$$

$$\hat{\pi}^{(k+1)} = \arg \max_{\pi^{(k)}} \sum_{i; y_i \in (0,1)} [\hat{\gamma}_i^{(k+1)} \log \pi^{(k)} + (1-\hat{\gamma}_i^{(k+1)}) \log(1-\pi^{(k)})]$$

$$= \sum_{i; y_i \in (0,1)} \hat{\gamma}_i^{(k+1)} / \sum_{i; y_i \in (0,1)} 1.$$

fim procedimento
até atingir a convergência.
Fim
Retorna $\hat{\pi}, \hat{\mu}_1, \hat{\phi}_1, \hat{\mu}_2$ e $\hat{\phi}_2$

E e M. Este novo passo usa as estimativas de $\hat{\gamma}_i$ obtidas no passo E e atribui a cada observação um componente de origem (Grun e Leisch, 2008). Neste caso, para mistura de duas distribuições beta, entre o passo E e o M, cada observação $y_i \in (0,1)$ atribui-se um correspondente $\hat{S}_i \in \{0,1\}$, $i = 1, ..., n$.

No algoritmo SEM a determinação de $\hat{S}_i$ é feita de forma estocástica, chamemos o passo adicional do algoritmo SEM de passo St. Para a mistura de duas distribuições beta o passo St é dado da seguinte forma: dado as estimativas obtidas por $\hat{\gamma}_i$ no passo E, defina

$$\hat{S}_i \sim Bernoulli(\hat{\gamma}_i)$$

em que $Bernoulli(\hat{\gamma}_i)$ denota a distribuição Bernoulli com probabilidade de sucesso $\hat{\gamma}_i$, isto é, $P(\hat{S}_i = 1) = \hat{\gamma}_i$, $i = 1, ..., n$. Posteriormente, o $\hat{S}_i$ é usado ao invés do $\hat{\gamma}_i$ no passo M.

Desta forma, o problema de otimização deixa de ser ponderado por $\hat{\gamma}_i$ no passo M, consequentemente, o problema de otimização fica mais fácil de ser implementado e resolvido, o mesmo ocorre com o algoritmo CEM.

Para o algoritmo CEM a atribuição de $\hat{S}_i$ é realizada de forma determinística, chamemos seu passo adicional de passo C. O passo C para a mistura de duas distribuições beta é dado da seguinte maneira: dada as estimativas $\hat{\gamma}_i$, $\hat{S}_i$ é determinado por

$$\hat{S}_i = \begin{cases} 1 & \text{se } \hat{\gamma}_i \geq 0,5, \\ 0 & \text{caso contrário.} \end{cases}$$

assim a observação y_i é atribuída ao componente com maior probabilidade condicional estimada desta observação ter sido originada.

Como mostram Friedman, Hastie e Tibshirani (2001), a iteração EM não decresce a log-verossimilhança, garantindo que o algoritmo EM atinge um máximo local, caso ocorra convergência. Ambas as variantes, SEM e CEM, foram propostas com a intenção de melhorar o desempenho do algoritmo EM, porque além do algoritmo EM convencional atingir apenas um máximo local este também tende a convergir lentamente. Grun e Leisch (2008) mencionam que o comportamento de convergência pode ser melhor com o algoritmo CEM do que o algoritmo EM, enquanto o algoritmo SEM pode escapar da convergência para um máximo local, no entanto, o algoritmo CEM não produz estimativas MV, mas o SEM proporciona boas aproximações do estimador MV.

Conforme já relatado o algoritmo EM pode ser lento, e esta característica pode ser potencializada quando utilizados valores iniciais ruins para os parâmetros. Uma maneira de atribuir estimativas iniciais para o algoritmo EM é especificar a densidade componente de origem de cada observação, particionando assim a amostra, e então estimar os parâmetros para cada grupo separadamente e utilizar estas estimativas como valor inicial no algoritmo EM. Uma forma bastante comum de especificar a densidade componente de origem de cada observação é a atribuição aleatória, outra alternativa é utilizar algum método das K-médias[1] para separar as observações (McLachlan e Peel, 2000). Bagnato e Punzo (2013) propõem o algoritmo *k*-bumps como alternativa de estratégia para inicialização do algoritmo EM.

Um valor inicial ruim além de potencializar a lentidão do algoritmo EM pode conduzir a convergência a um máximo local. De acordo com McLachlan e Peel (2000) um problema com os modelos de mistura é que a função de verossimilhança geralmente possui múltiplas raízes correspondentes aos máximos locais. E portanto, Grun e Leisch (2008) recomendam repetir o algoritmo EM a partir de um amplo conjunto de valores iniciais diferentes na busca de diversos máximos locais, e escolher como solução final para a raiz da função de verossimilhança aquele

[1] O método das K-médias visa dividir os pontos em k grupos, de modo que a soma de quadrados dos pontos para os centros dos *clusters* atribuídos seja minimizada.

com maior verossimilhança observado, embora este procedimento não garanta atingimento do máximo global, ainda é uma boa alternativa.

Devido a iteração do algoritmo EM não decrescer a log-verossimilhança, um critério de parada poderia ser o de avaliar o incremento na log-verossimilhança, e assim, interromper o processo se o incremento for menor do que uma quantidade pré-estabelecida.

Com as estimativas MV, $\hat{\alpha}, \hat{\pi}, \hat{\mu}_1$ e $\hat{\mu}_2$, para os parâmetros α, π, μ_1 e μ_2 e considerando a propriedade de invariância[2] do estimador de máxima verossimilhança temos que o estimador MV para o valor esperado de uma observação com distribuição beta bimodal inflacionada em zero pode ser obtido por

$$\hat{\mu} = \hat{\mu}_1\hat{\pi}(1-\hat{\alpha}) + \hat{\mu}_2(1-\hat{\pi})(1-\hat{\alpha}).$$

3.3 Modelo de regressão beta bimodal inflacionado em zero

Sejam $Y_1, ..., Y_n$ variáveis aleatórias independentes, em que cada Y_i, $i = 1, 2, ..., n$, possui função densidade beta bimodal inflacionada em zero da forma (3.2), com parâmetros $\alpha_i, \pi_i, \mu_{1i}, \phi_{1i}, \mu_{2i}$ ϕ_{2i}, respectivamente. Os modelos de regressão beta bimodal inflacionados em zero (RBBZ) são definidos pelos seguintes componentes sistemáticos:

$$\begin{aligned} g_0(\boldsymbol{\alpha}) &= \boldsymbol{\eta}_0 = X_0\boldsymbol{\beta}_0, \\ g_1(\boldsymbol{\mu}_1) &= \boldsymbol{\eta}_1 = X_1\boldsymbol{\beta}_1, \\ g_2(\boldsymbol{\mu}_2) &= \boldsymbol{\eta}_2 = X_2\boldsymbol{\beta}_2, \\ g_3(\boldsymbol{\pi}) &= \boldsymbol{\eta}_3 = X_3\boldsymbol{\beta}_3, \end{aligned} \tag{3.16}$$

em que $\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\pi}$ e $\boldsymbol{\eta}_k$, $k = 0, 1, 2$ e 3, são vetores de tamanho n, $\boldsymbol{\beta}_k^T = (\beta_{k1}, \beta_{k2}, ..., \beta_{kd_k})$ é um vetor de tamanho d_k, X_k é uma matriz de valores conhecidos da ordem $n \times d_k$. As funções $g_k(\cdot)$, são denominadas funções de ligação, relacionam os vetores de parâmetros $\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ e $\boldsymbol{\pi}$ às variáveis explanatórias em X_0, X_1, X_2 e X_3, respectivamente. As funções de ligação são conhecidas e devem ser estritamente monótonas e $g_k(\cdot) : (0,1) \longrightarrow \mathbb{R}$, ou seja, possuir domínio no intervalo $(0,1)$ e imagem nos reais, uma vez que estas funções ligam $0 < \alpha_i, \mu_{1i}, \mu_{2i}, \pi_i < 1$ ao preditor linear $\boldsymbol{\eta}_{ki}$. A função probito, $g(a) = \Phi^{-1}(a)$, e a função log-log, $g(a) = \log\{-\log(a)\}$, são exemplos de funções para $g_k(\cdot)$, para esta função uma família de transformações proposta por Aranda-Ordaz (1981) também pode ser utilizada, esta família é dada por

$$g(a) = \log\left\{\frac{(1-a)^{-\lambda} - 1}{\lambda}\right\}, \tag{3.17}$$

[2] Se $\hat{\theta}$ é o estimador de máxima verossimilhança para θ, então para qualquer função $g(\theta)$, o estimador de máxima verossimilhança para $g(\theta)$ é $g(\hat{\theta})$ (Casella e Berger, 2002).

em que λ, constante desconhecida, é um parâmetro de transformação tal que com $\lambda = 1$ a função (3.17) é equivalente a função logito, $g(a) = \log\{a/(1-a)\}$, e quando $\lambda \to 0$ temos a função complemento log-log, $g(a) = \log\{-\log(1-a)\}$.

Apesar de ser possível também relacionar preditores lineares aos parâmetros ϕ_1 e ϕ_2, aqui optamos por modelar apenas os parâmetros que compõem a esperança de uma variável aleatória beta bimodal inflacionada em zero.

O ajuste do modelo RBBZ, especificado por (3.16), é realizado pela estimação de máxima verossimilhança. Desta forma o logaritmo da função de máxima verossimilhança para uma amostra observada, $\boldsymbol{y} = (y_1, ..., y_n)$ da amostra aleatória $Y_1, ..., Y_n$, com função densidade beta bimodal inflacionada em zero com parâmetros $\boldsymbol{\varphi} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_1^T, \phi_1, \boldsymbol{\beta}_2^T, \phi_2)$, é dado por

$$\ell(\boldsymbol{\varphi}; \boldsymbol{y}) = \ell_1(\boldsymbol{\alpha}; \boldsymbol{y}) + \ell_2(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \phi_1, \boldsymbol{\mu}_2, \phi_2; \boldsymbol{y}),$$

em que

$$\ell_1(\boldsymbol{\alpha}; \boldsymbol{y}) = \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(y_i) \log(\alpha_i) + \big(1 - \mathbb{1}_{\{0\}}(y_i)\big) \log(1 - \alpha_i), \tag{3.18}$$

$$\ell_2(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \phi_1, \boldsymbol{\mu}_2, \phi_2; \boldsymbol{y}) = \sum_{i; y_i \in (0,1)} \log[\pi_i f(y_i; \mu_{1i}, \phi_1) + (1 - \pi_i) f(y_i; \mu_{2i}, \phi_2)]. \tag{3.19}$$

De maneira semelhante à estimação por máxima verossimilhança para os parâmetros da distribuição BBZ, a estimação MV para os parâmetros do modelo de regressão RBBZ pode ser realizado de forma separada, maximizando a função $\ell_1(\boldsymbol{\alpha}; \boldsymbol{y})$ com respeito à $\boldsymbol{\beta}_0$ e $\ell_2(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \phi_1, \boldsymbol{\mu}_2, \phi_2; \boldsymbol{y})$ com relação a $\boldsymbol{\beta}_3, \boldsymbol{\beta}_1, \phi_1, \boldsymbol{\beta}_2, \phi_2$, bem como o processo de estimação utilizado pode ser o mesmo. Assim, para

$$\hat{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta}_0} \ell_1(\boldsymbol{\alpha}; \boldsymbol{y}),$$

pode ser realizado diretamente por algum método numérico e para

$$(\hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_1, \hat{\phi}_1, \hat{\boldsymbol{\beta}}_2, \hat{\phi}_2) = \arg \max_{(\boldsymbol{\beta}_3, \boldsymbol{\beta}_1, \phi_1, \boldsymbol{\beta}_2, \phi_2)} \ell_2(\boldsymbol{\pi}, \boldsymbol{\mu}_1, \phi_1, \boldsymbol{\mu}_2, \phi_2; \boldsymbol{y})$$

o algoritmo EM pode ser utilizado.

Ainda neste capítulo, abordamos a construção de intervalos de confiança, testes de hipóteses e seleção de modelos para os modelos beta bimodal inflacionados em zero. No próximo capítulo um estudo de simulação é realizado com a intenção de avaliar o desempenho dos estimadores de máxima verossimilhança para os parâmetros das distribuição beta bimodal inflacionada em zero.

3.4 Identificabilidade

A identificabilidade é uma importante condição em um modelo estatístico. A falta desta condição acarreta que diferentes valores de parâmetros podem originar distribuições de

probabilidade idênticas. Neste sentido, quando há falta de identificabilidade no modelo o processo de estimação é prejudicado.

Huang (2005) menciona que um modelo é identificável se os valores dos parâmetros determinarem de forma exclusiva a distribuição de probabilidade dos dados e a distribuição de probabilidade dos dados determina os valores dos parâmetros.

De maneira mais formal Frühwirth-Schnatter (2006) define identificabilidade. Para tanto, considere uma família de distribuições paramétricas, indexadas por um parâmetro $\boldsymbol{\vartheta} \in \Theta$, que é definido sobre um espaço amostral $\mathscr{Y}$, esta família é dita ser identificável se quaisquer dois parâmetros $\boldsymbol{\vartheta}$ e $\boldsymbol{\vartheta}^*$ em Θ define a mesma lei de probabilidade em $\mathscr{Y}$, se e somente se $\boldsymbol{\vartheta}$ e $\boldsymbol{\vartheta}^*$ são idênticos. Em termos de densidades de probabilidades, correspondentes $f(y;\boldsymbol{\vartheta})$ e $f(y;\boldsymbol{\vartheta}^*)$, isso significa que se as densidades são idênticas para quase todo $y \in \mathscr{Y}$, então os parâmetros $\boldsymbol{\vartheta}$ e $\boldsymbol{\vartheta}^*$ precisão ser idênticos,

$$f(y;\boldsymbol{\vartheta}) = f(y;\boldsymbol{\vartheta}^*) \text{ para quase todo } y \in \mathscr{Y} \rightarrow \boldsymbol{\vartheta} = \boldsymbol{\vartheta}^*. \tag{3.20}$$

O fato do modelo beta bimodal inflacionado em zero ser baseado em mistura de distribuições implica que este modelo sofre dos mesmos problemas de falta de identificabilidade encontrado nos modelos de misturas finitas.

Uma não identificabilidade observada em modelos de mistura é consequência da invariância de $f(y;\boldsymbol{\vartheta})$ sobre a permutação dos índices dos componentes em $\boldsymbol{\vartheta}$, conhecido como *label switching*. Esta não identificabilidade é observada também no modelo beta bimodal inflacionado em zero. Considere uma distribuição beta bimodal inflacionada em zero com vetor de parâmetros $\boldsymbol{\vartheta} = (\alpha, \pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, tome $\boldsymbol{\vartheta}$ arbitrário em que $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, e defina o vetor de parâmetros $\boldsymbol{\vartheta}^* = (\alpha, 1-\pi, \boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$, obtido pela troca na ordem dos componentes beta. Embora os vetores de parâmetros sejam diferentes, a distribuição obtida por $\boldsymbol{\vartheta}$ é a mesma de $\boldsymbol{\vartheta}^*$, veja

$$\begin{aligned} f(y;\boldsymbol{\vartheta}) &= \alpha \mathbb{1}_{\{0\}}(y) + (1 - \mathbb{1}_{\{0\}}(y))(1-\alpha)[\pi f(y;\boldsymbol{\theta}_1) + (1-\pi) f(y;\boldsymbol{\theta}_2)] \\ &= \alpha \mathbb{1}_{\{0\}}(y) + (1 - \mathbb{1}_{\{0\}}(y))(1-\alpha)[(1-\pi) f(y;\boldsymbol{\theta}_2) + \pi f(y;\boldsymbol{\theta}_1)] \\ &= f(y;\boldsymbol{\vartheta}^*). \end{aligned}$$

Frühwirth-Schnatter (2006) argumenta que este não é um problema grave de identificabilidade, uma vez que a distinção está apenas em como os componentes estão arranjados, pois todos os parâmetros e as permutações estão relacionados um ao outro e, de fato, apenas diferem na forma como os componentes são organizados. Segundo McLachlan e Peel (2000) essa falta de identificação não é preocupante no curso normal dos eventos na montagem de modelos de mistura por máxima verossimilhança, por exemplo, através do algoritmo EM. Ao realizar a estimação através da máxima verossimilhança no caso da beta bimodal inflacionada em zero o único problema observado com o *label switching* é que os parâmetros podem apresentar trocados, não influenciando na inferência.

Outra não identificabilidade observada nos modelos de mistura é devido ao potencial superajuste, *overfitting*, que é o ajuste de muitos componentes no modelo. McLachlan e Peel (2000) relata como não identificabilidade devido superajuste modelar incorretamente uma mistura de $g-1$ componentes por uma mistura de g componentes, que pode ocorrer de duas maneiras: um dos pesos na mistura de g componentes pode ser ajustada como zero, ou duas densidades componentes na mistura de g componentes podem ser consideradas iguais.

O fato dos pesos componentes no modelo beta bimodal inflacionado serem maiores que zero, por definição, asseguram que superajuste de um componente ser vazio não ocorre, entretanto o outro caso não é garantido, ou seja, as densidades betas podem ser iguais. Exemplo, suponha que a verdadeira distribuição dos dados é uma distribuição beta inflacionada em zero com vetor de parâmetros (α, μ, ϕ), mas ao invés o modelo beta bimodal com vetor de parâmetros $\boldsymbol{\vartheta}^* = (\alpha^*, \pi^*, \mu_1^*, \phi_1^*, \mu_2^*, \phi_2^*)$ é utilizado, se $\alpha = \alpha^*$, $\mu = \mu_1^* = \mu_2^*$ e $\phi = \phi_1^* = \phi_2^*$ então para qualquer $\pi^* \in (0,1)$ temos que

$$\begin{aligned} f(y;\boldsymbol{\vartheta}^*) &= \alpha^* \mathbb{1}_{\{0\}}(y) + (1-\mathbb{1}_{\{0\}}(y))(1-\alpha^*)[\pi^* f(y;\mu_1^*,\phi_1^*) + (1-\pi^*) f(y;\mu_2^*,\phi_2^*)] \\ &= \alpha \mathbb{1}_{\{0\}}(y) + (1-\mathbb{1}_{\{0\}}(y))(1-\alpha)[\pi^* f(y;\mu,\phi) + (1-\pi^*) f(y;\mu,\phi)] \\ &= \alpha \mathbb{1}_{\{0\}}(y) + (1-\mathbb{1}_{\{0\}}(y))(1-\alpha) f(y;\mu,\phi) \\ &= f(y;\alpha,\mu,\phi). \end{aligned}$$

Segundo Frühwirth-Schnatter (2006) a identificabilidade pode ser atingida de uma maneira formal através da imposição de restrições no espaço paramétrico de tal forma que diferentes parâmetros não gerem a mesma distribuição e assim a condição (3.20) seja cumprida.

Podemos aplicar algumas das restrições no espaço paramétrico citadas por Frühwirth-Schnatter (2006) para resolver a não identificabilidade devido *label switching* e *overfitting* dos modelos beta bimodal inflacionado em zero e multivariados.

Para evitar a falta de identificabilidade devido ao potencial superajuste no modelo beta bimodal inflacionado em zero basta que os componentes beta sejam distintos, o que significa que ao menos um dos parâmetros em cada componente beta seja diferente, assim para a distribuição beta bimodal inflacionada em zero podemos impor que $\mu_1 \neq \mu_2$ e para o modelo de regressão beta bimodal inflacionado em zero que $\boldsymbol{\beta}_1$e $\boldsymbol{\beta}_2$ sejam diferentes, ou seja, $\beta_{1j} \neq \beta_{2j}$ para algum j.

A não identificabilidade *label switching* pode ser evitada no modelo beta bimodal inflacionado em zero impondo uma restrição de ordem em algum dos elementos dos vetores de parâmetros das densidades componentes beta. Para o modelo beta bimodal inflacionado em zero podemos impor $\mu_1 < \mu_2$ e para o modelo de regressão beta bimodal inflacionado em zero $\beta_{1j} < \beta_{2j}$ para um determinado j. Em particular, se utilizamos estas restrições para solucionar a não identificabilidade *label switching* resolvemos também a falta de identificabilidade devido ao potencial superajuste, uma vez que $\mu_1 < \mu_2$ implica $\mu_1 \neq \mu_2$, bem como $\beta_{1j} < \beta_{2j}$ para um determinado j implica $\beta_{1j} \neq \beta_{2j}$ para algum j.

De acordo com Frühwirth-Schnatter (2006) as distribuições de misturas finitas podem permanecer não identificáveis mesmo que uma restrição de identificabilidade formal exclua quaisquer dos problemas de não identificabilidade descritos anteriormente.

Uma vez que diferentes valores de parâmetros podem gerar idênticas distribuições, a não identificabilidade de um modelo estatístico impacta o processo de estimação dos parâmetros deste modelo, exemplo, ao utilizar o estimador de máxima verossimilhança podemos obter diferentes estimativas dos parâmetros relacionados a mesma máxima verossimilhança do conjunto de dados, ou seja, caso consigamos atingir o máximo global com o algoritmo EM, pode existir diferentes estimativas ligadas a este mesmo máximo global. Sendo assim, existe a necessidade de avaliar se as estimativas dos parâmetros obtida tem relação um pra um com a máxima verossimilhança encontrada. Uma prática comum para verificar se há uma relação um pra um entre as estimativas e a verossimilhança, envolve executar o algoritmo EM a partir de diferentes valores iniciais para a estimativas dos parâmetros, selecionar as soluções que levam ao mesmo maior máximo local, diferentes valores iniciais que produzem a mesma máxima verossimilhança devem resultar nas mesmas estimativas dos parâmetros, caso contrário o modelo é não identificável.

3.5 Intervalos de confiança e teste de hipóteses

As propriedades assintóticas do estimador de máxima verossimilhança (EMV) permitem a realização de teste de hipóteses e obtenção de intervalos de confiança assintóticos para os parâmetros do modelo de regressão beta bimodal inflacionado em zero. Aqui, assumimos $\boldsymbol{\varphi}$ identificável.

Sob certas condições de regularidade, veja Cox e Hinkley (1974), temos que $\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})$ converge em distribuição para distribuição normal multivariada da ordem d^+, $d^+ = d_0 + d_3 + d_1 + d_2 + 2$,

$$\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{\text{D}} N_{d^+}(0, J(\boldsymbol{\varphi})^{-1})$$

em que $\hat{\boldsymbol{\varphi}} = (\hat{\boldsymbol{\beta}}_0^T, \hat{\boldsymbol{\beta}}_3^T, \hat{\boldsymbol{\beta}}_1^T, \hat{\phi}_1, \hat{\boldsymbol{\beta}}_2^T, \hat{\phi}_2)^T$ é o estimador de máxima verossimilhança de $\boldsymbol{\varphi} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_3^T, \boldsymbol{\beta}_1^T, \phi_1, \boldsymbol{\beta}_2^T, \phi_2)^T$, $J(\boldsymbol{\varphi}) = \lim_{n\to\infty} \mathrm{I}(\boldsymbol{\varphi})/n$ existe e é não-singular, $\mathrm{I}(\boldsymbol{\varphi})$ é a matriz de informação de Fisher definida como

$$\mathrm{I}(\boldsymbol{\varphi}) = E_{\boldsymbol{\varphi}}\left[U(\boldsymbol{\varphi};\boldsymbol{Y})U^T(\boldsymbol{\varphi};\boldsymbol{Y})\right]$$

em que

$$U(\boldsymbol{\varphi};\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\varphi};\boldsymbol{y})}{\partial \boldsymbol{\varphi}}$$

é o vetor gradiente da função log-verossimilhança, conhecido como função escore, e $\boldsymbol{y} = (y_1, ..., y_n)$ vetor de dados observados. Sob condições de regularidade a matriz de informação de

Fisher pode ser expressa como

$$\mathrm{I}(\boldsymbol{\varphi}) = E_{\boldsymbol{\varphi}}\left[-\mathrm{H}(\boldsymbol{\varphi};\boldsymbol{Y})\right]$$

em que

$$\mathrm{H}(\boldsymbol{\varphi};\boldsymbol{y}) = \frac{\partial^2 \ell(\boldsymbol{\varphi};\boldsymbol{y})}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^T}$$

é a matriz Hessiana da função log-verossimilhança (McLachlan e Peel, 2000).

Assim pela normalidade assintótica do estimador de máxima verossimilhança $\hat{\boldsymbol{\varphi}}$, podemos construir intervalos de confiança de tipo assintótico para os parâmetros deste modelo de regressão. Visto que a matriz de covariância assintótica do estimador máxima verossimilhança é igual ao inverso da matriz de informação de Fisher, que pode ser aproximada por $\mathrm{I}(\hat{\boldsymbol{\varphi}})$, o erro padrão de $\hat{\varphi}_r = (\hat{\boldsymbol{\varphi}})_r$ é dado por

$$EP(\hat{\varphi}_r) \approx (\mathrm{I}^{-1}(\hat{\boldsymbol{\varphi}}))_{rr}^{1/2} \; r = 1, ..., d^+.$$

em que a notação $(A)_{rs}$ é usada para denotar o elemento na r-ésima linha e s-ésima coluna da matriz A. Usualmente, o inverso da matriz covariância do EMV é aproximada pela matriz de informação observada, definida como $\mathrm{I}(\hat{\boldsymbol{\varphi}};\boldsymbol{y}) = -\mathrm{H}(\hat{\boldsymbol{\varphi}};\boldsymbol{y})$ (McLachlan e Peel, 2000).

Desta forma, para amostras grandes, intervalos de confiança aproximados para cada um dos parâmetros com coeficiente de confiança de $(1-\mathfrak{c})$, $\mathfrak{c} \in (0,1)$, são dados por

$$IC(\varphi_r;\mathfrak{c}) = \left[\hat{\varphi}_r - z_{(1-\mathfrak{c}/2)}EP(\hat{\varphi}_r); \hat{\varphi}_r + z_{(1-\mathfrak{c}/2)}EP(\hat{\varphi}_r)\right]$$

em que $z_{(1-\mathfrak{c}/2)}$ representa o quantil $1-\mathfrak{c}/2$ da distribuição normal padrão, $N(0,1)$. Dizemos que $IC(\varphi_r;\mathfrak{c})$ é o intervalo de confiança assintótico com nível de $100(1-\mathfrak{c})\%$ para φ_r.

Geralmente, após ajustar um modelo de regressão, existe o interesse em investigar a importância de determinadas covariáveis no modelo, isto pode ser realizado através de teste de hipóteses. No modelo de regressão beta bimodal inflacionado em zero podemos utilizar o teste de razão de verossimilhança (TRV), para descrever tal teste, assumimos interesse em testar hipóteses sobre um subconjunto dos vetores de parâmetros $\boldsymbol{\beta}_0, \boldsymbol{\beta}_3, \boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$, consideremos partições destes vetores expressas por $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0t}^T, \boldsymbol{\beta}_{0*}^T)^T$, $\boldsymbol{\beta}_3 = (\boldsymbol{\beta}_{3t}^T, \boldsymbol{\beta}_{3*}^T)^T$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{1t}^T, \boldsymbol{\beta}_{1*}^T)^T$ e $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_{2t}^T, \boldsymbol{\beta}_{2*}^T)^T$, em que os vetores de parâmetros de interesse são $\boldsymbol{\beta}_{0t}, \boldsymbol{\beta}_{3t}, \boldsymbol{\beta}_{1t}$ e $\boldsymbol{\beta}_{2t}$, com respectivas dimensões d_{0t}, d_{3t}, d_{1t} e d_{2t}, tais que $d_{0t} < d_0$, $d_{3t} < d_3$, $d_{1t} < d_1$, $d_{2t} < d_2$, neste caso, as hipóteses do teste são representadas pela hipótese nula, $\mathrm{H}_0 : \boldsymbol{\beta}_{0t} = \boldsymbol{\beta}_{0t}^{\{0\}}, \boldsymbol{\beta}_{3t} = \boldsymbol{\beta}_{3t}^{\{0\}}, \boldsymbol{\beta}_{1t} = \boldsymbol{\beta}_{1t}^{\{0\}}, \boldsymbol{\beta}_{2t} = \boldsymbol{\beta}_{2t}^{\{0\}}$ e hipótese alternativa, H_1: descumprimento de ao menos uma das desigualdades de H_0, em que os vetores $\boldsymbol{\beta}_{0t}^{\{0\}}, \boldsymbol{\beta}_{3t}^{\{0\}}, \boldsymbol{\beta}_{1t}^{\{0\}}, \boldsymbol{\beta}_{2t}^{\{0\}}$ são valores especificados para os parâmetros de interesse.

A estatística do teste da razão de verossimilhanças é dada por

$$Q_{rv} = 2\{\ell(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_1, \hat{\phi}_1, \hat{\boldsymbol{\beta}}_2, \hat{\phi}_2;\boldsymbol{y}) - \ell(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}_3, \tilde{\boldsymbol{\beta}}_1, \tilde{\phi}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\phi}_2;\boldsymbol{y})\},$$

em que $\ell(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_1, \hat{\phi}_1, \hat{\boldsymbol{\beta}}_2, \hat{\phi}_2; \boldsymbol{y})$ é o logaritmo da função de verossimilhança aplicado no EMV de $\boldsymbol{\varphi}$ sem restrições, e $\ell(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}_3, \tilde{\boldsymbol{\beta}}_1, \tilde{\phi}_1, \tilde{\boldsymbol{\beta}}_2, \tilde{\phi}_2; \boldsymbol{y})$ é o logaritmo da função de verossimilhança aplicado no EMV dos parâmetros em teste, ou seja, EMV de $\boldsymbol{\varphi}$ sob hipótese nula.

As condições de regularidade que levam a normalidade assintótica do estimador de máxima verossimilhança levam também, sob H_0, $Q_{rv} \xrightarrow{\mathrm{D}} \chi^2_{d_{0t}+d_{3t}+d_{1t}+d_{2t}}$, desse modo, o teste pode ser realizado usando valores críticos aproximados de uma distribuição χ^2 com $d_{0t} + d_{3t} + d_{1t} + d_{2t}$ graus de liberdade.

Cabe salientar que tanto intervalos de confiança quanto teste da razão de verossimilhanças têm garantia de serem inferencialmente válidos de forma assintótica, assim no modelo de regressão beta bimodal inflacionado em zero o tamanho amostral n deve ser muito grande para aplicação da teoria assintótica.

3.6 Seleção de modelos

Na prática, é comum o ajuste de diversos modelos, o que nos leva a tarefa de selecionar um modelo estatístico dentre um conjunto de modelos candidatos. Quando, por exemplo, estamos diante do problema de selecionar entre dois modelos encaixados - modelo com quantidade menor de parâmetros é um caso particular do modelo com maior quantidade de parâmetros - podemos definir hipóteses e utilizar o teste da razão de verossimilhança, discutido na Seção 3.5, para selecionar o melhor modelo. Porém quando os modelos do conjunto de modelos candidatos não se tratam de modelos encaixados o TRV se torna inapropriado para a tarefa de selecionar o melhor modelo.

Existem diversos critérios que podem ser utilizados para avaliar a qualidade de modelos, na seleção de modelos não encaixados os critérios de informação são bastante utilizados para esta finalidade, o critério de informação de Akaike (AIC) e o critério de informação bayesiano (BIC) são alguns destes critérios. O AIC e BIC são definidos, respectivamente, por

$$\begin{aligned} \mathrm{AIC} &= -2\ell(\hat{\boldsymbol{\varphi}}; \boldsymbol{y}) + 2d^+, \\ \mathrm{BIC} &= -2\ell(\hat{\boldsymbol{\varphi}}; \boldsymbol{y}) + d^+ \log(n), \end{aligned}$$

em que $\ell(\hat{\boldsymbol{\varphi}}; \boldsymbol{y})$ é o valor da função log-verossimilhança do modelo com os parâmetros $\hat{\boldsymbol{\varphi}}$, d^+ é a quantidade total de parâmetros neste modelo e n quantidade de observações (McLachlan e Peel, 2000). A seleção de modelos envolve encontrar um modelo parcimonioso, com balanço ideal entre quantidade máxima de informação e mínimo de parâmetros, obtendo assim um modelo com melhor capacidade de generalização. Tanto o AIC quanto o BIC penalizam os modelos com muitos parâmetros, sendo que os modelos preferíveis são aqueles com valores menores de AIC e BIC.

Uma alternativa aos critérios AIC e BIC para selecionar um modelo consiste em dividar a amostra de dados disponível em amostra de desenvolvimento e amostra de validação, a amostra

de desenvolvimento deve ser utilizada no ajuste dos modelos e a amostra de validação na estimação da função log-verossimilhança dos modelos obtidos. Neste caso, o modelo a ser selecionado é o que apresentar o maior valor estimado para log-verossimilhança na amostra de validação.

Todas as abordagens aqui apresentadas para seleção de modelos possuem ampla aplicação e podem ser utilizadas na seleção de modelos de regressão beta bimodal inflacionados em zero.

CAPÍTULO

4

ESTUDO DE SIMULAÇÃO

Estudos de simulação foram realizados com o objetivo de obtermos certa segurança quanto à qualidade das estimativas obtidas a partir do estimador de máxima verossimilhança para os parâmetros tanto da distribuição BBZ como do modelo de regressão BBZ. Primeiramente, comparamos três diferentes estratégias de inicialização do algoritmo EM quanto a eficiência de tais estratégias no atingimento da solução ideal para a maximização da função log-verossimilhança, avaliamos também quanto a identificabilidade. Posteriormente, realizamos experimento de simulação para avaliar o comportamento dos estimadores de máxima verossimilhança, obtidos através do algoritmo EM com inicialização por método K-médias, dos parâmetros da distribuição beta bimodal inflacionada em zero com o aumento do tamanho amostral, similar estudo foi realizado para os parâmetros do modelo de regressão beta bimodal inflacionado em zero. Por fim, simulamos um banco de dados de LGD e comparamos o desempenho da regressão beta bimodal inflacionada em zero com a regressão beta, a regressão beta inflacionada em zero e o *support vector regression machines*.

Os estudos foram realizados com apoio computacional do software R. Em todas as estimações tomamos como critério de parada para o algoritmo EM: incremento menor que 10^{-6} no valor da função log-verossimilhança, ou o máximo de 200 iterações.

4.1 Inicialização do algoritmo EM

A solução para o problema de maximização da função log-verossimilhança da mistura de duas distribuições beta envolve o uso do algoritmo EM, e como visto na Seção 3.2, este algoritmo garante atingimento de máximos locais para este problema, porém nosso real interesse reside na obtenção do máximo global, caso exista. Neste sentido, em consonância com o relatado por Grun e Leisch (2008) que a convergência do algoritmo EM para a solução ideal depende da inicialização, avaliamos 3 diferentes estratégias de inicialização do algoritmo EM quanto a eficiência de tais estratégias no atingimento da solução ideal para a maximização da

função log-verossimilhança, avaliamos também quanto a identificabilidade, caso determinada estratégia tenha atingido ou ficado próximo do valor da função log-verossimilhança estabelecido como referência, averiguamos se estas soluções também possuíam os mesmos valores para as estimativas, ou pelo menos se estavam próximos.

A primeira estratégia, estratégia 1, consiste em separar, de forma aleatória, as observações pertencentes ao intervalo $(0,1)$ em dois grupos, assim estimar os parâmetros como dados completos, e com estas estimativas iniciar o algoritmo EM, realizar este procedimento 10 vezes e escolher a solução que produzir a maior log-verossimilhança.

A segunda estratégia, estratégia 2, utiliza um método K-médias para classificar inicialmente as observações pertencentes ao intervalo $(0,1)$ em dois grupos distintos, estimar os parâmetros como dados completos, e com estas estimativas iniciar o algoritmo EM.

A terceira estratégia, estratégia 3, utiliza os algoritmos SEM e CEM - são variações do algoritmo EM, em ambas um passo é adicionado entre os passos E e M, no SEM este novo passo é determinado de forma estocástica e no CEM de forma determinística - para posteriormente iniciar o EM. Esta estratégia consiste em separar, de forma aleatória, as observações pertencentes ao intervalo $(0,1)$ em dois grupos, assim estimar os parâmetros como dados completos, e com estas estimativas iniciar o algoritmo CEM, repetir este processo 5 vezes, realizar o mesmo procedimento com o SEM, dentre as 10 execuções escolher a solução que produzir o maior valor para a função log-verossimilhança, e por fim, com as estimativas da solução escolhida inciar o algoritmo EM.

As amostras foram simuladas de forma a preservar a variável latente que indica a origem da densidade componente de cada observação, desta maneira, possibilitando obter estimadores de máxima verossimilhança com dados completos para os parâmetros α, π, μ_1, ϕ_1, μ_2 e ϕ_2, assim utilizar tais estimativas como chute inicial para o algoritmo EM, e por fim tomar o valor da função log-verossimilhança e as estimativas obtidas pelo algoritmo EM como solução ideal.

Através do Algoritmo 2 simulamos uma amostra de tamanho 1000 para cada uma das 8 distribuições representadas na Figura 3. Para cada amostra obtivemos a solução ideal e estimamos os parâmetros pelas 3 diferentes estratégias relatadas. A Tabela 1 traz os resultados das estimativas para os parâmetros e os valores obtidos para a função log-verossimilhança da solução ideal e das 3 estratégias para os 8 cenários diferentes.

Devido ao estimador do parâmetro α possui solução analítica, em cada cenário o valor da estimativa deste parâmetro será a mesma em todas as estratégias realizadas, e por isso nos atentamos a avaliar as estimativas obtidas para os demais parâmetros.

Conforme mostra a Tabela 1, em todos os 8 cenários as estratégias 2 e 3 obtiveram valores para a função log-verossimilhança próximos do obtido pela solução ideal; já a estratégia 1 obteve valor da função log-verossimilhança próximo ao da solução ideal em 7 dos 8 cenários, apenas no primeiro cenário o valor ficou mais distante.

Algoritmo 2 – Simulação de observações de uma variável com distribuição beta bimodal inflacionada em zero

Entrada: Valores os parâmetros $\alpha, \pi, \mu_1, \phi_1, \mu_2$ e ϕ_2
Saída: Observações de uma variável aleatória com distribuição beta bimodal inflacionada em zero com parâmetros $\alpha, \pi, \mu_1, \phi_1, \mu_2$ e ϕ_2

Início
Forneça os valores para $\alpha, \pi, \mu_1, \phi_1, \mu_2, \phi_2$ e n (quantidade de observações desejada)
para i **faça** 1n
  gere uma variável $Z \sim Bernoulli(1-\alpha)$ ▷ $P(Z=1) = 1-\alpha$ e consequentemente $P(Z=0) = \alpha = P(Y=0)$.
  se $z = 0$ **então**
    $y_i = z = 0$
  senão
    gere uma variável $W \sim Bernoulli(\pi)$
    se $w = 1$ **então**
      gere $V \sim Beta(\mu_1, \phi_1)$ e atribua $y_i = v$
    senão
      gere $U \sim Beta(\mu_2, \phi_2)$ e atribua $y_i = u$
    fim se
  fim se
fim para
Fim
Retorna $y_1, y_2, ..., y_n$.

Nos cenários 5 e 6, observamos valores para as funções log-verossimilhança com pequena diferença entre as estratégias, porém apresentaram diferenças relevantes entre as estimativas dos parâmetros, isto não necessariamente mostra falta de identificabilidade, mas indica que na prática, diante destes cenários, as estimativas para os parâmetros podem mostrar instabilidade. Exemplo: no cenário 5, há uma pequena diferença entre o valor da função log-verossimilhança obtida pela estratégia 3, mas apresenta uma diferença relevante para a estimativa do parâmetro π obtida pela estratégia 3, $0,8061$, e obtida pela solução ideal, $0,392$; no cenário 6, observamos uma diferença relevante entre os valores dos parâmetros e as estimativas obtidas, inclusive as estimativas obtidas pela solução ideal, a dificuldade em obter estimativas mais precisas pode estar ligado com o fato destas densidades estarem sobrepostas, gerando assim duas distribuições difíceis de distinguir. Os cenários 5 e 6 correspondem as densidades $\boldsymbol{\vartheta}_5$ e $\boldsymbol{\vartheta}_6$ ilustradas na Figura 3, a densidade correspondente ao vetor $\boldsymbol{\vartheta}_5$ apresenta visual unimodalidade devido a proximidade dos parâmetros locação e relativa baixa precisão de ambas misturas beta, já a densidade $\boldsymbol{\vartheta}_6$ não apresenta moda no intervalo $(0,1)$ devido a valores baixos para os parâmetros de precisão das densidades que compõem a mistura.

Dentre as estratégias utilizadas, a estratégia 2, que utiliza método K-médias para inicializar o algoritmo EM, apresentou excelente desempenho, além de ser a que apresentou maior coerência com a solução ideal tanto dos valores das funções log-verossimilhança quanto das esti-

Tabela 1 – Resultado das 3 estratégias aplicadas em uma amostra de tamanho 1000 para diferentes distribuições.

	α	π	μ_1	ϕ_1	μ_2	ϕ_2	Log-ver.
Parâmetros cenário 1	0,25	0,40	0,30	10,00	0,90	12,00	
Solução ideal	0,2480	0,4016	0,3000	10,6426	0,8955	11,3800	-273,1847
Estratégia 1	0,2480	0,0000	0,6567	1,7928	0,6583	1,7968	-405,8042
Estratégia 2	0,2480	0,4016	0,3000	10,6445	0,8955	11,3781	-273,1846
Estratégia 3	0,2480	0,4016	0,3000	10,6441	0,8955	11,3785	-273,1846
Parâmetros cenário 2	0,25	0,10	0,30	10,00	0,90	12,00	
Solução ideal	0,2570	0,0929	0,2888	12,9202	0,8971	11,9301	132,3197
Estratégia 1	0,2570	0,0929	0,2887	12,9380	0,8971	11,9272	132,3199
Estratégia 2	0,2570	0,0929	0,2888	12,9203	0,8971	11,9301	132,3197
Estratégia 3	0,2570	0,0929	0,2889	12,9145	0,8971	11,9311	132,3196
Parâmetros cenário 3	0,25	0,80	0,30	10,00	0,90	12,00	
Solução ideal	0,2680	0,7855	0,3012	11,0740	0,9010	13,5735	-370,1852
Estratégia 1	0,2680	0,7855	0,3012	11,0728	0,9010	13,5785	-370,1852
Estratégia 2	0,2680	0,7855	0,3012	11,0746	0,9010	13,5710	-370,1853
Estratégia 3	0,2680	0,7855	0,3012	11,0740	0,9010	13,5735	-370,1852
Parâmetros cenário 4	0,25	0,40	0,40	10,00	0,80	12,00	
Solução ideal	0,2680	0,3661	0,4111	11,0410	0,7939	12,6570	-388,7692
Estratégia 1	0,2680	0,3661	0,4110	11,0440	0,7939	12,6544	-388,7691
Estratégia 2	0,2680	0,3661	0,4110	11,0453	0,7939	12,6533	-388,7691
Estratégia 3	0,2680	0,3661	0,4110	11,0425	0,7939	12,6557	-388,7692
Parâmetros cenário 5	0,25	0,40	0,50	10,00	0,70	12,00	
Solução ideal	0,2730	0,3920	0,5103	11,1851	0,7123	14,8357	-264,5521
Estratégia 1	0,2730	0,4938	0,5321	10,2814	0,7220	15,5501	-264,6264
Estratégia 2	0,2730	0,3989	0,5115	11,1979	0,7137	15,0020	-264,5582
Estratégia 3	0,2730	0,8061	0,5799	8,8319	0,7478	18,8685	-264,9037
Parâmetros cenário 6	0,25	0,40	0,30	3,00	0,80	5,00	
Solução ideal	0,2390	0,2891	0,1827	4,4067	0,7476	4,1449	-481,9071
Estratégia 1	0,2390	0,3469	0,2541	2,9274	0,7745	4,7461	-482,6322
Estratégia 2	0,2390	0,2760	0,1636	5,0599	0,7380	3,9326	-481,7758
Estratégia 3	0,2390	0,2681	0,1571	5,3181	0,7344	3,8550	-481,7573
Parâmetros cenário 7	0,25	0,40	0,30	13,00	0,80	15,00	
Solução ideal	0,2580	0,3854	0,3078	13,1572	0,7960	15,9444	-383,6949
Estratégia 1	0,2580	0,3854	0,3078	13,1607	0,7959	15,9417	-383,6948
Estratégia 2	0,2580	0,3854	0,3078	13,1623	0,7959	15,9405	-383,6948
Estratégia 3	0,2580	0,3854	0,3078	13,1623	0,7959	15,9405	-383,6948
Parâmetros cenário 8	0,20	0,30	0,40	30,00	0,90	45,00	
Solução ideal	0,2110	0,3194	0,3996	29,3403	0,9006	54,4586	235,1898
Estratégia 1	0,2110	0,3194	0,3996	29,3427	0,9006	54,4558	235,1898
Estratégia 2	0,2110	0,3194	0,3996	29,3404	0,9006	54,4585	235,1898
Estratégia 3	0,2110	0,3194	0,3996	29,3403	0,9006	54,4586	235,1898

mativas para os parâmetros, também apresentou melhor desempenho computacional, diferente das outras duas estratégias ele é executado uma única vez, e em geral, obteve convergência com menor numero de iterações.

Assim, devido os resultados expostos, recomendamos cuidado ao considerar este modelo para ajustar dados que não apresentem duas modas bem definidas no intervalo $(0,1)$, outra recomendação é utilizar o método K-médias para inicializar o algoritmo EM.

Na próxima seção realizamos estudo de simulação para avaliar o comportamento das estimativas para os parâmetros da distribuição beta bimodal inflacionada em zero com o aumento do tamanho amostral, utilizamos a estratégia 2 para inicialização do algoritmo EM.

4.2 Estimativas para distribuição BBZ

Realizamos experimento de simulação para ilustrar e comparar estimações de máxima verossimilhança obtidas a partir de amostras aleatórias da distribuição beta bimodal inflacionada em zero de tamanho n, com $n = 50, 200$ e 1000. Para tanto, cada amostra de tamanho n foi replicada 1000 vezes. Consideremos os parâmetros $\alpha, \pi, \mu_1, \phi_1, \mu_2, \phi_2$, cada replica foi obtida através da simulação de n observações da distribuição BBZ, assim temos $Y_i \sim BBIZ(\alpha, \pi, \mu_1, \phi_1, \mu_2, \phi_2)$ em que $i = 1, ..., n$, realizamos este procedimento considerando 8 configurações diferentes para os parâmetros, os resultados para cada cenário são mostrados nas Tabelas 2, 3, 4, 5, 6, 7, 8 e 9, estes cenários correspondem as distribuições ilustradas na Figura 3. Nestas tabelas apresentamos os resultados de estimativas para média $\hat{E}[\hat{\vartheta}] = m^{-1}\sum_{i=1}^{m}\vartheta_i$, viés $\hat{B}[\hat{\vartheta}] = \hat{E}[\hat{\vartheta}] - \vartheta$, viés relativo $\hat{B}_r[\hat{\vartheta}] = \hat{B}[\hat{\vartheta}]/\vartheta$, erro padrão $\hat{E}p = \sqrt{(m-1)^{-1}\sum_{i=1}^{m}(\hat{\vartheta}_i - \hat{E}[\hat{\vartheta}])^2}$ e raiz do erro quadrático médio REQM $= \sqrt{(m)^{-1}\sum_{i=1}^{m}(\hat{\vartheta}_i - \vartheta)^2}$, em que m quantidade de réplicas, ϑ valor verdadeiro do parâmetro e $\hat{\vartheta}$ seu estimador. Para a simulação de cada amostra veja o Algoritmo 2. Os valores dos parâmetros são considerados desconhecidos e assim os estimamos via método de máxima verossimilhança para cada uma das 1000 réplicas que compõem o conjunto de dados, neste caso, estimamos α diretamente e os demais parâmetros pelo algoritmo EM inicializado com estimativas obtidas após aplicar um método K-médias, referente a estratégia 2 utilizada na Seção 4.1.

Na Tabela 2 a distribuição BBZ, com vetor de parâmetros $\boldsymbol{\vartheta}_1 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0,3,\ \phi_1 = 10,\ \mu_2 = 0,9,\ \phi_2 = 12)$, possui duas distribuições beta com médias relativamente distantes e parâmetro π próximo a $0,5$, o que significa balanceamento das proporções da mistura de beta. Neste cenário, amostras de tamanho $n = 50$ já apresentam bons resultados, viés próximo a zero, para as estimativas dos parâmetros exceto para os parâmetros de precisão, ϕ_1 e ϕ_2, porém com o aumento de n há uma melhora substancial.

Reduzimos, com relação a $\boldsymbol{\vartheta}_1$, o valor de π para $0,1$ no vetor paramétrico $\boldsymbol{\vartheta}_2$ da distribuição, cujos resultados são apresentados na Tabela 3. Em geral, observamos uma piora

nas estimativas, especialmente para o parâmetro ϕ_1, mas não somente, π, μ_1 e ϕ_2 também apresentaram relativa piora. Visto que $\pi = 0,1$, numa amostra de tamanho $n = 50$ esperamos, em média 3 a 4 observações originadas na distribuição $Beta(\mu_1, \phi_1)$ e, provavelmente, a pequena quantidade de observações com origem nesta componente da mistura tenha provocado o referido problema nas estimativas. Cabe ressaltar ainda que, com tamanho amostral de 50 observações, em 44 das 1000 réplicas não foi possível obter estimativas, devido a problemas de convergência, o que também pode estar relacionado com uma quantidade insuficiente de observações com origem na distribuição $Beta(\mu_1, \phi_1)$.

Na Tabela 4 temos os resultados para a distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_3$, em que aumentamos π para $0,8$, neste cenário semelhante ao observado na simulação com $\boldsymbol{\vartheta}_2$, esperamos poucas observações com origem na distribuição $Beta(\mu_2, \phi_2)$, porém neste caso um pouco maior, o que provavelmente foi o que proporcionou melhores estimativas que no cenário com $\boldsymbol{\vartheta}_2$. Aqui, das 1000 réplicas apenas 1 não convergiu, com amostra de tamanho $n = 50$.

No cenário $\boldsymbol{\vartheta}_4$, resultados apresentados na Tabela 5, retornamos $\pi = 0,4$ e reduzimos a distância entre as médias, $\mu_1 = 0,4$ e $\mu_2 = 0,8$. Neste cenário há uma maior sobreposição das densidades componentes beta. Obtivemos estimativas com elevado erro padrão para os parâmetros ϕ_1 e ϕ_2, π teve uma leve piora.. Dente as 1000 réplicas para $n = 50$, apenas uma não obteve convergência.

A Tabela 6 traz os resultados do cenário com vetor de parâmetros $\boldsymbol{\vartheta}_5$, em que aproximamos as médias das distribuições beta, $\mu_1 = 0,5$ e $\mu_2 = 0,7$, como podemos observar na Figura 3, esta configuração dos parâmetros forma uma densidade unimodal no intervalo $(0,1)$ tamanha é a sobreposição das densidades componentes beta. Percebemos acentuado viés para os parâmetros de precisão, e um leve aumento do viés para o estimador de π com relação ao observado na Tabela 5, mas para os outros parâmetros não houve grande diferença na qualidade das estimativas. Dentre as 1000 réplicas, com amostras de tamanho $n = 50$, 17 não obtiveram convergência, já para o tamanho amostral $n = 200$ apenas uma não houve convergência.

No cenário para $\boldsymbol{\vartheta}_6$, veja Tabela 7, a distância entre as médias das distribuições beta é aumentada e reduz-se suas precisões de forma a obter uma mistura amodal no intervalo $(0,1)$, mesmo assim os vieses dos estimadores para os parâmetros de precisão são menores do que no cenário com vetor de parâmetros $\boldsymbol{\vartheta}_5$. Das 1000 réplicas de tamanho $n = 50$ apenas uma não obteve convergência.

Na Tabela 8 são apresentados os resultados da simulação para a distribuição com vetor de parâmetros $\boldsymbol{\vartheta}_7$, neste aumentamos as precisões, ϕ_1 para 13 e ϕ_2 para 15, e voltamos a observar bimodalidade em $(0,1)$. De forma geral, observamos expressiva redução dos vieses nos estimadores dos parâmetros.

Por fim na Tabela 9 temos os resultados referentes a distribuição com vetor de parâmetros $\boldsymbol{\vartheta}_8(\alpha = 0,2,\ \pi = 0,3,\ \mu_1 = 0,4,\ \phi_1 = 30,\ \mu_2 = 0,9,\ \phi_2 = 45)$. Aqui os estimadores dos

parâmetros μ_1, μ_2 e π possuem vieses semelhantes aos obtidos com o estimador de α, o qual possui solução analítica.

Observamos que as estimativas para $E[Y]$ e $Var[Y]$ não tiveram grandes diferenças entre as diversas distribuições utilizadas, como podemos ver nos gráficos da distribuição das estimativas nas Figuras 4 e 5.

De forma geral, se o parâmetro π está longe de $0,5$, significa que teremos poucas observações para uma das duas distribuições beta que compõem a mistura, assim para estimativas mais precisas precisamos de aumentar o tamanho da amostra. Quanto mais próximas as médias das distribuições beta estiverem mais difícil a separação das duas distribuições, o mesmo ocorre se tivermos pequenos valores para os parâmetros de precisão, e assim precisamos de uma quantidade maior de observações para boas estimativas.

Na próxima seção, estudo de simulação similar a este é realizado com a intenção de avaliar os estimadores de máxima verossimilhança para modelos de regressão regressão beta bimodal inflacionado em zero.

Tabela 2 – Resultado de 1000 réplicas das amostras de tamanho *n* da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_1 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0,3,\ \phi_1 = 10,\ \mu_2 = 0,9,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
	50	0,2481	-0,0019	-0,0075	0,0616	0,0616
$\alpha = 0,25$	200	0,2487	-0,0013	-0,0052	0,0306	0,0306
	1000	0,2501	0,0001	0,0005	0,0135	0,0135
	50	0,3982	-0,0018	-0,0045	0,0845	0,0844
$\pi = 0,4$	200	0,4004	0,0004	0,0011	0,0414	0,0414
	1000	0,3993	-0,0007	-0,0018	0,0181	0,0181
	50	0,3020	0,0020	0,0066	0,0497	0,0497
$\mu_1 = 0,3$	200	0,2997	-0,0003	-0,0009	0,0225	0,0225
	1000	0,2997	-0,0003	-0,0011	0,0095	0,0095
	50	20,3199	10,3199	1,0320	178,1171	178,3268
$\phi_1 = 10$	200	10,7518	0,7518	0,0752	2,6137	2,7185
	1000	10,1885	0,1885	0,0188	1,0201	1,0368
	50	0,8988	-0,0012	-0,0013	0,0262	0,0262
$\mu_2 = 0,9$	200	0,8996	-0,0004	-0,0004	0,0107	0,0107
	1000	0,8998	-0,0002	-0,0002	0,0047	0,0047
	50	15,1089	3,1089	0,2591	8,2478	8,8105
$\phi_2 = 12$	200	12,5652	0,5652	0,0471	2,6837	2,7413
	1000	12,0777	0,0777	0,0065	1,0635	1,0658
	50	0,4972	0,0022	0,0044	0,0550	0,0550
$E[Y] = 0,495$	200	0,4955	0,0005	0,0009	0,0275	0,0275
	1000	0,4951	0,0001	0,0001	0,0125	0,0125
	50	0,1521	-0,0032	-0,0206	0,0126	0,0130
$Var[Y] = 0,1553$	200	0,1543	-0,0010	-0,0063	0,0061	0,0061
	1000	0,1552	-0,0001	-0,0007	0,0026	0,0026

Tabela 3 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_2 = (\alpha = 0,25,\ \pi = 0,1,\ \mu_1 = 0,3,\ \phi_1 = 10,\ \mu_2 = 0,9,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2507	0,0007	0,0028	0,0625	0,0624
	200	0,2484	-0,0016	-0,0065	0,0297	0,0297
	1000	0,2503	0,0003	0,0012	0,0137	0,0137
$\pi = 0,1$	50	0,1438	0,0438	0,4381	0,1129	0,1211
	200	0,1080	0,0080	0,0798	0,0333	0,0342
	1000	0,1007	0,0007	0,0070	0,0142	0,0143
$\mu_1 = 0,3$	50	0,4012	0,1012	0,3372	0,2000	0,2240
	200	0,3574	0,0574	0,1914	0,1398	0,1510
	1000	0,3212	0,0212	0,0707	0,0867	0,0892
$\phi_1 = 10$	50	$\approx 1,9 \cdot 10^{12}$	$\approx 1,9 \cdot 10^{12}$	$\approx 1,9 \cdot 10^{11}$	$\approx 2,8 \cdot 10^{13}$	$\approx 2,8 \cdot 10^{13}$
	200	12,8615	2,8615	0,2861	13,4710	13,7650
	1000	10,1288	0,1288	0,0129	3,3327	3,3335
$\mu_2 = 0,9$	50	0,9050	0,0050	0,0056	0,0190	0,0197
	200	0,9014	0,0014	0,0016	0,0087	0,0088
	1000	0,9005	0,0005	0,0005	0,0040	0,0040
$\phi_2 = 12$	50	20,3132	8,3132	0,6928	39,0598	39,9147
	200	13,1321	1,1321	0,0943	2,7879	3,0077
	1000	12,3008	0,3008	0,0251	1,1542	1,1922
$E[Y] = 0,63$	50	0,6326	0,0026	0,0041	0,0582	0,0582
	200	0,6349	0,0049	0,0078	0,0298	0,0302
	1000	0,6318	0,0018	0,0029	0,0140	0,0141
$Var[Y] = 0,1627$	50	0,1593	-0,0034	-0,0207	0,0208	0,0211
	200	0,1610	-0,0017	-0,0102	0,0101	0,0102
	1000	0,1625	-0,0002	-0,0013	0,0045	0,0045

Tabela 4 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_3 = (\alpha = 0,25,\ \pi = 0,8,\ \mu_1 = 0,3,\ \phi_1 = 10,\ \mu_2 = 0,9,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2501	0,0001	0,0004	0,0619	0,0619
	200	0,2483	-0,0017	-0,0067	0,0309	0,0309
	1000	0,2502	0,0002	0,0008	0,0134	0,0134
$\pi = 0,8$	50	0,7868	-0,0132	-0,0165	0,0774	0,0785
	200	0,7969	-0,0031	-0,0038	0,0360	0,0361
	1000	0,8012	0,0012	0,0015	0,0152	0,0153
$\mu_1 = 0,3$	50	0,2959	-0,0041	-0,0135	0,0300	0,0303
	200	0,2991	-0,0009	-0,0028	0,0140	0,0140
	1000	0,2998	-0,0002	-0,0006	0,0061	0,0061
$\phi_1 = 10$	50	12,8107	2,8107	0,2811	6,9195	7,4653
	200	10,5527	0,5527	0,0553	1,9477	2,0237
	1000	10,0970	0,0970	0,0097	0,7139	0,7201
$\mu_2 = 0,9$	50	0,8689	-0,0311	-0,0345	0,0979	0,1027
	200	0,8873	-0,0127	-0,0141	0,0572	0,0585
	1000	0,8985	-0,0015	-0,0016	0,0170	0,0171
$\phi_2 = 12$	50	$\approx 1,3 \cdot 10^4$	$\approx 1,3 \cdot 10^4$	$\approx 1,1 \cdot 10^3$	$\approx 4 \cdot 10^5$	$\approx 4 \cdot 10^5$
	200	13,6140	1,6140	0,1345	6,4069	6,6039
	1000	12,2614	0,2614	0,0218	2,2228	2,2370
$E[Y] = 0,315$	50	0,3122	-0,0028	-0,0089	0,0430	0,0430
	200	0,3143	-0,0007	-0,0022	0,0222	0,0222
	1000	0,3140	-0,0010	-0,0030	0,0098	0,0098
$Var[Y] = 0,0888$	50	0,0855	-0,0032	-0,0365	0,0181	0,0184
	200	0,0874	-0,0014	-0,0155	0,0098	0,0098
	1000	0,0882	-0,0006	-0,0067	0,0041	0,0042

Tabela 5 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_4 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0,4,\ \phi_1 = 10,\ \mu_2 = 0,8,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2498	-0,0002	-0,0009	0,0622	0,0622
	200	0,2495	-0,0005	-0,0020	0,0305	0,0305
	1000	0,2497	-0,0003	-0,0014	0,0134	0,0134
$\pi = 0,4$	50	0,4043	0,0043	0,0108	0,1630	0,1630
	200	0,3976	-0,0024	-0,0061	0,0988	0,0988
	1000	0,3956	-0,0044	-0,0110	0,0400	0,0402
$\mu_1 = 0,4$	50	0,3994	-0,0006	-0,0014	0,0914	0,0913
	200	0,3996	-0,0004	-0,0010	0,0544	0,0544
	1000	0,4012	0,0012	0,0029	0,0229	0,0229
$\phi_1 = 10$	50	40,5912	30,5912	3,0591	241,9985	243,8041
	200	12,2375	2,2375	0,2237	7,4093	7,7363
	1000	10,2651	0,2651	0,0265	1,8191	1,8374
$\mu_2 = 0,8$	50	0,8006	0,0006	0,0008	0,0563	0,0563
	200	0,7998	-0,0002	-0,0003	0,0314	0,0314
	1000	0,8007	0,0007	0,0008	0,0135	0,0135
$\phi_2 = 12$	50	45,5185	33,5185	2,7932	447,7260	448,7553
	200	14,7807	2,7807	0,2317	32,3233	32,4266
	1000	12,3259	0,3259	0,0272	1,8062	1,8345
$E[Y] = 0,48$	50	0,4821	0,0021	0,0044	0,0502	0,0502
	200	0,4822	0,0022	0,0046	0,0254	0,0255
	1000	0,4824	0,0024	0,0050	0,0112	0,0114
$Var[Y] = 0,1177$	50	0,1159	-0,0018	-0,0151	0,0134	0,0135
	200	0,1175	-0,0002	-0,0018	0,0064	0,0064
	1000	0,1180	0,0003	0,0027	0,0028	0,0028

Tabela 6 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_5 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0.5,\ \phi_1 = 10,\ \mu_2 = 0,7,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2493	-0,0007	-0,0027	0,0627	0,0626
	200	0,2495	-0,0005	-0,0020	0,0303	0,0303
	1000	0,2493	-0,0007	-0,0029	0,0134	0,0134
$\pi = 0,4$	50	0,4203	0,0203	0,0507	0,2523	0,2530
	200	0,3749	-0,0251	-0,0627	0,2013	0,2027
	1000	0,3679	-0,0321	-0,0802	0,0827	0,0886
$\mu_1 = 0,5$	50	0,4760	-0,0240	-0,0480	0,1036	0,1063
	200	0,4967	-0,0033	-0,0066	0,0736	0,0736
	1000	0,5070	0,0070	0,0141	0,0292	0,0300
$\phi_1 = 10$	50	$\approx 7,2 \cdot 10^3$	$\approx 7,2 \cdot 10^3$	716,3321	$\approx 2,2 \cdot 10^5$	$\approx 2,2 \cdot 10^5$
	200	17,7909	7,7909	0,7791	40,3281	41,0539
	1000	10,2428	0,2428	0,0243	2,4224	2,4334
$\mu_2 = 0,7$	50	0,7192	0,0192	0,0275	0,0673	0,0700
	200	0,7024	0,0024	0,0035	0,0461	0,0461
	1000	0,7021	0,0021	0,0030	0,0182	0,0184
$\phi_2 = 12$	50	96,1115	84,1115	7,0093	999,4444	$\approx 10^3$
	200	18,5252	6,5252	0,5438	43,5576	44,0221
	1000	12,5303	0,5303	0,0442	1,8999	1,9715
$E[Y] = 0,465$	50	0,4677	0,0027	0,0058	0,0453	0,0453
	200	0,4700	0,0050	0,0107	0,0224	0,0229
	1000	0,4725	0,0075	0,0162	0,0108	0,0132
$Var[Y] = 0,0934$	50	0,0916	-0,0018	-0,0193	0,0129	0,0131
	200	0,0936	0,0003	0,0029	0,0064	0,0064
	1000	0,0947	0,0013	0,0139	0,0029	0,0032

Tabela 7 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_6 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0.3,\ \phi_1 = 3,\ \mu_2 = 0,8,\ \phi_2 = 5)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2501	0,0001	0,0002	0,0625	0,0625
	200	0,2492	-0,0008	-0,0032	0,0306	0,0306
	1000	0,2491	-0,0009	-0,0036	0,0142	0,0142
$\pi = 0,4$	50	0,3611	-0,0389	-0,0973	0,1832	0,1872
	200	0,3656	-0,0344	-0,0859	0,1328	0,1371
	1000	0,3737	-0,0263	-0,0657	0,0737	0,0782
$\mu_1 = 0,3$	50	0,2758	-0,0242	-0,0807	0,1391	0,1411
	200	0,2903	-0,0097	-0,0324	0,1143	0,1147
	1000	0,3017	0,0017	0,0056	0,0698	0,0698
$\phi_1 = 3$	50	30,2861	27,2861	9,0954	296,8398	297,9432
	200	5,2246	2,2246	0,7415	8,8136	9,0857
	1000	3,3453	0,3453	0,1151	1,3227	1,3664
$\mu_2 = 0,8$	50	0,7884	-0,0116	-0,0145	0,0788	0,0796
	200	0,7913	-0,0087	-0,0109	0,0540	0,0547
	1000	0,7996	-0,0004	-0,0005	0,0287	0,0287
$\phi_2 = 5$	50	14,9944	9,9944	1,9989	53,4602	54,3600
	200	6,4951	1,4951	0,2990	10,3051	10,4079
	1000	5,1742	0,1742	0,0348	1,0370	1,0510
$E[Y] = 0,45$	50	0,4563	0,0063	0,0140	0,0575	0,0578
	200	0,4608	0,0108	0,0239	0,0295	0,0314
	1000	0,4627	0,0127	0,0282	0,0137	0,0187
$Var[Y] = 0,1403$	50	0,1383	-0,0019	-0,0137	0,0126	0,0128
	200	0,1413	0,0011	0,0078	0,0063	0,0064
	1000	0,1420	0,0017	0,0123	0,0029	0,0034

Tabela 8 – Resultado de 1000 réplicas das amostras de tamanho *n* da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_7 = (\alpha = 0,25,\ \pi = 0,4,\ \mu_1 = 0,3,\ \phi_1 = 13,\ \mu_2 = 0,8,\ \phi_2 = 15)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,25$	50	0,2496	-0,0004	-0,0017	0,0632	0,0632
	200	0,2485	-0,0015	-0,0059	0,0313	0,0313
	1000	0,2498	-0,0002	-0,0007	0,0133	0,0133
$\pi = 0,4$	50	0,4031	0,0031	0,0077	0,0981	0,0981
	200	0,3974	-0,0026	-0,0065	0,0460	0,0460
	1000	0,3986	-0,0014	-0,0034	0,0207	0,0207
$\mu_1 = 0,3$	50	0,3046	0,0046	0,0154	0,0554	0,0556
	200	0,3001	0,0001	0,0005	0,0228	0,0228
	1000	0,3002	0,0002	0,0008	0,0096	0,0096
$\phi_1 = 13$	50	19,5622	6,5622	0,5048	17,5087	18,6898
	200	14,1314	1,1314	0,0870	3,9441	4,1013
	1000	13,2287	0,2287	0,0176	1,5141	1,5305
$\mu_2 = 0,8$	50	0,7999	-0,0001	-0,0001	0,0344	0,0344
	200	0,7994	-0,0006	-0,0008	0,0149	0,0149
	1000	0,8003	0,0003	0,0004	0,0059	0,0059
$\phi_2 = 15$	50	21,2522	6,2522	0,4168	22,8530	23,6818
	200	15,7263	0,7263	0,0484	3,3485	3,4247
	1000	15,1999	0,1999	0,0133	1,3769	1,3907
$E[Y] = 0,45$	50	0,4513	0,0013	0,0029	0,0512	0,0512
	200	0,4518	0,0018	0,0039	0,0254	0,0254
	1000	0,4509	0,0009	0,0019	0,0109	0,0110
$Var[Y] = 0,1215$	50	0,1190	-0,0025	-0,0206	0,0117	0,0119
	200	0,1208	-0,0007	-0,0060	0,0057	0,0057
	1000	0,1215	0,0000	0,0002	0,0025	0,0025

Tabela 9 – Resultado de 1000 réplicas das amostras de tamanho n da distribuição BBZ com vetor paramétrico $\boldsymbol{\vartheta}_8 = (\alpha = 0,2,\ \pi = 0,3,\ \mu_1 = 0,4,\ \phi_1 = 30,\ \mu_2 = 0,9,\ \phi_2 = 45)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\alpha = 0,2$	50	0,2015	0,0015	0,0075	0,0576	0,0576
	200	0,1995	-0,0005	-0,0026	0,0275	0,0275
	1000	0,1991	-0,0009	-0,0044	0,0124	0,0124
$\pi = 0,3$	50	0,3020	0,0020	0,0066	0,0718	0,0718
	200	0,2986	-0,0014	-0,0045	0,0371	0,0371
	1000	0,2994	-0,0006	-0,0019	0,0160	0,0160
$\mu_1 = 0,4$	50	0,4020	0,0020	0,0049	0,0314	0,0315
	200	0,4000	0,0000	0,0001	0,0135	0,0135
	1000	0,3999	-0,0001	-0,0003	0,0058	0,0058
$\phi_1 = 30$	50	42,4178	12,4178	0,4139	32,9375	35,1852
	200	32,0575	2,0575	0,0686	7,1070	7,3954
	1000	30,4499	0,4499	0,0150	2,8479	2,8818
$\mu_2 = 0,9$	50	0,9005	0,0005	0,0005	0,0084	0,0085
	200	0,9000	0,0000	0,0000	0,0042	0,0042
	1000	0,9000	-0,0000	-0,0000	0,0019	0,0019
$\phi_2 = 45$	50	51,3763	6,3763	0,1417	16,7579	17,9222
	200	46,4215	1,4215	0,0316	6,3627	6,5165
	1000	45,2342	0,2342	0,0052	2,8021	2,8105
$E[Y] = 0,6$	50	0,5989	-0,0011	-0,0019	0,0531	0,0531
	200	0,6009	0,0009	0,0016	0,0256	0,0256
	1000	0,6008	0,0008	0,0014	0,0116	0,0117
$Var[Y] = 0,135$	50	0,1326	-0,0024	-0,0178	0,0165	0,0167
	200	0,1342	-0,0007	-0,0055	0,0080	0,0081
	1000	0,1346	-0,0004	-0,0027	0,0036	0,0036

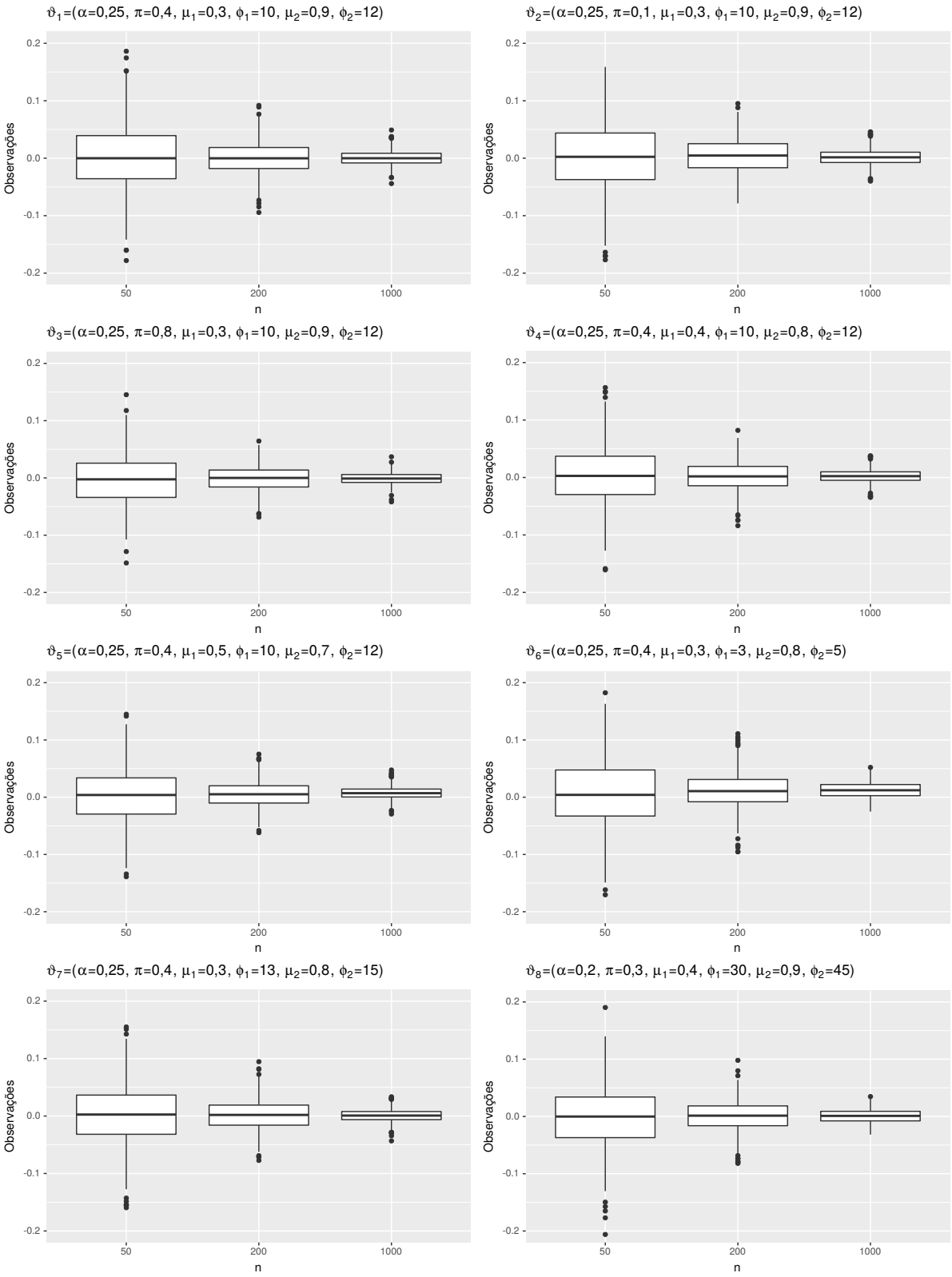

Figura 4 – Boxplot do viés das estimativas das esperanças de 1000 réplicas com tamanhos amostrais 50, 200 e 1000 para diferentes distribuições BBZ.

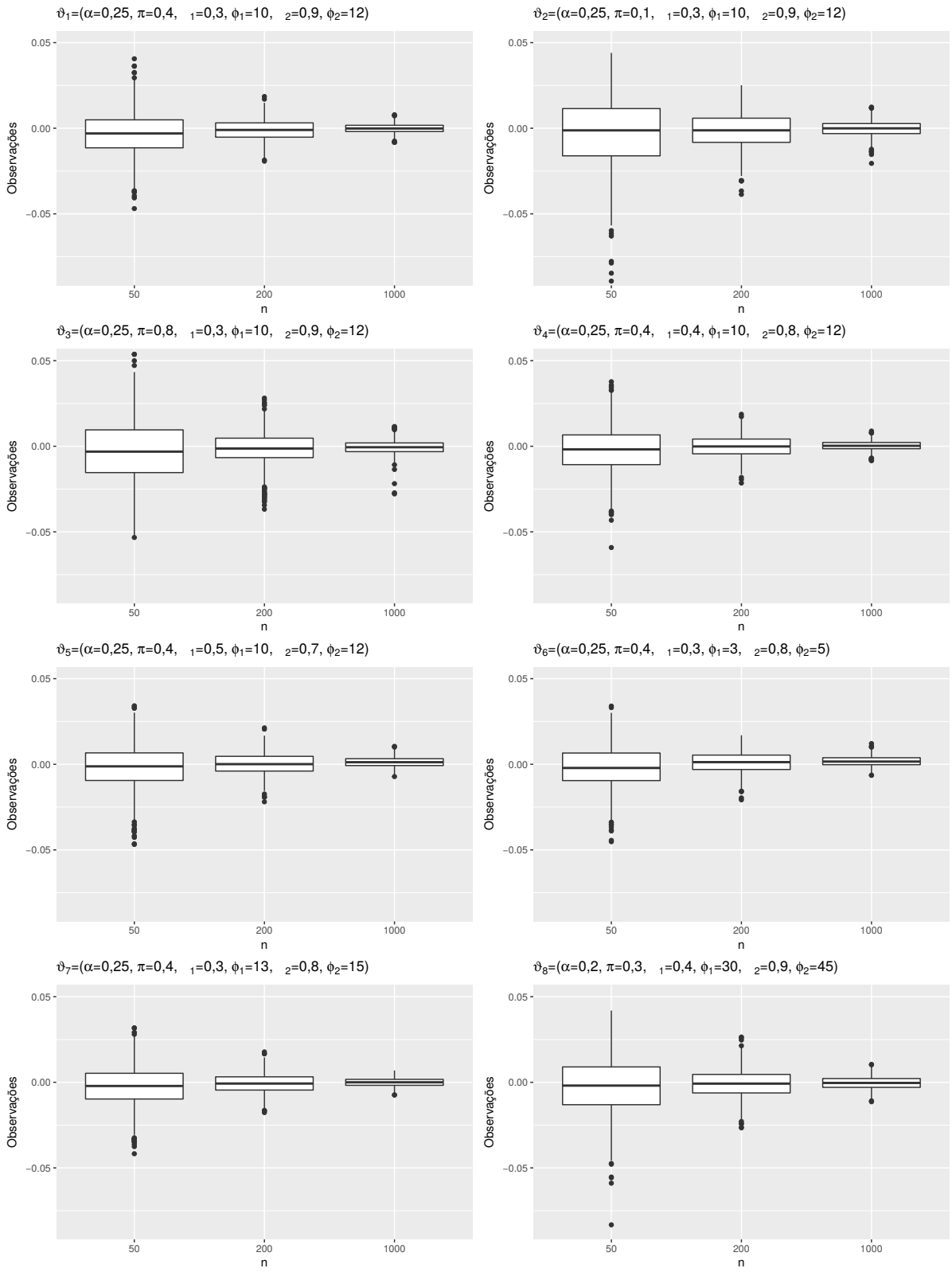

Figura 5 – Boxplot do viés das estimativas das variâncias de 1000 réplicas com tamanhos amostrais 50, 200 e 1000 para diferentes distribuições BBZ.

4.3 Estimativas para modelo de regressão BBZ

Realizamos estudo de simulação para avaliar o desempenho dos estimadores de máxima verossimilhança dos parâmetros do modelo de regressão regressão beta bimodal inflacionado em zero. Para tanto, simulamos amostras de tamanho $n = 200$ e 1000 para 8 diferentes cenários, para cada cenário e tamanho amostral foram realizadas 1000 réplicas. Diferente do estudo realizado na Seção 4.2, não consideramos amostras de tamanho 50, devido o aumento na quantidade de parâmetros a serem estimados. As Tabelas 10, 11, 12, 13, 14, 15, 16 e 17 trazem os resultados das estimativas para média, viés, viés relativo, erro padrão, raiz do erro quadrático médio de cada parâmetro estimado para cada cenário e considerando as 1000 réplicas. Cada amostra de tamanho n foi simulada considerando um modelo de regressão BBZ com os seguintes preditores lineares:

$$\begin{aligned} g_0(\alpha_i) &= \eta_{0i} = \beta_{00} + \beta_{01}x_{0i}, \\ g_1(\mu_{1i}) &= \eta_{1i} = \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i}, \\ g_2(\mu_{2i}) &= \eta_{2i} = \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i}, \\ g_3(\pi_i) &= \eta_{3i} = \beta_{30} + \beta_{31}x_{3i}, \end{aligned}$$

em que $g_k(\cdot)$, $k = 0,1,2$ e 3, funções de ligação logito. Assim a variável resposta $Y_i \sim BBZ(\boldsymbol{\vartheta}_i)$, $\boldsymbol{\vartheta}_i = (\alpha_i, \pi_i, \mu_{1i}, \phi_1, \mu_{2i}, \phi_2)$, $i = 1, ..., n$. Os valores das covariáveis x_{0i}, x_{1i}, x_{2i} e x_{3i} são realizações independentes de uma variável aleatória uniforme com parâmetros 0 e 1, $U(0,1)$. O Algoritmo 3 traz mais detalhes da simulação de cada amostra.

As estimativas de máxima verossimilhança dos parâmetros para cada amostra foram obtidas por meio da função `regbbz()` utilizando método K-médias para inicializar o algoritmo EM, mais detalhes sobre a função `regbbz()` veja Apêndice Subseção A.2.5. Com o propósito de evitar o problema de *label switching* nas réplicas, impomos uma restrição de ordem nos parâmetros, $\beta_{10} < \beta_{20}$, desta maneira, após a obtenção das estimativas, rearranjamos o vetor paramétrico de forma que $\hat{\beta}_{10} < \hat{\beta}_{20}$.

De forma geral, com o aumento do tamanho amostral n observamos redução do erro padrão, raiz do erro quadrático médio e valor absoluto dos vieses. As estimativas se mostraram satisfatórias com vieses relativo próximos a zero, para amostras de tamanho 1000, em todos os cenários, já para as amostras de tamanho 200 tivemos 2 cenários em que as estimativas para os parâmetros de precisão apresentaram elevado viés, veja nas Tabelas 14 e 15, entretanto para os demais parâmetros as estimativas apresentaram bom desempenho.

Na próxima seção simulamos uma base de dados de LGD e comparamos o desempenho da regressão beta bimodal inflacionada em zero com outras três regressões, a regressão beta, regressão beta inflacionada em zero e o *support vector regression machines*.

Algoritmo 3 – Simulação de observações de uma variável com distribuição beta bimodal inflacionada em zero considerando uma estrutura de regressão

Entrada: Valores os parâmetros $\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{30}, \beta_{31}, \phi_1, \phi_2$ e n.
Saída: Observações de uma variável aleatória com distribuição beta bimodal inflacionada em zero com parâmetros $\alpha_i, \pi_i, \mu_{1i}, \phi_1, \mu_{2i}$ e ϕ_2.

Início
Forneça os valores para os parâmetros $\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{30}, \beta_{31}, \phi_1, \phi_2$ e a quantidade de observações desejada n.
Gere n observações $U(0,1)$ para x_0, x_1, x_2 e x_3 de forma independente.
Calcule

$$\begin{aligned}
\alpha_i &= g_0^{-1}(\beta_{00} + \beta_{01}x_{0i}) \\
\mu_{1i} &= g_1^{-1}(\beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i}) \\
\mu_{2i} &= g_2^{-1}(\beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i}) \\
\pi_i &= g_3^{-1}(\beta_{30} + \beta_{31}x_{3i})
\end{aligned}$$

para $i = 1, 2, ..., n$.
para i **faça**1n
    gere uma variável $Z \sim Bernoulli(1-\alpha_i)$ ▷ $P(Z=1) = 1-\alpha_i$ e consequentemente $P(Z=0) = \alpha_i = P(Y_i = 0)$.
    se $z = 0$ **então**
        $y_i = z = 0$
    senão
        gere uma variável $W \sim Bernoulli(\pi_i)$
        se $w = 1$ **então**
            gere $V \sim Beta(\mu_{1i}, \phi_1)$ e atribua $y_i = v$
        senão
            gere $U \sim Beta(\mu_{2i}, \phi_2)$ e atribua $y_i = u$
        fim se
    fim se
fim para
Fim
Retorna $y_1, y_2, ..., y_n$.

Tabela 10 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_1 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -0,2,\ \beta_{10} = -0,2,\ \beta_{11} = -1,3,\ \beta_{12} = 0,\ \phi_1 = 10,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 2,8,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4185	-0,0185	0,0462	0,3450	0,3453
	1000	-0,4043	-0,0043	0,0107	0,1393	0,1393
$\beta_{01} = -1,4$	200	-1,3849	0,0151	-0,0108	0,6090	0,6089
	1000	-1,4002	-0,0002	0,0001	0,2643	0,2642
$\beta_{30} = -0,3$	200	-0,3119	-0,0119	0,0398	0,3620	0,3620
	1000	-0,2904	0,0096	-0,0318	0,1488	0,1490
$\beta_{31} = -0,2$	200	-0,1758	0,0242	-0,1209	0,6316	0,6317
	1000	-0,2178	-0,0178	0,0889	0,2562	0,2567
$\beta_{10} = -0,2$	200	-0,1783	0,0217	-0,1083	0,2859	0,2865
	1000	-0,1875	0,0125	-0,0623	0,1207	0,1213
$\beta_{11} = -1,3$	200	-1,3281	-0,0281	0,0216	0,3285	0,3296
	1000	-1,3093	-0,0093	0,0072	0,1456	0,1458
$\beta_{12} = 0$	200	-0,0135	-0,0135		0,3457	0,3458
	1000	-0,0139	-0,0139		0,1428	0,1434
$\phi_1 = 10$	200	11,2198	1,2198	0,1220	3,0203	3,2559
	1000	10,1801	0,1801	0,0180	1,0546	1,0694
$\beta_{20} = 0,8$	200	0,8107	0,0107	0,0134	0,2420	0,2421
	1000	0,8094	0,0094	0,0118	0,1068	0,1071
$\beta_{21} = 0$	200	-0,0107	-0,0107		0,3240	0,3240
	1000	-0,0076	-0,0076		0,1376	0,1377
$\beta_{22} = 2,8$	200	2,8165	0,0165	0,0059	0,3380	0,3383
	1000	2,8000	-0,0000	-0,0000	0,1437	0,1437
$\phi_2 = 12$	200	13,2297	1,2297	0,1025	2,9950	3,2362
	1000	12,2533	0,2533	0,0211	1,1586	1,1854

Tabela 11 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_2 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -3,8,\ \beta_{10} = -0,2,\ \beta_{11} = -1,3,\ \beta_{12} = 0,\ \phi_1 = 10,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 2,8,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4114	-0,0114	0,0285	0,3374	0,3374
	1000	-0,4060	-0,0060	0,0150	0,1329	0,1329
$\beta_{01} = -1,4$	200	-1,4159	-0,0159	0,0113	0,6109	0,6108
	1000	-1,3929	0,0071	-0,0051	0,2604	0,2604
$\beta_{30} = -0,3$	200	-0,1521	0,1479	-0,4930	0,6616	0,6776
	1000	-0,2982	0,0018	-0,0061	0,1923	0,1922
$\beta_{31} = -3,8$	200	-3,3980	0,4020	-0,1058	2,6605	2,6894
	1000	-3,8395	-0,0395	0,0104	0,5179	0,5192
$\beta_{10} = -0,2$	200	-0,1605	0,0395	-0,1976	0,5385	0,5397
	1000	-0,1918	0,0082	-0,0409	0,2159	0,2159
$\beta_{11} = -1,3$	200	-1,0981	0,2019	-0,1553	0,7890	0,8140
	1000	-1,3192	-0,0192	0,0148	0,2475	0,2481
$\beta_{12} = 0$	200	0,3835	0,3835		1,0665	1,1329
	1000	0,0009	0,0009		0,2585	0,2583
$\phi_1 = 10$	200	13,6048	3,6048	0,3605	7,9404	8,7167
	1000	10,6242	0,6242	0,0624	1,9438	2,0406
$\beta_{20} = 0,8$	200	0,8383	0,0383	0,0479	0,2325	0,2355
	1000	0,8033	0,0033	0,0042	0,0861	0,0861
$\beta_{21} = 0$	200	-0,1849	-0,1849		0,6605	0,6855
	1000	-0,0064	-0,0064		0,1083	0,1084
$\beta_{22} = 2,8$	200	2,5427	-0,2573	-0,0919	1,0352	1,0662
	1000	2,8046	0,0046	0,0016	0,1196	0,1196
$\phi_2 = 12$	200	12,5916	0,5916	0,0493	3,3743	3,4241
	1000	12,1415	0,1415	0,0118	0,8428	0,8541

Tabela 12 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_3 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = 3,4,\ \beta_{10} = -0,2,\ \beta_{11} = -1,3,\ \beta_{12} = 0,\ \phi_1 = 10,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 2,8,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,3781	0,0219	-0,0549	0,3286	0,3292
	1000	-0,4001	-0,0001	0,0003	0,1357	0,1357
$\beta_{01} = -1,4$	200	-1,4605	-0,0605	0,0432	0,5864	0,5892
	1000	-1,4094	-0,0094	0,0067	0,2636	0,2637
$\beta_{30} = -0,3$	200	-0,2933	0,0067	-0,0225	0,4316	0,4315
	1000	-0,3018	-0,0018	0,0061	0,1695	0,1694
$\beta_{31} = 3,4$	200	3,2421	-0,1579	-0,0464	1,6764	1,6830
	1000	3,4243	0,0243	0,0072	0,3976	0,3981
$\beta_{10} = -0,2$	200	-0,2205	-0,0205	0,1027	0,1985	0,1995
	1000	-0,1999	0,0001	-0,0005	0,0796	0,0796
$\beta_{11} = -1,3$	200	-1,2150	0,0850	-0,0654	0,4851	0,4923
	1000	-1,3026	-0,0026	0,0020	0,1011	0,1011
$\beta_{12} = 0$	200	0,1403	0,1403		0,7096	0,7230
	1000	0,0008	0,0008		0,0996	0,0995
$\phi_1 = 10$	200	10,4919	0,4919	0,0492	2,4446	2,4924
	1000	10,0859	0,0859	0,0086	0,6506	0,6559
$\beta_{20} = 0,8$	200	0,7612	-0,0388	-0,0485	0,4675	0,4688
	1000	0,7899	-0,0101	-0,0126	0,1716	0,1719
$\beta_{21} = 0$	200	-0,0612	-0,0612		0,6148	0,6175
	1000	0,0115	0,0115		0,2193	0,2195
$\beta_{22} = 2,8$	200	2,7713	-0,0287	-0,0103	0,8279	0,8280
	1000	2,8239	0,0239	0,0085	0,2223	0,2235
$\phi_2 = 12$	200	15,4901	3,4901	0,2908	7,3007	8,0887
	1000	12,5537	0,5537	0,0461	1,9784	2,0535

Tabela 13 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_4 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -0,2,\ \beta_{10} = -0,2,\ \beta_{11} = -0,4,\ \beta_{12} = 0,\ \phi_1 = 10,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 1,2,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4093	-0,0093	0,0232	0,3494	0,3494
	1000	-0,4009	-0,0009	0,0021	0,1381	0,1381
$\beta_{01} = -1,4$	200	-1,4192	-0,0192	0,0137	0,6116	0,6116
	1000	-1,4106	-0,0106	0,0076	0,2705	0,2706
$\beta_{30} = -0,3$	200	-0,3056	-0,0056	0,0187	0,5791	0,5788
	1000	-0,2853	0,0147	-0,0491	0,2110	0,2114
$\beta_{31} = -0,2$	200	-0,2177	-0,0177	0,0884	0,8115	0,8113
	1000	-0,2168	-0,0168	0,0842	0,3038	0,3041
$\beta_{10} = -0,2$	200	-0,1798	0,0202	-0,1010	0,3505	0,3509
	1000	-0,1936	0,0064	-0,0322	0,1511	0,1512
$\beta_{11} = -0,4$	200	-0,4109	-0,0109	0,0273	0,4042	0,4042
	1000	-0,4047	-0,0047	0,0117	0,1567	0,1567
$\beta_{12} = 0$	200	0,0184	0,0184		0,4391	0,4393
	1000	-0,0066	-0,0066		0,1565	0,1566
$\phi_1 = 10$	200	13,3296	3,3296	0,3330	11,5554	12,0200
	1000	10,3552	0,3552	0,0355	1,6222	1,6598
$\beta_{20} = 0,8$	200	0,7854	-0,0146	-0,0182	0,2647	0,2650
	1000	0,8000	0,0000	0,0000	0,1178	0,1177
$\beta_{21} = 0$	200	-0,0160	-0,0160		0,3209	0,3211
	1000	0,0007	0,0007		0,1340	0,1340
$\beta_{22} = 1,2$	200	1,2099	0,0099	0,0082	0,3976	0,3975
	1000	1,2068	0,0068	0,0057	0,1295	0,1297
$\phi_2 = 12$	200	13,6398	1,6398	0,1366	5,4135	5,6538
	1000	12,2938	0,2938	0,0245	1,4990	1,5267

Tabela 14 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_5 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -0,2,\ \beta_{10} = -0,2,\ \beta_{11} = 0,4,\ \beta_{12} = 0,\ \phi_1 = 10,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 0,1,\ \phi_2 = 12)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4100	-0,0100	0,0250	0,3414	0,3414
	1000	-0,3995	0,0005	-0,0012	0,1357	0,1357
$\beta_{01} = -1,4$	200	-1,4218	-0,0218	0,0156	0,6022	0,6023
	1000	-1,4083	-0,0083	0,0059	0,2620	0,2620
$\beta_{30} = -0,3$	200	-0,2146	0,0854	-0,2846	5,6784	5,6762
	1000	-0,2639	0,0361	-0,1204	0,6253	0,6261
$\beta_{31} = -0,2$	200	-0,3048	-0,1048	0,5239	10,1334	10,1288
	1000	-0,2141	-0,0141	0,0704	0,6467	0,6466
$\beta_{10} = -0,2$	200	-0,2827	-0,0827	0,4135	0,5261	0,5323
	1000	-0,2059	-0,0059	0,0294	0,2286	0,2285
$\beta_{11} = 0,4$	200	0,4956	0,0956	0,2391	0,6627	0,6693
	1000	0,4376	0,0376	0,0941	0,2277	0,2307
$\beta_{12} = 0$	200	0,0984	0,0984		0,5765	0,5845
	1000	-0,0107	-0,0107		0,2062	0,2064
$\phi_1 = 10$	200	147,2675	137,2675	13,7268	2332,8457	2335,7160
	1000	10,9780	0,9780	0,0978	3,8502	3,9706
$\beta_{20} = 0,8$	200	0,9218	0,1218	0,1523	0,4417	0,4580
	1000	0,8177	0,0177	0,0221	0,1717	0,1725
$\beta_{21} = 0$	200	-0,1153	-0,1153		0,5078	0,5205
	1000	-0,0325	-0,0325		0,1745	0,1774
$\beta_{22} = 0,1$	200	0,0082	-0,0918	-0,9178	0,5569	0,5641
	1000	0,1034	0,0034	0,0344	0,1652	0,1651
$\phi_2 = 12$	200	49,7182	37,7182	3,1432	277,8760	280,2865
	1000	12,7980	0,7980	0,0665	3,0765	3,1768

Tabela 15 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_6 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -0,2,\ \beta_{10} = -0,2,\ \beta_{11} = -1,3,\ \beta_{12} = 0,\ \phi_1 = 3,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 1,2,\ \phi_2 = 5)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4069	-0,0069	0,0174	0,3384	0,3383
	1000	-0,4006	-0,0006	0,0016	0,1293	0,1293
$\beta_{01} = -1,4$	200	-1,4148	-0,0148	0,0105	0,6049	0,6048
	1000	-1,4089	-0,0089	0,0064	0,2532	0,2532
$\beta_{30} = -0,3$	200	-0,3002	-0,0002	0,0007	0,9105	0,9101
	1000	-0,2892	0,0108	-0,0361	0,3229	0,3229
$\beta_{31} = -0,2$	200	-0,1326	0,0674	-0,3372	1,1378	1,1392
	1000	-0,2047	-0,0047	0,0234	0,3427	0,3425
$\beta_{10} = -0,2$	200	-0,2830	-0,0830	0,4149	0,6588	0,6637
	1000	-0,1911	0,0089	-0,0446	0,3122	0,3122
$\beta_{11} = -1,3$	200	-1,0696	0,2304	-0,1772	0,9972	1,0230
	1000	-1,2900	0,0100	-0,0077	0,3487	0,3487
$\beta_{12} = 0$	200	0,3187	0,3187		0,8994	0,9538
	1000	0,0192	0,0192		0,3248	0,3252
$\phi_1 = 3$	200	6,1718	3,1718	1,0573	13,4587	13,8209
	1000	3,2646	0,2646	0,0882	0,9381	0,9743
$\beta_{20} = 0,8$	200	0,7782	-0,0218	-0,0273	0,4060	0,4064
	1000	0,7709	-0,0291	-0,0364	0,1701	0,1725
$\beta_{21} = 0$	200	-0,2963	-0,2963		0,9345	0,9799
	1000	0,0014	0,0014		0,2422	0,2421
$\beta_{22} = 1,2$	200	0,9817	-0,2183	-0,1819	0,8386	0,8662
	1000	1,2099	0,0099	0,0082	0,2539	0,2540
$\phi_2 = 5$	200	6,0455	1,0455	0,2091	8,2974	8,3589
	1000	5,0993	0,0993	0,0199	0,8753	0,8805

Tabela 16 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_7 = (\beta_{00} = -0,4,\ \beta_{01} = -1,4,\ \beta_{30} = -0,3,\ \beta_{31} = -0,2,\ \beta_{10} = -0,2,\ \beta_{11} = -1,3,\ \beta_{12} = 0,\ \phi_1 = 13,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 1,2,\ \phi_2 = 15)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4016	-0,0016	0,0040	0,3364	0,3363
	1000	-0,4003	-0,0003	0,0007	0,1369	0,1368
$\beta_{01} = -1,4$	200	-1,4230	-0,0230	0,0164	0,6002	0,6003
	1000	-1,4022	-0,0022	0,0016	0,2679	0,2678
$\beta_{30} = -0,3$	200	-0,3059	-0,0059	0,0197	0,3746	0,3745
	1000	-0,2969	0,0031	-0,0103	0,1651	0,1651
$\beta_{31} = -0,2$	200	-0,2147	-0,0147	0,0737	0,6315	0,6313
	1000	-0,2126	-0,0126	0,0631	0,2804	0,2805
$\beta_{10} = -0,2$	200	-0,1811	0,0189	-0,0947	0,2841	0,2846
	1000	-0,1947	0,0053	-0,0266	0,1160	0,1161
$\beta_{11} = -1,3$	200	-1,3250	-0,0250	0,0192	0,3261	0,3269
	1000	-1,3033	-0,0033	0,0025	0,1333	0,1333
$\beta_{12} = 0$	200	-0,0129	-0,0129		0,3067	0,3068
	1000	-0,0075	-0,0075		0,1356	0,1357
$\phi_1 = 13$	200	14,7212	1,7212	0,1324	4,5816	4,8921
	1000	13,3255	0,3255	0,0250	1,5014	1,5355
$\beta_{20} = 0,8$	200	0,8057	0,0057	0,0071	0,2050	0,2050
	1000	0,8017	0,0017	0,0022	0,0914	0,0913
$\beta_{21} = 0$	200	-0,0002	-0,0002		0,2476	0,2475
	1000	0,0003	0,0003		0,1138	0,1137
$\beta_{22} = 1,2$	200	1,2062	0,0062	0,0052	0,2493	0,2493
	1000	1,1995	-0,0005	-0,0004	0,1053	0,1052
$\phi_2 = 15$	200	16,2930	1,2930	0,0862	3,7545	3,9692
	1000	15,2340	0,2340	0,0156	1,4103	1,4288

Tabela 17 – Resultado de 1000 réplicas das amostras de tamanho n da regressão BBZ com vetor paramétrico $\boldsymbol{\varphi}_8 = (\beta_{00} = -0,4,\ \beta_{01} = -1,8,\ \beta_{30} = -0,3,\ \beta_{31} = -1,\ \beta_{10} = -0,2,\ \beta_{11} = -0,5,\ \beta_{12} = 0,\ \phi_1 = 30,\ \beta_{20} = 0,8,\ \beta_{21} = 0,\ \beta_{22} = 3,\ \phi_2 = 45)$.

Parâmetro	n	Média	Viés	Viés rel.	Erro padrão	REQM
$\beta_{00} = -0,4$	200	-0,4136	-0,0136	0,0339	0,3498	0,3499
	1000	-0,4030	-0,0030	0,0075	0,1410	0,1410
$\beta_{01} = -1,8$	200	-1,8052	-0,0052	0,0029	0,6429	0,6426
	1000	-1,8037	-0,0037	0,0020	0,2757	0,2756
$\beta_{30} = -0,3$	200	-0,3016	-0,0016	0,0052	0,3416	0,3414
	1000	-0,3016	-0,0016	0,0053	0,1408	0,1408
$\beta_{31} = -1$	200	-1,0148	-0,0148	0,0148	0,6309	0,6308
	1000	-1,0023	-0,0023	0,0023	0,2671	0,2670
$\beta_{10} = -0,2$	200	-0,1948	0,0052	-0,0261	0,1530	0,1530
	1000	-0,2014	-0,0014	0,0069	0,0685	0,0684
$\beta_{11} = -0,5$	200	-0,4994	0,0006	-0,0013	0,1971	0,1970
	1000	-0,5000	-0,0000	0,0000	0,0858	0,0858
$\beta_{12} = 0$	200	-0,0114	-0,0114		0,1913	0,1916
	1000	0,0032	0,0032		0,0830	0,0830
$\phi_1 = 30$	200	33,5177	3,5177	0,1173	7,7777	8,5327
	1000	30,6473	0,6473	0,0216	2,8518	2,9230
$\beta_{20} = 0,8$	200	0,8048	0,0048	0,0060	0,1125	0,1126
	1000	0,8015	0,0015	0,0019	0,0533	0,0533
$\beta_{21} = 0$	200	-0,0002	-0,0002		0,1564	0,1563
	1000	-0,0057	-0,0057		0,0735	0,0736
$\beta_{22} = 3$	200	3,0028	0,0028	0,0009	0,1733	0,1733
	1000	3,0053	0,0053	0,0018	0,0809	0,0810
$\phi_2 = 45$	200	47,3262	2,3262	0,0517	7,0639	7,4337
	1000	45,5152	0,5152	0,0114	2,9905	3,0330

4.4 Comparação dos modelos, considerando dados simulados de LGD

Para comparar os modelos apresentados no Capítulo 2 com o modelo proposto no Capítulo 3, simulamos um banco de dados de LGD considerando três subpopulações diferentes quanto a capacidade de pagamento. Uma vez que a falta de capacidade de pagamento pode motivar a inadimplência, assumimos três níveis de capacidade de pagamento. As observações de $LGD = 0$ representam indivíduos que tiveram sua capacidade de pagamento levemente afetada, observações provenientes da distribuição $Beta(\mu_1, \phi_1)$ tiveram capacidade de pagamento afetada moderadamente, e as observações provenientes da subpopulação com distribuição $Beta(\mu_2, \phi_2)$ tiveram sua capacidade de pagamento afetada severamente. A capacidade de pagamento não é uma variável observada diretamente mas sabemos que pode ser afetada por um descontrole financeiro, problemas pessoais - como a perda do emprego - entre outros. Para representar este comportamento criamos as seguintes variáveis regressoras:

- v_1: situação emprego, variável $v_1 = 1$ se empregado e $v_1 = 0$ se desempregado, $v_1 \in \{0,1\}$ e segue distribuição uniforme discreta;
- v_2: qualificação profissional, $v_2 = 1$ se profissional qualificado e $v_2 = 0$ se não possui qualificação, $v_2 \in \{0,1\}$ e segue distribuição uniforme discreta;
- v_3: percentual pago de toda a dívida calculado no momento de descumprimento, $v_3 \in (0,1)$ e segue distribuição uniforme;
- v_4: outras dívidas, $v_4 = 0$ se não possui dívidas, $v_4 = 1$ se possui de 1 até 3 dívidas e $v_4 = 2$ se possui mais de 4 dívidas, $v_4 \in \{0,1,2\}$ e segue distribuição uniforme discreta;
- v_5: histórico de pagamentos (quantidade de atrasos), $v_5 = 0$ se não possui atrasos, $v_5 = 1$ se possui de 1 até 3 atrasos e $v_5 = 2$ mais de 4 atrasos em seu histórico, $v_5 \in \{0,1,2\}$ e segue distribuição uniforme discreta;
- v_6: garantias, $v_6 = 0$ operação não possui garantias, $v_6 = 1$ operação com garantias fidejussórias[1], $v_6 = 2$ garantias reais[2], $v_6 \in \{0,1,2\}$ e segue distribuição uniforme discreta;
- v_7: razão entre o valor da parcela sobre o valor da renda mensal apurada no momento da concessão do crédito, consideramos que a política de crédito da instituição não permite operação com parcela maior ou igual a 30% da renda, $v_7 \in (0,3)$ e segue distribuição uniforme;

[1] Garantias fidejussórias, ou garantias pessoais, quando alguém assume obrigação, ao garantir o cumprimento de obrigação alheia. Exemplos: fiança, aval, etc.

[2] Garantias reais, quando alguma propriedade é destinada para assegurar o cumprimento da obrigação contraída.

- v_8: montante do crédito tomado, considerando que a linha de crédito possui valor mínimo igual a R$1.000,00 e máximo de R$100.000,00, $v_8 \in (1.000, 100.000)$ e segue distribuição uniforme;
- v_9: propriedades, $v_9 = 0$ não possui propriedades, $v_9 = 1$ possui propriedades, $v_9 \in \{0,1\}$ e segue distribuição uniforme discreta;
- v_{10}: faixa etária do tomador no momento do *default*, $v_{10} = 0$ de 18 a 30 anos, $v_{10} = 1$ de 31 a 50 anos e $v_{10} = 2$ se idade acima de 50 anos, $v_{10} \in \{0,1,2\}$ e segue distribuição uniforme discreta;
- v_{11}: estado civil, $v_{11} = 0$ se solteiro, $v_{11} = 1$ indivíduo casado e $v_{11} = 2$ para os demais casos, $v_{11} \in \{0,1,2\}$ e segue distribuição uniforme discreta.

Para as variáveis v_4, v_5, v_6, v_{10} e v_{11} foram obtidas variáveis *dummy* da seguinte forma: de v_4 obtemos $v_{4.1} = 1$ se não possui dívidas, $v_{4.1} = 0$ caso contrário, e $v_{4.2} = 1$ se possui de 1 até 3 dívidas, $v_{4.2} = 0$ caso contrário; de v_5 obtemos $v_{5.1} = 1$ se não possui atraso, $v_{5.1} = 0$ caso contrário, e $v_{5.2} = 1$ se possui de 1 a 3 atrasos, $v_{5.2} = 0$ caso contrário; de v_6 obtemos $v_{6.1} = 1$ se não possui garantias, $v_{6.1} = 0$ caso contrário, e $v_{6.2} = 1$ se possui garantias fidejussórias, $v_{6.2} = 0$ caso contrário; de v_{10} obtemos $v_{10.1} = 1$ se idade de 18 a 30 anos, $v_{10.1} = 0$ caso contrário, e $v_{10.2} = 1$ se idade de de 31 a 50 anos, $v_{10.2} = 0$ caso contrário; e de v_{11} obtemos $v_{11.1} = 1$ se estado civil solteiro, $v_{11.1} = 0$ caso contrário, e $v_{11.2} = 1$ se casado, $v_{11.2} = 0$ caso contrário. Para cada uma destas variáveis foram obtidas 10 mil realizações, assim a partir do conjunto de covariáveis obtidas geramos 10 mil observações de LGD, assumindo que a variável resposta $\text{LGD}_i \sim BBZ(\alpha_i, \pi_i, \mu_{1i}, \phi_1, \mu_{2i}, \phi_2)$, $i = 1, ..., 10.000$, com

$$
\begin{aligned}
g_0(\alpha_i) &= -4 + 3v_{1i} + 2v_{2i},\\
g_1(\mu_{1i}) &= -1,3 - v_{3i} - 0,8v_{4.1i} - 0,4v_{4.2i} + 0,9v_{5.1i} + 0,45v_{5.2i} + 0,5v_{6.1i} + 0,45v_{6.2i} + 2v_{7i},\\
g_2(\mu_{2i}) &= 1 - 1,2v_{3i} - 1,1v_{4.1i} - 0,55v_{4.2i} + v_{5.1i} + 0,5v_{5.2i} + 0,9v_{6.1i} + 0,45v_{6.2i} + 2,5v_{7i},\\
g_3(\pi_i) &= -3.5 + 4v_{1i} + 3v_{2i},
\end{aligned}
$$

$\phi_1 = 50$, $\phi_2 = 60$, e as funções g_0, g_1, g_2 e g_3 são funções de ligação logito. Obtemos assim um banco de dados simulado em que nem todas as covariáveis disponíveis são importantes, as covariáveis v_8, v_9, v_{10} e v_{11} não tem ligação com nenhum dos parâmetros. A Figura 6 mostra o histograma das observações simuladas de LGD.

Devido a presença relevante de observações iguais a zero, talvez na Figura 6 não fique evidente a existência de duas modas no intervalo $(0,1)$, neste sentido a Figura 7 traz histograma da observações excluindo as observações iguais a zero, a qual deixa claro a bimodalidade destas observações no intervalo $(0,1)$.

O conjunto de dados foi particionado em três subconjuntos: dados para desenvolvimento (8 mil observações), dados para validação (1 mil observações) e dados para teste (1 mil observações). As base de dados desenvolvimento e validação foram utilizados para obter um modelo

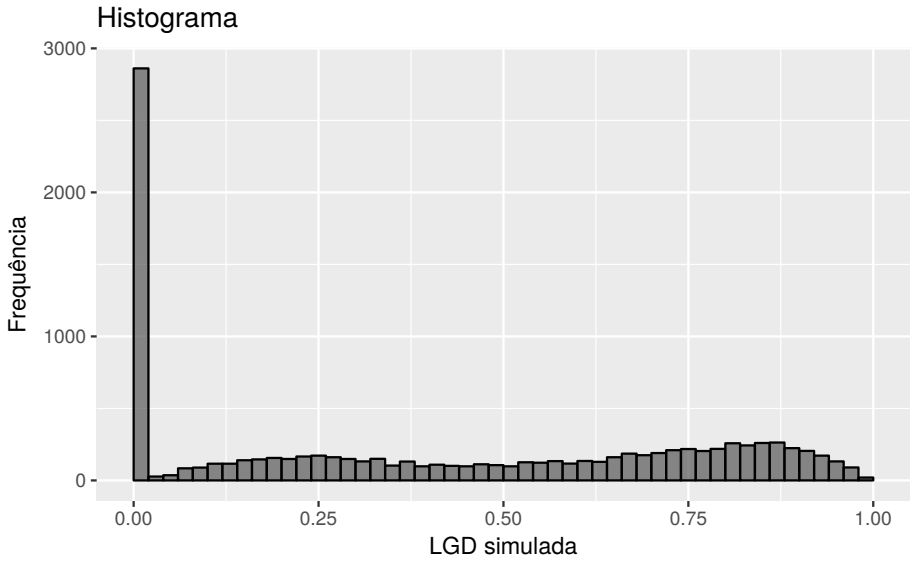

Figura 6 – Histograma para dados simulados de LGD.

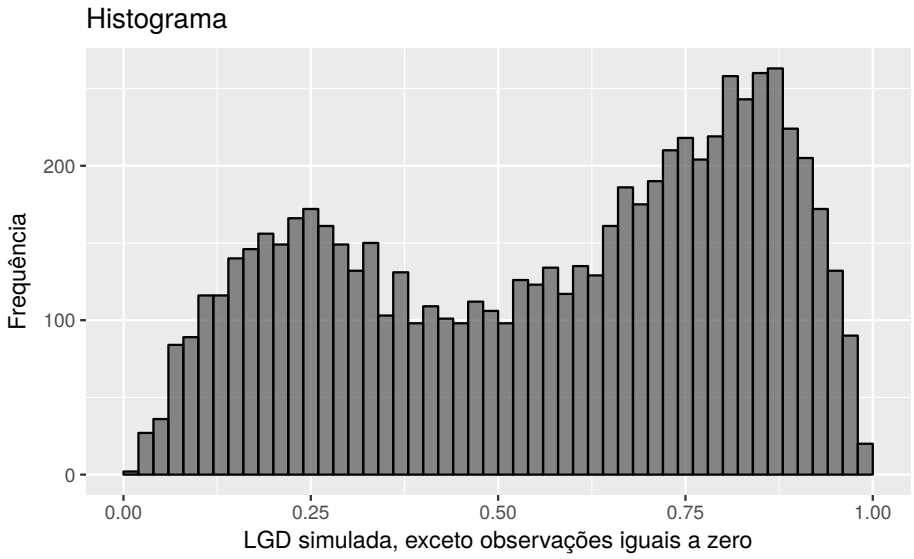

Figura 7 – Histograma para dados simulados de LGD, excluídas as observações iguais a zero.

final para cada um dos modelos apresentados no Capítulo 2 - regressão beta, regressão beta inflacionada em zero, *support vector regression machines* - e o modelo proposto no Capítulo 3, regressão beta bimodal inflacionado em zero. A base de dados de teste foi utilizada para a comparação entre os modelos.

Para aplicação do modelo de regressão beta foi necessário atribuir um pequeno valor (utilizamos 10^{-4}) às observações de LGD iguais a zero, devido este modelo ser definido apenas para variáveis respostas no intervalo $(0, 1)$. Primeiramente obtivemos o modelo de regressão beta, considerando todas as variáveis, modelo de regressão beta completo, veja Tabela 18.

Tabela 18 – Estimativas dos parâmetros do modelo de regressão beta completo.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor z	$P(>\|z\|)$
	Intercepto	0,4425	0,0615	7,1979	***
	v_1	-1,8318	0,0280	-65,3595	***
	v_2	-0,9265	0,0261	-35,5126	***
	v_3	-0,3970	0,0441	-9,0120	***
	$v_{4.1}$	-0,3221	0,0312	-10,3075	***
	$v_{4.2}$	-0,1839	0,0313	-5,8842	***
	$v_{5.1}$	0,2807	0,0311	9,0193	***
	$v_{5.2}$	0,1742	0,0314	5,5522	***
$g(\mu_i)$	$v_{6,1}$	0,3037	0,0313	9,6967	***
	$v_{6.2}$	0,1599	0,0312	5,1287	***
	v_7	0,8811	0,1474	5,9779	***
	v_8	-0,0000	0,0000	-0,1564	0,8757
	v_9	0,0352	0,0255	1,3817	0,1671
	$v_{10.1}$	0,0217	0,0312	0,6957	0,4866
	$v_{10.2}$	0,0213	0,0311	0,6844	0,4937
	$v_{11.1}$	-0,0235	0,0310	-0,7578	0,4486
	$v_{11.2}$	0,0246	0,0313	0,7854	0,4322
ϕ	Intercepto	1,5266	0,0229	66,6065	***

Posteriormente, ajustamos outro modelo de regressão beta considerando apenas as variáveis que apresentaram p-valor abaixo de 0,01 no modelo contendo todas as variáveis, as estimativas para os parâmetros do modelo de regressão beta reduzido estão na Tabela 19. A função de ligação logito foi utilizada em ambos modelos de regressão beta ajustados, o software R foi utilizados nos ajustes destes modelos. Para obtenção das estimativas utilizamos a função `betareg()` do pacote `betareg`, erro padrão e p-valor foram obtidos com a aplicação da função `summary()` nos modelos ajustados.

Tabela 19 – Estimativas dos parâmetros do modelo de regressão beta reduzido.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor z	$P(>\|z\|)$
	Intercepto	0,4727	0,0491	9,6222	***
	v_1	-1,8318	0,0280	-65,3781	***
	v_2	-0,9277	0,0261	-35,5836	***
	v_3	-0,3982	0,0440	-9,0395	***
$g(\mu_i)$	$v_{4,1}$	-0,3216	0,0312	-10,2903	***
	$v_{4.2}$	-0,1829	0,0313	-5,8504	***
	$v_{5.1}$	0,2794	0,0311	8,9782	***
	$v_{5.2}$	0,1743	0,0314	5,5569	***
	$v_{6.1}$	0,3036	0,0313	9,6965	***
	$v_{6.2}$	0,1601	0,0312	5,1381	***
	v_7	0,8757	0,1474	5,9430	***
ϕ	Intercepto	1,5258	0,0229	66,6123	***

Como critério para seleção entre estes dois modelos utilizamos o AIC, calculamos

também a estimativa para a função log-verossimilhança na amostra de validação. Os valores para AIC no modelo com todas as variáveis foi de $-22.162,06$ e no modelo com menos variáveis $-22.169,1$. O valor para log-verossimilhança na amostra de validação do modelo beta completo foi $1.601,836$ e no modelo beta reduzido $1.601,527$. Considerando AIC menor no segundo modelo e a pequena diferença entre os valores do log-verossimilhança na amostra de validação, mantemos o segundo modelo.

A Tabela 20, traz estimativas do modelo de regressão beta inflacionado em zero considerando todas as variáveis, modelo completo.

Tabela 20 – Estimativas dos parâmetros do modelo de regressão beta inflacionado em zero completo.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor t	$P(>\|t\|)$
	Intercepto	-4,0740	0,1598	-25,4981	***
	v_1	3,1599	0,0825	38,2993	***
	v_2	2,0078	0,0685	29,3000	***
	v_3	0,2055	0,1091	1,8836	0,0597
	$v_{4.1}$	0,0201	0,0778	0,2589	0,7958
	$v_{4.2}$	0,0307	0,0773	0,3975	0,6910
	$v_{5.1}$	0,0028	0,0774	0,0365	0,9709
	$v_{5.2}$	-0,0757	0,0782	-0,9676	0,3333
$g_0(\alpha_i)$	$v_{6.1}$	-0,1792	0,0780	-2,2978	0,0216
	$v_{6.2}$	-0,0557	0,0773	-0,7209	0,4710
	v_7	-0,1820	0,3682	-0,4944	0,6210
	v_8	0,0000	0,0000	0,1244	0,9010
	v_9	-0,1028	0,0634	-1,6208	0,1051
	$v_{10.1}$	-0,0417	0,0776	-0,5369	0,5913
	$v_{10.2}$	-0,1017	0,0776	-1,3114	0,1898
	$v_{11.1}$	0,1510	0,0771	1,9588	0,0502
	$v_{11.2}$	0,0105	0,0783	0,1334	0,8939
	Intecepto	0,6772	0,0466	14,5293	***
	v_1	-1,1291	0,0236	-47,9422	***
	v_2	-0,6121	0,0224	-27,2901	***
	v_3	-0,8694	0,0371	-23,4554	***
	$v_{4.1}$	-0,7785	0,0262	-29,7422	***
	$v_{4.2}$	-0,3746	0,0263	-14,2478	***
	$v_{5.1}$	0,7604	0,0263	28,9156	***
	$v_{5.2}$	0,3796	0,0259	14,6675	***
$g_1(\mu_i)$	$v_{6.1}$	0,6264	0,0263	23,8089	***
	$v_{6.2}$	0,3825	0,0262	14,6124	***
	v_7	1,9037	0,1224	15,5590	***
	v_8	0,0000	0,0000	0,0955	0,9239
	v_9	-0,0014	0,0213	-0,0659	0,9475
	$v_{10.1}$	0,0136	0,0260	0,5247	0,5998
	$v_{10.2}$	-0,0173	0,0259	-0,6663	0,5052
	$v_{11.1}$	0,0449	0,0260	1,7256	0,0845
	$v_{11.2}$	0,0460	0,0258	1,7784	0,0754
ϕ	Intercepto	5,6251	0,1046	53,8	***

Na sequência, outro modelo de regressão beta inflacionado em zero foi ajustado considerando apenas as variáveis que apresentaram p-valor abaixo de 0,01 obtido no modelo contendo todas as variáveis, as estimativas para os parâmetros do modelo de regressão beta inflacionado em zero reduzido são mostradas na Tabela 21. A função de ligação logito foi utilizada em ambos modelos de regressão beta inflacionados em zero ajustados, utilizamos a função `gamlss()` do pacote `gamlss` do software R para obtenção das estimativas e a função `summary()` para obter erros padrão e p-valores.

Tabela 21 – Estimativas dos parâmetros do modelo de regressão beta inflacionado em zero reduzido.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor t	$P(>\|t\|)$
	Intercepto	-4,0963	0,0875	-46,7920	***
$g_0(\alpha_i)$	v_1	3,1492	0,0792	39,7661	***
	v_2	1,9991	0,0673	29,7100	***
	Intecepto	0,7206	0,0399	18,0435	***
	v_1	-1,1280	0,0235	-47,9741	***
	v_2	-0,6121	0,0224	-27,3106	***
	v_3	-0,8695	0,0371	-23,4628	***
$g_1(\mu_i)$	$v_{4.1}$	-0,7788	0,0262	-29,7628	***
	$v_{4.2}$	-0,3741	0,0263	-14,2284	***
	$v_{5.1}$	0,7599	0,0263	28,9160	***
	$v_{5.2}$	0,3788	0,0259	14,6337	***
	$v_{6.1}$	0,6281	0,0263	23,8849	***
	$v_{6.2}$	0,3826	0,0262	14,6156	***
	v_7	1,9024	0,1223	15,5509	***
ϕ	Intercepto	5,61850	0,09893	56,79	***

Como critério para seleção entre os dois modelos foi utilizado o AIC, calculamos também as estimativas para a função log-verossimilhança na amostra de validação. O valor para o AIC do modelo beta inflacionado em zero completo foi de $1758,329$ e para o modelo reduzido foi de $1744,617$. O valor para log-verossimilhança na amostra de validação do modelo completo foi $-184,695$ e no modelo reduzido $-182,7251$. Selecionamos o modelo reduzido, considerando que este obteve menor valor de AIC e maior valor de log-verossimilhança na amostra de validação.

Na obtenção dos modelos de regressão beta bimodal inflacionado em zero, ajustamos um modelo com todas as variáveis para os preditores $g_0(\alpha_i), g_3(\pi_i), g_1(\mu_{1i})$ e $g_1(\mu_{2i})$, modelo beta bimodal inflacionado em zero completo, estimativas para este modelo estão nas Tabelas 22 e 23.

Em seguida, consideramos um segundo modelo de regressão beta bimodal inflacionado em zero contendo apenas os parâmetros que apresentaram p-valor abaixo de 0,01 no modelo de regressão beta bimodal inflacionado em zero completo. As estimativas para este modelo reduzido são apresentadas na Tabela 24.

As estimativas para os parâmetros do modelo de regressão beta bimodal inflacionado em zero foram obtidas com auxílio do software R através da função `regbbz()` descrita no apêndice,

Tabela 22 – Estimativas dos parâmetros do modelo de regressão beta bimodal inflacionado em zero completo.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor z	$P(>\|z\|)$
	Intercepto	-4,0740	0,1671	-24,3740	***
	v_1	3,1599	0,0796	39,7195	***
	v_2	2,0078	0,0677	29,6470	***
	v_3	0,2055	0,1091	1,8833	0,0597
	$v_{4.1}$	0,0201	0,0778	0,2588	0,7958
	$v_{4.2}$	0,0307	0,0773	0,3975	0,6910
	$v_{5.1}$	0,0028	0,0774	0,0365	0,9709
	$v_{5.2}$	-0,0757	0,0782	-0,9674	0,3333
$g_0(\alpha_i)$	$v_{6.1}$	-0,1792	0,0780	-2,2978	0,0216
	$v_{6.2}$	-0,0557	0,0773	-0,7208	0,4711
	v_7	-0,1820	0,3682	-0,4944	0,6210
	v_8	0,0000	0,0000	0,3026	0,7622
	v_9	-0,1028	0,0634	-1,6209	0,1050
	$v_{10.1}$	-0,0417	0,0776	-0,5369	0,5913
	$v_{10.2}$	-0,1017	0,0776	-1,3114	0,1897
	$v_{11.1}$	0,1510	0,0771	1,9590	0,0501
	$v_{11.2}$	0,0105	0,0783	0,1334	0,8939
	(Intercept)	-1,2953	0,0348	-37,2282	***
	v_1	0,0015	0,0181	0,0809	0,9355
	v_2	0,0261	0,0173	1,5117	0,1306
	v_3	-1,0131	0,0255	-39,6605	***
	$v_{4.1}$	-0,8364	0,0184	-45,5623	***
	$v_{4.2}$	-0,3756	0,0170	-22,0587	***
	$v_{5.1}$	0,9000	0,0185	48,7219	***
	$v_{5.2}$	0,4390	0,0190	23,1054	***
$g_1(\mu_{1i})$	$v_{6.1}$	0,5389	0,0183	29,4503	***
	$v_{6.2}$	0,4638	0,0184	25,1502	***
	v_7	1,8675	0,0844	22,1156	***
	v_8	-0,0000	-	-	-
	v_9	0,0191	0,0146	1,3119	0,1896
	$v_{10.1}$	-0,0334	0,0179	-1,8656	0,0621
	$v_{10.2}$	-0,0251	0,0176	-1,4242	0,1544
	$v_{11.1}$	0,0085	0,0178	0,4774	0,6331
	$v_{11.2}$	0,0105	0,0177	0,5914	0,5542
$\log(\phi_1)$	Intercepto	3,9038	0,0251	155,2500	***

Subseção A.2.5, e a função `summary()` forneceu os erros padrão e p-valores. Como critério para seleção entre os dois modelos utilizamos o AIC, calculamos também a estimativa para a função log-verossimilhança na amostra de validação. O valor para o AIC do modelo de regressão beta bimodal inflacionado em zero completo foi de $-6149,988$ e o modelo reduzido foi de $-6190,082$. O valor para log-verossimilhança na amostra de validação do modelo completo foi $297,9921$ e no modelo reduzido $301,6612$. Selecionamos o modelo reduzido, considerando que este obteve menor valor de AIC e maior valor de log-verossimilhança na amostra de validação.

Tabela 23 – Estimativas dos parâmetros do modelo de regressão beta bimodal inflacionado em zero completo (Continuação da Tabela 22).

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor z	$P(>\|z\|)$
	Intercepto	0,9866	0,0215	45,8382	***
	v_1	-0,0173	0,0145	-1,1906	0,2338
	v_2	-0,0073	0,0113	-0,6495	0,5160
	v_3	-1,1704	0,0179	-65,2948	***
	$v_{4.1}$	-1,0767	0,0128	-84,1269	***
	$v_{4.2}$	-0,5342	0,0133	-40,0487	***
	$v_{5.1}$	0,9992	0,0127	78,7966	***
	$v_{5.2}$	0,5234	0,0120	43,5047	***
$g_2(\mu_{2i})$	$v_{6.1}$	0,9026	0,0128	70,7314	***
	$v_{6.2}$	0,4302	0,0120	35,8490	***
	v_7	2,4292	0,0587	41,3932	***
	v_8	-0,0000	-	-	-
	v_9	0,0025	0,0102	0,2450	0,8064
	$v_{10.1}$	-0,0049	0,0124	-0,3968	0,6915
	$v_{10.2}$	-0,0102	0,0125	-0,8210	0,4116
	$v_{11.1}$	0,0132	0,0125	1,0531	0,2923
	$v_{11.2}$	0,0093	0,0124	0,7460	0,4556
$\log(\phi_2)$	Intercepto	4,0696	0,0223	182,7500	***
	Intercepto	-3,3084	0,1868	-17,7140	***
	v_1	4,0898	0,1306	31,3185	***
	v_2	3,0025	0,1295	23,1954	***
	v_3	-0,1593	0,1212	-1,3143	0,1887
	$v_{4.1}$	-0,1138	0,0857	-1,3283	0,1841
	$v_{4.2}$	-0,0648	0,0861	-0,7524	0,4518
	$v_{5.1}$	0,0931	0,0863	1,0784	0,2808
	$v_{5.2}$	0,0945	0,0858	1,1017	0,2706
$g_3(\pi_i)$	$v_{6.1}$	0,0117	0,0868	0,1350	0,8926
	$v_{6.2}$	-0,0681	0,0866	-0,7869	0,4313
	v_7	-0,2768	0,4052	-0,6830	0,4946
	v_8	0,0000	-	-	-
	v_9	0,0225	0,0703	0,3205	0,7486
	$v_{10.1}$	-0,0929	0,0858	-1,0820	0,2793
	$v_{10.2}$	0,0281	0,0858	0,3275	0,7433
	$v_{11.1}$	-0,1650	0,0860	-1,9181	0,0551
	$v_{11.2}$	-0,1630	0,0858	-1,9000	0,0574

No ajuste do *support vector regression machines* utilizamos a função `svm()` do pacote `e1071`, neste ajuste unimos as bases de dados de desenvolvimento e validação, uma vez que a função utilizada no software R já realiza a seleção automática do modelo, consideramos o kernel gaussiano.

Com a amostra de teste comparamos as predições dos diferentes modelos selecionados. Utilizamos duas medidas de desempenho: erro quadrático médio (EQM) e desvio absoluto médio (DAM), resultados da comparação estão na Tabela 25. Para as duas medidas, valores menores

Tabela 24 – Estimativas dos parâmetros do modelo de regressão beta bimodal inflacionado em zero reduzido.

Preditor	Variáveis relacionadas	Estimativa	Erro padrão	Valor z	$P(>\|z\|)$
	Intercepto	-4,0963	0,0875	-46,7942	***
$g_0(\alpha_i)$	v_1	3,1492	0,0792	39,7681	***
	v_2	1,9991	0,0673	29,7105	***
	Intercepto	-1,2872	0,0265	-48,6030	***
	v_3	-1,0142	0,0255	-39,6990	***
	$v_{4.1}$	-0,8365	0,0184	-45,4670	***
	$v_{4.2}$	-0,3760	0,0170	-22,0710	***
$g_1(\mu_{1i})$	$v_{5.1}$	0,9018	0,0185	48,7400	***
	$v_{5.2}$	0,4409	0,0190	23,1680	***
	$v_{6.1}$	0,5379	0,0183	29,3710	***
	$v_{6.2}$	0,4614	0,0185	24,9910	***
	v_7	1,8679	0,0845	22,1020	***
$\log(\phi_1)$	Intercepto	3,8988	0,0313	124,6500	***
	Intercepto	0,9774	0,0180	54,1590	***
	v_3	-1,1701	0,0179	-65,3230	***
	$v_{4.1}$	-1,0764	0,0128	-84,0820	***
	$v_{4.2}$	-0,5340	0,0133	-40,0430	***
$g_2(\mu_{2i})$	$v_{5.1}$	0,9987	0,0127	78,8400	***
	$v_{5.2}$	0,5228	0,0120	43,4730	***
	$v_{6.1}$	0,9026	0,0128	70,7300	***
	$v_{6.2}$	0,4302	0,0120	35,8630	***
	v_7	2,4281	0,0587	41,3930	***
$\log(\phi_2)$	Intercepto	4,0686	0,0231	175,8000	***
	Intercepto	-3,6206	0,1240	-29,2000	***
$g_3(\pi_i)$	v_1	4,0740	0,1303	31,2660	***
	v_2	2,9993	0,1293	23,2000	***

geralmente indicam o modelo com melhor ajuste.

Apresentamos as medidas EQM e DAM a seguir

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}\left(\text{LGD}_i - \hat{\text{LGD}}_i\right)^2,$$

$$\text{DAM} = \frac{1}{n}\sum_{i=1}^{n}\mid \text{LGD}_i - \hat{\text{LGD}}_i \mid,$$

em que n é o número de observações, $\hat{\text{LGD}}_i$ é a i-ésima perda predita e LGD_i é a i-ésima perda observada.

Conforme mostra a Tabela 25, quando consideramos como medida de acurácia o erro quadrático médio o modelo beta bimodal inflacionado em zero apresentou o melhor resultado na amostra de teste e dentre todos os modelos o SVR apresentou o pior resultado. Entretanto, quando consideramos o desvio absoluto médio como medida de acurácia o modelo SVR apresentou o melhor resultado, seguido pelo modelo de regressão beta bimodal inflacionado em zero, a

Tabela 25 – Resultados do EQM e DAM na amostra de teste para os modelos de regressão beta, beta inflacionado em zero, beta bimodal inflacionado em zero e SVR.

	EQM	DAM
Beta	0,0603	0,1960
Beta inflacionado em zero	0,0534	0,1822
Beta bimodal inflacionado em zero	0,0533	0,1773
Support vector regression machines	0,0631	0,1759

melhora no desempenho do modelo SVR na medida DAM é explicada pela função de perda utilizada na construção do SVR, veja a Equação 2.7. O modelo de regressão beta apresentou o pior desempenho neste estudo, já a regressão beta inflacionada em zero apresentou relativa proximidade com o desempenho do modelo de regressão beta bimodal inflacionado em zero.

Ainda que os dados tenham sido gerados de uma distribuição beta bimodal inflacionada em zero, conferindo assim certa vantagem ao modelo de regressão BBZ, consideramos que este modelo de regressão apresentou desempenho satisfatório neste estudo, evidenciando-o como alternativa de modelagem para estimação da LGD que possua inflação em zero e bimodalidade.

CAPÍTULO

5

CONSIDERAÇÕES FINAIS

Neste trabalho propomos estimar a LGD através do modelo de regressão beta bimodal inflacionado em zero com objetivo de obter acurácia semelhante às obtidas com abordagens que tem como finalidade principal a predição, e possua interpretação tão viável quanto um modelo de regressão linear. Apesar do modelo de regressão beta bimodal inflacionado em zero, aqui desenvolvido, ser motivado por dados de LGD, este modelo possui utilidade em todo problema de regressão em que seja plausível a distribuição da variável resposta ser beta bimodal inflacionada em zero.

O trabalho contextualizou o problema de estimação da LGD, apresentou, de forma breve, a regressão beta, regressão beta inflacionado em zero e *support vector regression machines*. Desenvolvemos a distribuição beta bimodal inflacionada em zero e o modelo de regressão para esta distribuição, comparamos três diferentes estratégias de inicialização do algoritmo EM quanto a eficiência de tais estratégias no atingimento da solução ideal para a maximização da função log-verossimilhança, realizamos estudo de simulação avaliando os estimadores de máxima verossimilhança da distribuição beta bimodal inflacionada em zero, bem como para os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão desta distribuição. Abordamos a construção de intervalos de confiança assintóticos, testes de hipóteses e seleção de modelos para os modelos beta bimodal inflacionados em zero. Comparamos o desempenho da regressão beta, regressão beta inflacionada em zero, *support vector regression machines* e a regressão beta bimodal inflacionada em zero, comparação esta realizada com uma base dados de LGD simulada, devido à clara dificuldade no acesso a dados de LGD de uma instituição financeira para estudos acadêmicos.

Por fim, recomendamos o modelo de regressão beta bimodal inflacionado em zero como alternativa para modelagem estimação da LGD, bem como qualquer variável resposta, que apresente inflação de zeros e bimodalidade.

REFERÊNCIAS

ALTMAN, E. I.; KALOTAY, E. A. Ultimate recovery mixtures. **Journal of Banking & Finance**, Elsevier, v. 40, p. 116–129, 2014. Citado nas páginas 29, 30 e 31.

ANDERSON, R. **The credit scoring toolkit**: theory and practice for retail credit risk management and decision automation. New York: Oxford University Press, 2007. Citado na página 25.

ANNIBAL, C. A. **Inadimplência do Setor Bancário Brasileiro**: uma avaliação de suas medidas. Brasília, 2009. Citado na página 27.

ARANDA-ORDAZ, F. J. On two families of transformations to additivity for binary response data. **Biometrika**, Oxford University Press, v. 68, n. 2, p. 357–363, 1981. Citado na página 55.

BACEN. **Resolução nº 3.721**. 2009. <http://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=Res&ano=2009&numero=003721>. Acesso em 22 fev. 2017. Citado nas páginas 25 e 26.

____. **Circular nº 3.648**. 2013. <https://www.bcb.gov.br/pre/normativos/busca/normativo.asp?tipo=circ&ano=2013&numero=3648>. Acesso em 22 de fev. 2017. Citado nas páginas 26, 27 e 28.

BAGNATO, L.; PUNZO, A. Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. **Computational Statistics**, v. 28, n. 4, p. 1571–1597, Aug 2013. ISSN 1613-9658. Disponível em: <https://doi.org/10.1007/s00180-012-0367-4>. Citado na página 54.

BANCO DO BRASIL. **Acordo de Basiléia**. 2008. <http://www.bb.com.br/portalbb/page51,136,3696,0,0,1,8.bb?codigoNoticia=7724>. Acesso em 22 de fev. 2017. Citado nas páginas 26 e 27.

BASTOS, J. A. Forecasting bank loans loss-given-default. **Journal of Banking & Finance**, Elsevier, v. 34, p. 2510–2517, 2010. Citado na página 31.

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). **Statistical science**, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. Citado na página 33.

BUARQUE, A. **Minidicionário Aurélio da Língua Portuguesa**. Rio de Janeiro: Nova Fronteira, 1993. Citado na página 27.

CALABRESE, R. Predicting bank loan recovery rates with a mixed continuous-discrete model. **Applied Stochastic Models in Business and Industry**, Wiley Online Library, v. 30, n. 2, p. 99–114, 2014. Citado nas páginas 30 e 35.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2. ed. Pacific Grove: Duxbury, 2002. Citado na página 55.

CASTLE, K. Van de; KEISMAN, D. Recovering your money: Insights into losses from defaults. **Standard & Poor's Credit Week**, June 16, p. 29–34, 1999. Citado na página 29.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v. 20, p. 273–297, 1995. Citado na página 37.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. London: Chapman and Hall, 1974. Citado na página 59.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in r. **Journal of Statistical Software**, v. 34, n. 2, p. 1–24, 2010. ISSN 1548-7660. Disponível em: <https://www.jstatsoft.org/v034/i02>. Citado na página 34.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the royal statistical society**, JSTOR, v. 39, p. 1–38, 1977. Citado na página 52.

DRUCKER, H.; BURGES, C. J. C.; KAUFMAN, L.; SMOLA, A.; VAPNIK, V. Support vector regression machines. In: **Proceedings of the 9th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 1996. (NIPS'96), p. 155–161. Citado na página 37.

FERNANDES, G.; TOMAZELLA, P. Loss given default: Definições e métodos de modelagem. **Tecnologia de Crédito**, v. 83, p. 6–19, 2013. Citado nas páginas 28 e 29.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado na página 34.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. 2. ed. New York: Springer, 2001. (Series in statistics). Citado nas páginas 38, 39, 51 e 54.

FRÜHWIRTH-SCHNATTER, S. **Finite mixture and Markov switching models**. New York: Springer, 2006. (Series in Statistics). Citado nas páginas 42, 43, 57, 58 e 59.

FRYE, J. Depressing recoveries. **Risk Magazine**, v. 13, n. 11, p. 108–111, 2000. Citado na página 29.

GRUN, B.; LEISCH, F. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. **Journal of Statistical Software**, v. 28, 2008. Citado nas páginas 53, 54 e 63.

GUPTON, G. M.; STEIN, R. M.; SALAAM, A.; BREN, D. LOSSCALC^{TM}: Model for predicting loss given default (LGD). **Moody's KMV**, New York, 2002. Citado na página 34.

HUANG, G. Model identifiability. 2005. In: EVERITT, B. S.; HOWELL, D. C. (Ed.). **Encyclopedia of Statistics in Behavioral Science**. Chichester, John Wiley & Sons, 2005. v. 3, p. 1249–1251. ISBN 0-470-86080-4. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa399>. Citado na página 57.

MARTINEZ, R. O. **Modelos de regressão beta inflacionados**. Tese (Doutorado) — Universidade de São Paulo, São Paulo, 2008. Citado nas páginas 30, 35, 36, 42, 43 e 46.

MCLACHLAN, G.; PEEL, D. **Finite mixture models**. New York: John Wiley & Sons, 2000. Citado nas páginas 54, 57, 58, 60 e 61.

QI, M.; YANG, X. Loss given default of high loan-to-value residential mortgages. **Journal of Banking & Finance**, Elsevier, v. 33, p. 788–799, 2009. Citado na página 29.

QI, M.; ZHAO, X. Comparison of modeling methods for loss given default. **Journal of Banking & Finance**, Elsevier, v. 35, p. 2842–2855, 2011. Citado na página 31.

SCHUERMANN, T. What do we know about loss given default? **SSRN Electronic Journal**, Elsevier BV, 2004. Disponível em: <https://doi.org/10.2139/ssrn.525702>. Citado nas páginas 27, 28 e 29.

SERVIGNY, A. de; RENAULT, O. **The Standard & Poor's Guide to Measuring and Managing Credit Risk**. McGraw-Hill Education, 2004. ISBN 0071417559. Disponível em: <https://www.amazon.com/Standard-Poors-Measuring-Managing-Credit/dp/0071417559?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0071417559>. Citado na página 30.

SILVA, A. C. M. da; MARINS, J. T. M.; NEVES, M. B. E. das. **Loss Given Default: um estudo sobre perdas em operações prefixadas no mercado brasileiro**. Brasília, 2009. Citado nas páginas 28 e 29.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and computing**, v. 14, n. 3, p. 199–222, 2004. ISSN 1573-1375. Disponível em: <https://doi.org/10.1023/B:STCO.0000035301.49549.88>. Citado nas páginas 15 e 38.

SUYKENS, J. A.; VANDEWALLE, J. Least squares support vector machine classifiers. **Neural processing letters**, Springer, v. 9, n. 3, p. 293–300, 1999. Disponível em: <https://doi.org/10.1023/a:1018628609742>. Citado na página 37.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. **Credit scoring and its applications**. Philadelphia: Society for Industrial and Applied Mathematics, 2002. Citado na página 27.

WONG, S.; LAI, T. L. Statistical models for the basel II internal ratings-based approach to measuring credit risk of retail products. **Statistics and Its Interface**, v. 1, p. 229–241, 2008. Citado na página 34.

YAO, X.; CROOK, J.; ANDREEVA, G. Support vector regression for loss given default modelling. **European Journal of Operational Research**, Elsevier BV, v. 240, n. 2, p. 528–538, jan 2015. Disponível em: <https://doi.org/10.1016/j.ejor.2014.06.043>. Citado nas páginas 29, 30, 31 e 37.

ZOHOVA, I. **Application of Scoring Approach in the LGD Estimation**. 2015. <https://www.business-school.ed.ac.uk/crc/wp-content/uploads/sites/55/2017/02/Application-of-Scoring-Approach-in-the-LGD-Estimation-Ivana-Zohova.pdf>. Acesso em 10 de abr. 2018. Citado na página 28.

APÊNDICE

A

A.1 Mistura de duas distribuições beta com médias iguais

Considere uma variável aleatória W proveniente de uma mistura de duas distribuições beta com médias iguais, a distribuição desta variável aleatória possui função densidade da forma

$$db_W(w;\pi,\mu,\phi_1,\phi_2) = \pi f_W(w;\mu,\phi_1) + (1-\pi) f_W(w;\mu,\phi_2),$$

em que $0 < w, \pi < 1$, $f_W(w;\mu,\phi_1)$ e $f_W(w;\mu,\phi_2)$ são funções densidade da distribuição beta da forma (2.2) referentes às duas subpopulações misturadas aleatoriamente com proporções π e $(1-\pi)$, respectivamente.

A esperança e a variância de W são dadas, respectivamente, por

$$\begin{aligned} E[W] &= \mu, \\ Var[W] &= \mu(1-\mu)\left(\frac{\pi}{\phi_1+1} + \frac{1-\pi}{\phi_2+1}\right). \end{aligned}$$

Na Figura 8 apresentamos quatro gráficos da densidade beta com parâmetros $\mu = 0,5$ e $\phi = 12$, representados pela linha cheia, comparando com funções densidades da distribuição de mistura de duas beta com médias iguais de parâmetros $\mu = 0,5$, $\phi_1 = 12$, $\phi_2 = 2$, e π com diferentes valores: $0,2$ (Gráfico 1), $0,5$ (Gráfico 2), $0,8$ (Gráfico 3) e $0,95$ (Gráfico 4), representadas pela linha tracejada.

Já na Figura 9 apresentamos quatro gráficos da densidade beta com parâmetros $\mu = 0,5$ e $\phi = 12$, representados pela linha cheia, comparando com funções densidades da distribuição de mistura de duas beta com médias iguais de parâmetros $\mu = 0,5$, $\phi_1 = 12$, $\pi = 0.5$, e ϕ_2 com diferentes valores: $1,8$ (Gráfico 1), 4 (Gráfico 2), 20 (Gráfico 3) e 100 (Gráfico 4), representadas pela linha tracejada.

Os parâmetros π e ϕ_2 permitem manejo da forma da densidade nos extremos, quando aumentamos π o peso nos extremos é reduzido, e ϕ_2 apresentou maior concentração nos extremos quando $\phi_2 < \phi_1$, mas quando $\phi_2 > \phi_1$ os pesos nos extremos apresentaram menor concentração em relação a distribuição beta com parâmetros $\mu = 0,5$ e $\phi = 12$.

A distribuição de mistura de duas distribuições beta é uma alternativa para ajustar dados em um intervalo fechado em que os extremos não são acomodados pela distribuição beta, devido a flexibilidade de forma e peso nos extremos que a mistura de duas distribuições beta com médias iguais proporciona, como visto na Figura 8 e na Figura 9.

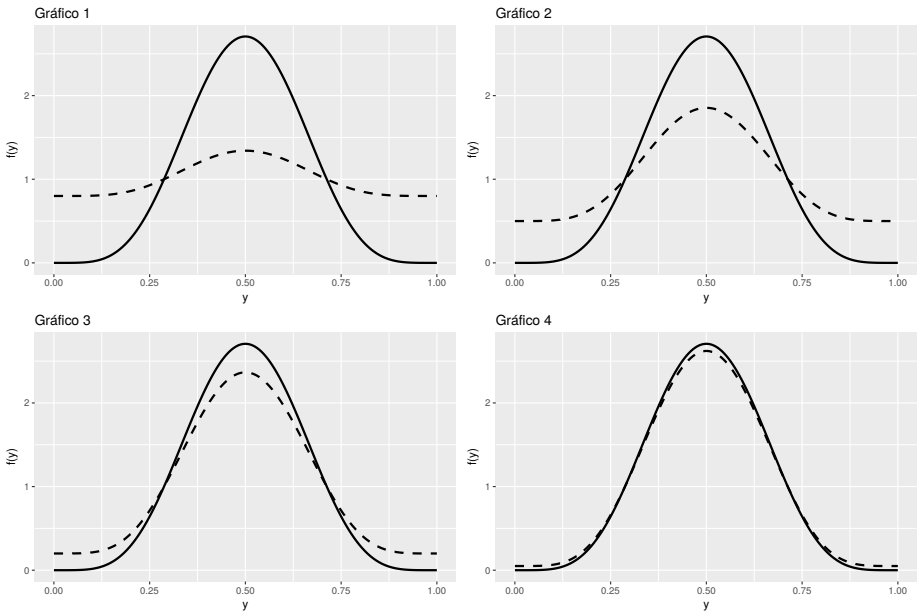

Figura 8 – Gráficos da função de densidade beta e mistura de duas distribuições beta com médias iguais variando o componente peso.

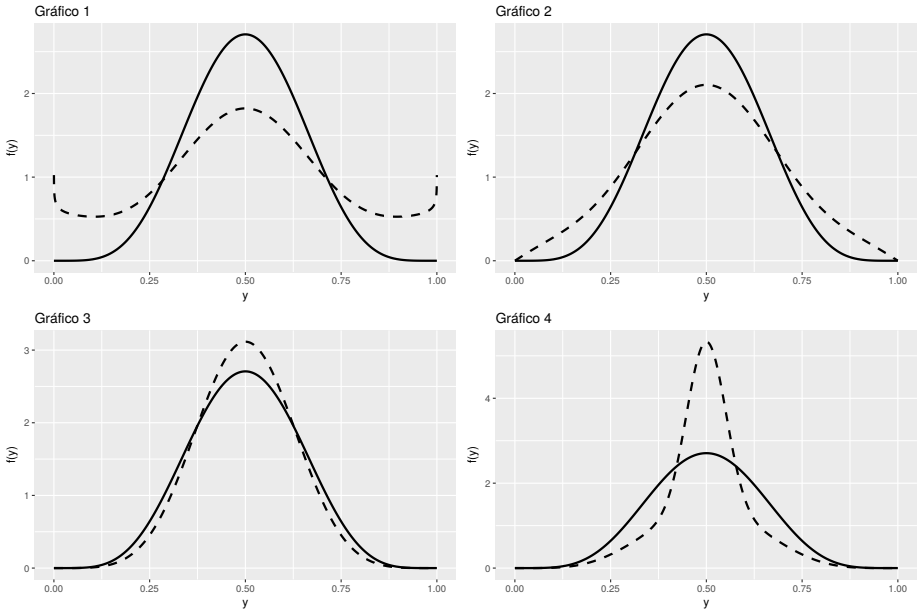

Figura 9 – Gráficos da função de densidade beta e mistura de duas distribuições beta com médias iguais variando o parâmetro de precisão de uma das distribuições da mistura.

A.2 Códigos em R

A.2.1 Função densidade BBZ

Código-fonte 1 – Função densidade da distribuição beta bimodal inflacionada em zero

```
dbbz <- function (x, alpha, Pi,  mu_1, phi_1, mu_2, phi_2)
{
  if (any(alpha <= 0) | any(alpha >= 1))
    stop(paste("alpha must be between 0 and 1", "\n", ""))
  if (any(Pi <= 0) | any(Pi >= 1))
    stop(paste("Pi must be between 0 and 1", "\n", ""))
  if (any(mu_1 <= 0) | any(mu_1 >= 1))
    stop(paste("mu_1 must be between 0 and 1", "\n", ""))
  if (any(mu_2 <= 0) | any(mu_2 >= 1))
    stop(paste("mu_2 must be between 0 and 1", "\n", ""))
  if (any(phi_1 <= 0))
    stop(paste("phi_1 must be greated than 0", "\n", ""))
  if (any(phi_2 <= 0))
    stop(paste("phi_2 must be greated than 0", "\n", ""))
  if (any(x < 0) | any(x >= 1))
    stop(paste("x must be 0<=x<1, i.e. [0 to 1)", "\n", ""))
  a1 <- mu_1 * phi_1
  b1 <- (1 - mu_1) * phi_1
  a2 <- mu_2 * phi_2
  b2 <- (1 - mu_2) * phi_2
  d <- rep(0, length(x))
  d <- ifelse((x > 0), (1-alpha)*(Pi*dbeta(x, shape1 = a1,
   shape2 = b1) + (1-Pi)*dbeta(x, shape1 = a2, shape2 = b2)),
   0)
  d <- ifelse((x == 0), alpha, d)
  d
}
```

A.2.2 Função distribuição acumulada BBZ

Código-fonte 2 – Função distribuição acumulada da distribuição beta bimodal inflacionada em zero

```
pbbz <- function (q, alpha, Pi,  mu_1, phi_1, mu_2, phi_2)
{
  if (any(alpha <= 0) | any(alpha >= 1))
```

```
 4:     stop(paste("alpha must be between 0 and 1", "\n", ""))
 5:   if (any(Pi <= 0) | any(Pi >= 1))
 6:     stop(paste("Pi must be between 0 and 1", "\n", ""))
 7:   if (any(mu_1 <= 0) | any(mu_1 >= 1))
 8:     stop(paste("mu_1 must be between 0 and 1", "\n", ""))
 9:   if (any(mu_2 <= 0) | any(mu_2 >= 1))
10:     stop(paste("mu_2 must be between 0 and 1", "\n", ""))
11:   if (any(phi_1 <= 0))
12:     stop(paste("phi_1 must be greated than 0", "\n", ""))
13:   if (any(phi_2 <= 0))
14:     stop(paste("phi_2 must be greated than 0", "\n", ""))
15:   if (any(q < 0) | any(q >= 1))
16:     stop(paste("y must be 0<=y<1, i.e. [0 to 1)", "\n", ""))
17:   a1 <- mu_1 * phi_1
18:   b1 <- (1 - mu_1) * phi_1
19:   a2 <- mu_2 * phi_2
20:   b2 <- (1 - mu_2) * phi_2
21:   p <- ifelse((q > 0), alpha + (1-alpha)*(Pi*pbeta(q, shape1 =
       a1, shape2 = b1) + (1-Pi)*pbeta(q, shape1 = a2, shape2 = b2)
       ),0)
22:   p <- ifelse((q == 0), alpha, p)
23:   p
24: }
```

A.2.3 Função quantil da distribuição BBZ

Código-fonte 3 – Função quantil da distribuição beta bimodal inflacionada em zero

```
 1: qbbz <- function (p, alpha, Pi,  mu_1, phi_1, mu_2, phi_2)
 2: {
 3:   if (any(alpha <= 0) | any(alpha >= 1))
 4:     stop(paste("alpha must be between 0 and 1", "\n", ""))
 5:   if (any(Pi <= 0) | any(Pi >= 1))
 6:     stop(paste("Pi must be between 0 and 1", "\n", ""))
 7:   if (any(mu_1 <= 0) | any(mu_1 >= 1))
 8:     stop(paste("mu_1 must be between 0 and 1", "\n", ""))
 9:   if (any(mu_2 <= 0) | any(mu_2 >= 1))
10:     stop(paste("mu_2 must be between 0 and 1", "\n", ""))
11:   if (any(phi_1 <= 0))
12:     stop(paste("phi_1 must be greated than 0", "\n", ""))
13:   if (any(phi_2 <= 0))
```

```
14:     stop(paste("phi_2 must be greated than 0", "\n", ""))
15:   if (any(p <= 0) | any(p >= 1))
16:     stop(paste("p must be between 0 and 1", "\n", ""))
17:   a1 <- mu_1 * phi_1
18:   b1 <- (1 - mu_1) * phi_1
19:   a2 <- mu_2 * phi_2
20:   b2 <- (1 - mu_2) * phi_2
21:   q <- base::sapply(p,function(p){
22:     f = function(x) { alpha + (1-alpha)*( Pi*pbeta(x, shape1 =
      a1, shape2 = b1) + (1-Pi)*pbeta(x, shape1 = a2, shape2 = b2)
      ) - p}
23:     suppressWarnings(q <- ifelse((p <= (alpha)), 0, ( uniroot(f
      ,c(0.0001,0.9999))$root )))
24:     q
25:   })
26:   q
27: }
```

A.2.4 Funções que geram valores aleatórios da distribuição BBZ

Código-fonte 4 – Função gera valores aleatórios da distribuição beta bimodal inflacionada em zero

```
 1: rbbz <- function (n, alpha, Pi,  mu_1, phi_1, mu_2, phi_2)
 2: {
 3:   if (any(alpha <= 0) | any(alpha >= 1))
 4:     stop(paste("alpha must be between 0 and 1", "\n", ""))
 5:   if (any(Pi <= 0) | any(Pi >= 1))
 6:     stop(paste("Pi must be between 0 and 1", "\n", ""))
 7:   if (any(mu_1 <= 0) | any(mu_1 >= 1))
 8:     stop(paste("mu_1 must be between 0 and 1", "\n", ""))
 9:   if (any(mu_2 <= 0) | any(mu_2 >= 1))
10:     stop(paste("mu_2 must be between 0 and 1", "\n", ""))
11:   if (any(phi_1 <= 0))
12:     stop(paste("phi_1 must be greated than 0", "\n", ""))
13:   if (any(phi_2 <= 0))
14:     stop(paste("phi_2 must be greated than 0", "\n", ""))
15:   if (any(n <= 0))
16:     stop(paste("n must be a positive integer", "\n", ""))
17:   a1 <- mu_1 * phi_1
18:   b1 <- (1 - mu_1) * phi_1
```

```
19:   a2 <- mu_2 * phi_2
20:   b2 <- (1 - mu_2) * phi_2
21:   n <- ceiling(n)
22:   y <- vector(length = n)
23:   for(i in 1:n){
24:     y[i] <- rbinom(1,1,1-alpha)
25:     if( y[i] == 1 ){ y[i] <- rbinom( 1, 1, Pi)
26:     ifelse( y[i] == 1 , y[i] <- rbeta(1, shape1 = a1, shape2 =
      b1), y[i] <- rbeta(1, shape1 = a2, shape2 = b2) )
27:     }
28:   }
29:   y
30: }
```

Código-fonte 5 – Função gera valores aleatórios da distribuição beta bimodal inflacionada em zero e preserva a origem do componente densidade

```
 1: rbbz_p <- function (n, alpha, Pi,  mu_1, phi_1, mu_2, phi_2)
 2: {
 3:   if (any(alpha <= 0) | any(alpha >= 1))
 4:     stop(paste("alpha must be between 0 and 1", "\n", ""))
 5:   if (any(Pi <= 0) | any(Pi >= 1))
 6:     stop(paste("Pi must be between 0 and 1", "\n", ""))
 7:   if (any(mu_1 <= 0) | any(mu_1 >= 1))
 8:     stop(paste("mu_1 must be between 0 and 1", "\n", ""))
 9:   if (any(mu_2 <= 0) | any(mu_2 >= 1))
10:     stop(paste("mu_2 must be between 0 and 1", "\n", ""))
11:   if (any(phi_1 <= 0))
12:     stop(paste("phi_1 must be greated than 0", "\n", ""))
13:   if (any(phi_2 <= 0))
14:     stop(paste("phi_2 must be greated than 0", "\n", ""))
15:   if (any(n <= 0))
16:     stop(paste("n must be a positive integer", "\n", ""))
17:   a1 <- mu_1 * phi_1
18:   b1 <- (1 - mu_1) * phi_1
19:   a2 <- mu_2 * phi_2
20:   b2 <- (1 - mu_2) * phi_2
21:   n <- ceiling(n)
22:   y <- matrix(nrow = n, ncol = 2)
23:   for(i in 1:n){
24:     y[i,] <- c( rbinom(1,1,1-alpha), 0 )
```

```
25:      if( y[i,1] == 1 ){ y[i,1] <- rbinom( 1, 1, Pi)
26:      ifelse( y[i,1] == 1, y[i,] <- c( rbeta(1, shape1 = a1,
       shape2 = b1), 1 ), y[i,] <- c( rbeta(1, shape1 = a2, shape2
       = b2) , 2 ) )
27:      }
28:    }
29:    y
30: }
```

A.2.5 Função para a regressão BBZ

O problema de otimização (3.18) é o mesmo problema encontrado quando temos uma amostra aleatória de distribuição bernoulli com $P(Y^* = 1) = \alpha$ e $P(Y^* = 0) = 1 - \alpha$, em que $y_i^* = \mathbb{1}_{\{0\}}(y_i)$, para $i = 1, ..., n$, assim a função de log verossimilhança é

$$\begin{aligned}
\ell(\boldsymbol{\alpha}; \boldsymbol{y}^*) &= \sum_{i=1}^{n} y_i^* \log(\alpha_i) + \left(n - \sum_{i=1}^{n} y_i^* \right) \log(1 - \alpha_i), \\
&= \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(y_i) \log(\alpha_i) + \left(n - \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(y_i) \right) \log(1 - \alpha_i) \\
&= \ell_1(\boldsymbol{\alpha}; \boldsymbol{y}).
\end{aligned}$$

Desta forma para encontrar as estimativas dos parâmetros relacionados a α utilizamos a função `glm()` do pacote `stats` do software R. Para encontrar as estimativas do problema (3.19), utilizamos a função `betamix()` do pacote `betareg` do software R.

Código-fonte 6 – Função ajusta um modelo de regressão beta bimodal inflacionado em zero.

```
1: regbbz <- function(formula_alpha, formula_2beta, link_alpha="
     logit", link_2beta="logit", cluster = NULL, FLXcontrol =
     list(minprior=0), dados){
2:   dados_alpha = data.frame(y = as.numeric(dados[,1] == 0),
     dados[,-1]  )
3:   mod_alpha <- stats::glm(formula = formula_alpha, family =
     binomial(link = link_alpha), data = dados_alpha)
4:   dados_2beta = subset(dados, dados$y != 0)
5:   mod_2beta <- betareg::betamix(formula = formula_2beta,  k=2,
     cluster = cluster, link = link_2beta, FLXcontrol =
     FLXcontrol, data = dados_2beta)
6:   return (list(mod_alpha,mod_2beta))
7: }
```