Universidade de Brasília

Instituto de Exatas

Departamento de Estatística

Modelagem da LGD Via Mistura de Distribuições *Kumaraswamy*

Thiago Morais de Carvalho

Brasília

2015

Thiago Morais de Carvalho - 13/0000949

Modelagem da LGD Via Mistura de Distribuições *Kumaraswamy*

Dissertação apresentada à Universidade de Brasília como requisito para obtenção do título de mestre em Estatística.

Orientação: *Prof.ª Dra. Cira Etheowalda Guevara Otiniano*

Brasília - DF

2015

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este trabalho teve como motivação inicial a necessidade de apresentar uma maneira alternativa de acompanhar o desempenho de um modelo de risco de crédito para uma instituição financeira (IF) brasileira.

Técnicas de acompanhamento para verificação do desempenho, ao longo do tempo, de modelos de risco de clientes praticadas até hoje, no escopo dos modelos de crédito avaliados e monitorados por uma determinada IF, são os testes Kolmogorov- Smirnov (KS), Índice de Gini e Curva ROC.

Com o advento do Acordo de Basileia, mais especificamente o Basileia II, surgiu a necessidade de criação de modelos referentes aos parâmetros de risco de crédito: probabilidade de descumprimento (PD) que vem de "*probability of default*", exposição no momento do descumprimento (EAD), que vem de "*exposure at default*" e perda dado o descumprimento (LGD) que vem de "*loss given default*".

Os anos de 2013 e 2014, nesta IF, concentraram os esforços para a construção de tais modelos. Entretanto, ao contrário do que é verificado para os modelos de risco de clientes, esses novos modelos não dispõem de um conjunto de técnicas estatísticas definidas para seu monitoramento.

Dessa forma, definiu-se o objetivo deste trabalho, que consiste em modelar os dados de LGD de forma a obtenção de uma distribuição paramétrica, possibilitando assim o conhecimento do comportamento deste tipo de dados e sua avaliação sob o teste KS com a finalidade de subsidiar as decisões de revisão ou remodelagem dos modelos de LGD.

É importante ressaltar que a partir da aplicação do teste KS tem-se apenas indícios de uma possível alteração das características populacionais para a tomada de decisão e não uma conclusão sobre a calibragem ou remodelagem do modelo. Para isto, deve-se alinhar essa avaliação a outros testes.

Neste sentido, [Calabrese, 2014] apresentou um modelo para a LGD em períodos de crise, utilizando a mistura de duas distribuições *Beta*, o que nos motivou a utilizar a mistura de duas distribuições *Kumaraswamy*.

Assim como a distribuição *Beta*, a distribuição *Kumaraswamy* tem suporte no intervalo $(0, 1)$. A vantagem da utilização da distribuição *Kumaraswamy* frente a *Beta* consiste no comportamento igualmente variado que a distribuição pode assumir, mas com forma fechada de sua função de distribuição acumulada, que facilita seu uso em procedimentos computacionais. Além disso, há o fato de que dentro de um ambiente não-acadêmico existe uma certa resistência à técnicas mais complexas ou a conceitos não-auto explicáveis.

Portanto, este trabalho foi realizado com a intenção de fornecer uma técnica alternativa, simples, de monitoramento de modelos de LGD, proporcionando ao leitor entendimento sobre o parâmero em si, a técnica de mistura de distribuições, em particular a mistura de distribuições *Kumaraswamy*. Utilizamos o algoritmo EM para obter as estimativas de máxima verossimilhança dos parâmetros do modelo. O desempenho do algoritmo foi avaliado por meio de simulação de Monte Carlo, e em seguida o modelo foi aplicado em dois conjuntos de dados reais.

Este trabalho está estruturado em 6 capítulos. O capítulo 1 corresponde à introdução. O capítulo 2 trata dos conceitos preliminares que serão utilizados ao longo do estudo, como por exemplo, o conceito de LGD, a apresentação da distribuição de probabilidade utilizada na modelagem via mistura, bem como a técnica de mistura de distribuições, a propriedade de identificabilidade e o teste de Kolmogorov-Smirnov. O capítulo 3 conceitua e desenvolve a modelagem por meio de mistura de distribuições para o modelo de mistura considerado, *Kumaraswamy* com *Kumaraswamy*. O capítulo 4 apresenta o algoritmo EM, sua utilização e os resultados dos estudos de simulação que o validaram. No capítulo 5, o modelo desenvolvido é aplicado em dois conjuntos de dados reais de perda e são verificados

os seus resultados quanto ao ajuste. Finalmente, no capítulo 6 é apresentada a conclusão e as ações para trabalhos futuros.

Capítulo 2

Conceitos Preliminares

Este capítulo trata de alguns conceitos que são necessários ao bom entendimento deste trabalho. Portanto, nas próximas seções, o leitor entrará em contato com as definições e os detalhes do que vem a ser o parâmetro de risco de crédito LGD, a distribuição de probabilidade que será utilizada na modelagem dos dados de perda (*Kumaraswamy*) e a técnica estatística de mistura de distribuições de probabilidade bem como suas propriedades.

2.1 *Loss Given Default* - LGD

A LGD é elemento essencial na abordagem IRB avançada (Basileia II), visto que as instituições financeiras devem calcular seu valor através de modelos internos ao invés de utilizar os valores fixados pelo regulador de mercado [Bennett *et al.* , 2005]. Em termos gerais, a perda dado o descumprimento, em inglês "*Loss Given Default*", corresponde à parcela de uma exposição não percebida pelo credor em virtude do inadimplemento (descumprimento ou *default*) do tomador, isto é, em virtude da incapacidade de pagamento por parte do tomador do empréstimo.

O conceito se torna mais simples quando a ele é associado o conceito de recuperação de crédito. Essa recuperação de crédito é definida como a parcela de uma dívida que é paga ao credor em cumprimento à obrigação de crédito firmada entre este e o tomador do empréstimo, ou seja, corresponde aos valores que retornam ao credor. Em termos práticos, se por algum motivo uma parcela do valor emprestado não retorna ao credor,

tem-se aí uma perda. Quando se observa essa perda com relação ao montante devido no momento em que o devedor está numa situação de inadimplência, isto é, quando este já está em situação de atraso e demonstra incapacidade de pagamento, tem-se aí a LGD.

Portanto, em termos algébricos, temos : LGD = valor da exposição - valor recuperado. Em termos percentuais, corresponde à $1 - tr$, onde tr corresponde à taxa de recuperação dos valores emprestados num determinado produto de crédito.

O Banco Central do Brasil (Bacen), em sua circular nº 3.648, de 4 de março de 2013, conceitua a LGD como o percentual em relação ao parâmetro EAD (valor da exposição no momento do descumprimento) observado, da perda econômica dado o 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.

Segundo [Jacobs Jr & Karagozoglu, 2007], a LGD pode ser definida de diversas formas em função do arcabouço institucional, do contexto de modelagem ou, ainda, conforme o tipo de instrumento de crédito utilizado. No caso de empréstimos bancários, a LGD é definida como o percentual de perdas de uma exposição de risco no momento da inadimplência e, uma vez que tenha ocorrido o evento, a LGD inclui três tipos de perdas: a perda do principal, a perda decorrente dos custos de empréstimos não pagos (inclusive custos de oportunidade) e a perda relacionada às despesas relativas ao processo de cobrança e recuperação do crédito.

O parâmetro LGD é o valor de perda *ex post* expresso como um percentual da EAD para uma exposição de crédito, caso o tomador do empréstimo esteja em *default* (descumprimento, isto é, classificado como insolvente). No caso de um tomador que não esteja em *default*, a LGD é a estimativa *ex ante* da perda também expressa como um percentual da EAD, ou seja, é uma variável aleatória que deve ser estimada através de modelos internos na abordagem IRB avançada [Bennett *et al.* , 2005].

Para o cálculo de LGD das exposições que não estejam em *default*, existem métodos subjetivos, que são baseados na experiência e julgamento por parte dos especialistas diretamente envolvidos no processo, e métodos objetivos, que utilizam dados de perdas realizadas para o desenvolvimento de modelos. Os métodos objetivos podem ser explícitos

ou implícitos [Bennett *et al.* , 2005].

Os métodos explícitos são aqueles em que as informações analisadas em bases de dados permitem o cálculo direto da LGD. São dois os métodos utilizados [Bennett *et al.* , 2005]:

- *Market* LGD, abordagem que se baseia na observação dos preços de mercado de títulos ou empréstimos negociáveis logo após o *default*;
- *Workout* LGD, abordagem baseada no desconto dos fluxos de caixa resultantes do processo de recuperação desde a data do *default* até o final do período de recuperação de crédito.

Já nos métodos implícitos, de acordo com [Bennett *et al.* , 2005], os valores de LGD são derivados de perdas e estimativas de PD, e não podem ser calculados diretamente pelas informações existentes nas bases de dados. Existem dois métodos distintos utilizados:

- *implied market* LGD, abordagem na qual as estimativas de LGD são derivadas dos preços de mercado de títulos arriscados que não estejam em *default* através de um modelo de precificação de ativos. Esta abordagem é útil em portfólios com poucas observações de inadimplência;
- *implied historical* LGD, método utilizado no cálculo de LGD das carteiras do varejo, que consiste na inferência da LGD das perdas realizadas e de estimativas de PD.

Na abordagem IRB avançada, existem duas filosofias de LGD que os bancos podem adotar. Na filosofia conhecida como PIT (*point-in-time*), que é normalmente utilizada pelos bancos, a LGD é uma medida cíclica que reflete a LGD esperada, normalmente, nos próximos 12 meses. Por outro lado, na filosofia denominada TTC (*through-the-cycle*), a LGD estimada é uma medida acíclica definida pela média no ciclo econômico e relativamente constante ao longo do ciclo [Miu & Ozdemir, 2006].

2.2 Distribuição de Probabilidade *Kumaraswamy*

Nesta seção é apresentada a distribuição de probabilidade *Kumaraswamy*, utilizada no escopo deste trabalho. É uma distribuição de probabilidade contínua, com suporte

no intervalo $[0, 1]$, bastante conhecida no meio estatístico e em hidrologia, graças ao seu criador [Kumaraswamy, 1980].

Conforme [Jones, 2009], tratamos da distribuição de probabilidade constituída de dois parâmetros, que chamaremos de distribuição *Kumaraswamy*, denotada por *Kumaraswamy*(a, b), onde $a > 0$ e $b > 0$, são parâmetros de forma.

Algumas propriedades relativas à forma da distribuição *Kumaraswamy* são:

- $a > 1$ e $b > 1$ - unimodal;
- $a < 1$ e $b < 1$ - uniantimodal;
- $a > 1$ e $b \leq 1$ - crescente;
- $a \leq 1$ e $b > 1$ - decrescente;
- $a = 1$ e $b = 1$ - constante.

Nos primeiros dois casos, a moda e a antimoda estão no valor

$$x_0 = \left(\frac{a-1}{ab-1}\right)^{1/a}.$$

Conforme pode ser observado na figura 2.1, para várias combinações de valores dos parâmetros a e b, a distribuição *Kumaraswamy* pode assumir diversas formas. Essa é uma característica semelhante à observada na distribuição *Beta*, $B(\alpha, \beta)$, cuja densidade é dada por

$$\begin{aligned} g(x; \alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \end{aligned} \tag{2.1}$$

mas a *Kumaraswamy* tem algumas vantagens no que se refere à tratabilidade matemática. Sua função densidade de probabilidade é dada por

$$f(x; a, b) = abx^{a-1}(1-x^a)^{b-1}, \quad 0 < x < 1 \tag{2.2}$$

Devido à flexibilidade de seus parâmetros, a e b, reais e não negativos, a distribuição *Kumaraswamy* pode assumir diversas formas. A figura 2.1 ilustra essa característica, para alguns pares de valores dos parâmetros.

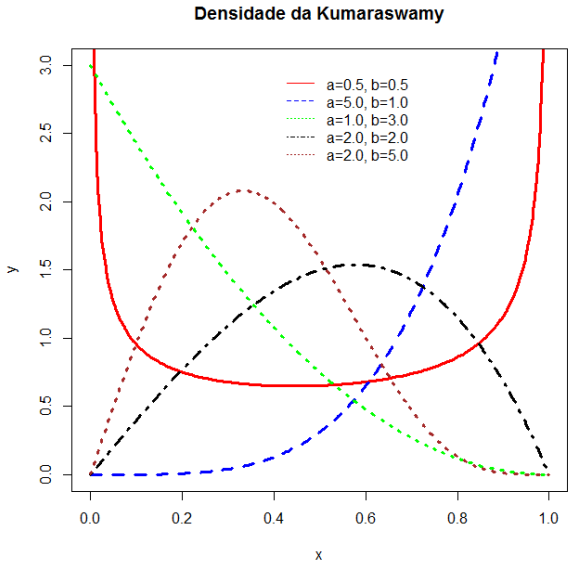


Figura 2.1: Função densidade de probabilidade *Kumaraswamy* para diversas combinações de parâmetros

O comportamento assintótico da densidade *Kumaraswamy* nos limites de seu intervalo é dado por

$$\begin{aligned} f(x) &\sim x^{a-1}, \quad \text{quando} \quad x \to 0 \\ f(x) &\sim (1-x)^{b-1} \quad \text{quando} \quad x \to 1. \end{aligned} \tag{2.3}$$

A função densidade quantílica $q(y) = Q'(y)$ é dada por:

$$\begin{aligned} q(y) &= \frac{(1-y)^{1/b-1}\left[1-(1-y)^{1/b}\right]^{1/a-1}}{ab} \\ &= f(1-y; 1/b, 1/a). \end{aligned} \tag{2.4}$$

Os momentos de ordem r da distribuição *Kumaraswamy* são obtidos imediatamente a partir de

$$\begin{aligned} E(X^r) &= \int_0^1 x^r dF(x) \\ &= \int_0^1 x^r f(x) dx \\ &= bB(1+r/a;b), \quad \forall r > -a. \end{aligned} \tag{2.5}$$

Note que na expressão acima está presente uma função especial, a função *Beta*.

Em particular, temos os dois primeiros momentos, dos quais se obtêm a média e a variância, como segue

$$E(X) = bB\left(1+\frac{1}{a};b\right) \tag{2.6}$$

$$Var(X) = bB\left(1+\frac{2}{a};b\right) - \left[bB\left(1+\frac{1}{a},b\right)\right]^2 \tag{2.7}$$

A expressão da assimetria para a distribuição *Kumaraswamy* é dada por:

$$\begin{aligned} \tau_3 &= \frac{\lambda_3}{\lambda_2} \\ &= \frac{B\left(1+\frac{1}{a};b\right) - 6B\left(1+\frac{1}{a},2b\right) + 6B\left(1+\frac{1}{a};3b\right)}{B\left(1+\frac{1}{a};b\right) - 2B\left(1+\frac{1}{a},2b\right)}, \end{aligned} \tag{2.8}$$

onde

$$\lambda_r = \frac{b}{r}\sum_{l=1}^{r}(-1)^{l-1} l \binom{r}{l}\binom{r+l-2}{r-1} B\left(bl;1+\frac{1}{a}\right). \tag{2.9}$$

Conforme [Jones, 2009], no que se refere à assimetria, quando um valor de a é fixado, à medida que o parâmetro b sofre incremento a assimetria que começa negativa, vai mudando de comportamento até o momento em que se torna positiva. Relação inversa é observada quando b é fixado e há incremento no parâmetro a, ou seja, a assimetria que inicialmente era positiva torna-se negativa num certo nível de a.

Para a curtose, [Jones, 2009] informa a seguinte relação:

$$\tau_4 = \frac{B\left(1+\frac{1}{a};b\right) - 12B\left(1+\frac{1}{a};2b\right) + 30B\left(1+\frac{a}{a};3b\right) - 20B\left(1+\frac{1}{a};4b\right)}{B\left(1+\frac{1}{a};b\right) - 2B\left(1+\frac{1}{a};2b\right)}. \quad (2.10)$$

A expressão da função de distribuição acumulada (fda) para a distribuição *Kumaraswamy* é dada por:

$$F(x;a,b) = 1-(1-x^a)^b; \quad 0 \leq x \leq 1,\ a>0\ e\ b>0. \quad (2.11)$$

A figura 2.2 ilustra o comportamento da fda *Kumaraswamy* para alguns pares de valores dos parâmetros:

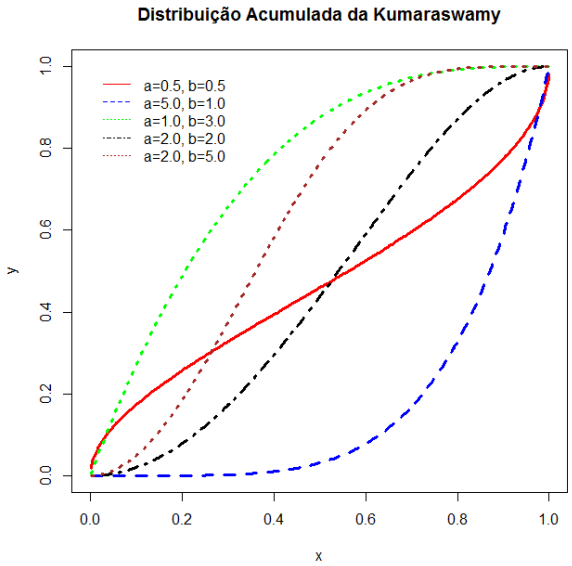


Figura 2.2: Função Distribuição de Probabilidade Acumulada *Kumaraswamy* para diversas combinações de parâmetros

A função de distribuição da *Kumaraswamy* é facilmente invertida a fim de se obter a função quantílica

$$\begin{aligned} Q(y) &= F^{-1}(y) \\ &= \left[1-(1-y)^{1/b}\right]^{1/a}, \quad 0<y<1. \end{aligned} \quad (2.12)$$

2.3 Mistura de Distribuições

Segundo [McLachlan & Peel, 2000], a história dos modelos de mistura finita remonta há mais de um século com Pearson, em 1894, cujos modelos de misturas estavam baseados em uma mistura de duas componentes normais univariadas.

Os modelos de mistura de distribuições são métodos flexíveis de modelar funções de distribuições de probabilidade complexas a partir de elementos mais simples e tratáveis [Horta, 2009]. Por causa dessa flexibilidade, estes modelos estão sendo cada vez mais explorados como uma maneira conveniente de modelar formas desconhecidas de distribuições [McLachlan & Peel, 2000].

Uma situação útil para entender o contexto em que se insere a técnica de mistura de distribuições pode ser verificada quando um conjunto de dados amostrais, proveniente de uma única população, apresenta o fenômeno da superdispersão, ou seja, grande disparidade entre a variância teórica (calculada a partir do modelo) e a variância amostral (calculada a partir da sequência amostral). Pressupor a existência de subpopulações dentro de uma população geral, pode minimizar tal disparidade e é aí que reside uma das principais motivações para o uso dos modelos de mistura de distribuições.

Quanto ao uso, nas últimas décadas, verificou-se um aumento considerável no potencial de aplicação dos modelos de mistura de distribuições em várias áreas do conhecimento, como por exemplo, em astronomia, biologia, genética, medicina, psiquiatria, economia, engenharia, finanças, dentre outras, especialmente quando na modelagem de populações heterogêneas

Grande parte dos conceitos, definições e demonstrações de mistura de distribuições apresentados neste trabalho, têm origem nos estudos de [McLachlan & Peel, 2000].

Definição

Um modelo de misturas finitas é uma combinação linear convexa de funções de distribuição acumuladas.

Mais precisamente, se $\mathcal{F}$ é uma família de funções de distribuição acumuladas tal que $F_1, \ldots, F_n \in \mathcal{F}$, então

$$H(x;\Theta) \;=\; \sum_{i=1}^{n} p_i F_i(x;\theta_i) \tag{2.13}$$

é a mistura finita de n componentes $F_1, \ldots, F_n$, com pesos $p_1 > 0, \ldots, p_n > 0$, tal que $\sum_{i=1}^{n} p_i = 1$, e $\Theta = (\theta_1, \cdots, \theta_n)$ é um vetor dos parâmetros θ_i da componente F_i.

Neste trabalho, o conjunto

$$\mathcal{H} \;=\; \left\{ H : H(x;\Theta) = \sum_{i=1}^{n} p_i F_i(x;\theta_i); F_i(x;\theta_i) \in \mathcal{F} \right\} \tag{2.14}$$

denota a classe das misturas finitas da família $\mathcal{F}$, onde o vetor Θ contém todos os parâmetros desconhecidos do modelo de mistura, ou seja,

$$\Theta \;=\; (p_1, \ldots, p_{k-1}, \theta_1, \ldots, \theta_k). \tag{2.15}$$

Neste caso, as componentes F_i pertencem à mesma família $\mathcal{F}$, mas em geral $\mathcal{F}$ pode ser a união de diferentes famílias de funções de distribuição.

Teoria e aplicações de misturas de distribuições podem ser encontradas nos trabalhos de [McLachlan & Peel, 2000].

Problemas

Segundo [Horta, 2009], existem algumas dificuldades na utilização dos modelos de mistura que devem ser levadas em consideração no momento de construção do modelo:

- a escolha das distribuições $f_1(\cdot), \ldots, f_k(\cdot)$ a serem empregadas para compor a distribuição que descreve a mistura;
- a procura de técnicas de estimação dos parâmetros da distribuição da mistura;
- a determinação prática do tamanho de amostras para a estimação dos parâmetros;
- a decisão de utilizar um k conhecido ou não, e caso se opte por incluí-lo no conjunto paramétrico desconhecido, escolher a técnica mais adequada para estimá-lo.

2.4 Propriedade da Identificabilidade

Identificabilidade é uma propriedade relativa ao modelo e tem a missão de verificar se a partir de conjuntos de parâmetros distintos, é possível observar resultados também distintos para o modelo. Caso se verifique, diz-se que o modelo é identificável.

De forma a clarificar essa ideia, pode-se imaginar a seguinte situação: seja θ um conjunto de parâmetros de um modelo X, que por sua vez, pode ser expressado por meio de sua função densidade de probabilidade, $f(x, \theta)$. Se para outro conjunto de parâmetros, θ^*, relativo ao mesmo modelo, for obtido resultado diferente de quando da aplicação de θ, tem-se que o modelo em questão é identificável. Em termos matemáticos, pode-se escrever que se para $\theta \neq \theta^*$ é verificado que $f(x, \theta) \neq f(x, \theta^*)$, então o modelo é identificável.

Neste capítulo será abordada, por meio de conceitos, definições, exemplos e demonstrações, a propriedade de identificabilidade. Num primeiro momento, de forma geral, sob a ótica de [McLachlan & Peel, 2000]; e de forma mais detalhada, sob o enfoque de [Atienza *et al.* , 2006].

De acordo com [McLachlan & Peel, 2000], a estimação de Θ, conjunto de parâmetros, com base nas observações x_i, somente é significativa se Θ é identificável. Em geral, uma família paramétrica de densidades $f(x_i; \Theta)$ é identificável se distintos valores do parâmetro Θ determinam distintos membros da família de densidades $\{F(x_i; \Theta) : \Theta \in \Omega\}$, onde Ω é um espaço paramétrico específico, tal que:

$$F(x_i; \Theta) \quad = \quad F(x_i; \Theta^*) \tag{2.16}$$

se e somente se, $\Theta = \Theta^*$.

Uma abordagem mais recente de identificabilidade foi apresentada por [Atienza *et al.* , 2006] onde é apresentada uma condição suficiente para a identificabilidade de misturas finitas de distribuições. Tal condição é menos restritiva que as condições desenvolvidas por [Teicher, 1963] e pode ser aplicada a um conjunto mais amplo de famílias de mistura. Essa caracterização é apropriada também, para os casos de misturas com diferentes famílias de distribuição.

Para o processo de estimação ser bem definido é necessário que H (conforme expressão 2.14) seja identificável, isto é, deve existir uma única caracterização para a classe dos modelos considerados. Dificuldades particulares na aplicação dessa propriedade aparecem na classe de modelos de misturas.

A classe $\mathcal{F}$ de funções distribuição, a classe de misturas finitas de $\mathcal{F}$ e $\mathcal{H}$, são indentificáveis se, e somente se, para todo $H, \hat{H} \in \mathcal{H}$,

$$\begin{aligned} H &= \sum_{j=1}^{k} p_j F_j \\ \hat{H} &= \sum_{j=1}^{\hat{k}} \hat{p}_j \hat{F}_j, \end{aligned}$$

a igualdade $H = \hat{H}$ implica em $k = \hat{k}$ e que $(p_1F_1), \ldots, (p_kF_k)$'s são uma permutação de $(\hat{p}_1\hat{F}_1), \ldots, (\hat{p}_k\hat{F}_k)$.

Problemas de identificabilidade relativos à misturas finitas tem sido bastante investigados. Vários autores contribuíram nessa área como, por exemplo, [Teicher, 1963] que criou uma condição suficiente para que uma mistura das famílias *Normal* e *Gama* seja identificável; ou [Ahmad, 1988] que provou a identificabilidade para misturas de densidades *Weibull*, *Lognormal*, *Chi* e *Pareto* pela modificação do teorema de [Teicher, 1963], dado por [Chandra, 1977], para a função geradora de momentos de $logX$; ou [Barndorff-Nielsen, 1965] que criou uma condição suficiente para misturas da família exponencial.

O problema da identificabilidade em modelos de misturas tem sido o foco de interesse nas últimas décadas. O trabalho de [Teicher, 1963] é o ponto inicial para a investigação da identificabilidade em modelos de mistura. Entretanto, essa identificabilidade é estabelecida sob algumas condições restritivas que não são aplicáveis a algumas famílias multiparamétricas. Em seu trabalho, [Atienza *et al.* , 2006] propõem um novo resultado que relaxa os pre-requisitos do teorema de [Teicher, 1963], que dizia:

Teorema 1 *Seja $\mathcal{F}$ uma família de funções de distribuição univariadas com transformações $\phi(t)$, tal que t pertence a um domínio de definição $S(\phi)$ e a aplicação $F \to \phi$ é linear. Suponha que exista uma ordem total em $\mathcal{F}$, denotada por $\prec$, de modo que $F_1 \prec F_2$ implica em:*

1. *os domínios* $S(\phi_1)$ *e* $S(\phi_2)$ *são tais que:*

$$S(\phi_1) \subset S(\phi_2), \tag{2.17}$$

2. *Existe* $t_1 \in \bar{S}(\phi_1)$ *(onde* $\bar{S}(\phi_1)$ *é o complemento de* $S(\phi_1)$*) com* t_1 *independente de* $S(\phi_2)$ *tal que*

$$\lim_{t \to t(\phi_1)} \frac{\phi_{\phi_2}(t)}{\phi_{\phi_1}(t)} \quad = \quad 0. \tag{2.18}$$

O ponto principal do trabalho de [Atienza *et al.* , 2006] gira em torno de uma nova versão do teorema de [Teicher, 1963], com hipóteses mais fracas, que permitiram o estudo de problemas de identificabilidade num contexto mais amplo. E isso é bastante útil no estudo de mistura de diferentes famílias de distribuição, com uso amplo em áreas como a biologia, medicina e fenômenos sociais.

A seguir apresentamos os principais resultados dados por [Atienza *et al.* , 2006]. Para isto, considere $A^{'}$ o conjunto de pontos de acumulação de $A \subset R^d$, que consiste em todos os pontos para os quais cada ponto vizinho contém um número infinito de pontos distintos de A.

Teorema 2 *Seja* $\mathcal{F}$ *uma família de distribuições. Seja* M *uma função linear que transforma qualquer* $F \in \mathcal{F}$ *numa função real* ϕ_F *com domínio* $S(F) \subset R^d$*. Seja* $S_0(F) = \{t \in S(F) : \phi_F(t) \neq 0\}$*. Suponha que exista uma ordem total* $\prec$ *em* $\mathcal{F}$*, tal que para todo* $F \in \mathcal{F}$ *existe* $t(F) \in S_0(F)'$*, verificando:*

1. *Se* $F_1, F_2, \ldots, F_n \in \mathcal{F}$ *com* $F_1 \prec F_i$ *para* $2 \leq i \leq n$*, então*

$$t(F_1) \in \left[S_0(F_1) \cap \left[\cap_{i=2}^{n} S(F_i)\right]\right]'.$$

2. *Se* $F_1 \prec F_2$*, então* $\lim_{t \to t(F_1)} \frac{\phi_{F_2}(t)}{\phi_{F_1}(t)} = 0$*.*

Então, a classe $\mathcal{H}$ *de todas as misturas de distribuição finitas de* $\mathcal{F}$ *é identificável.*

De posse deste resultado, deve-se considerar os seguintes pontos:

1. Essa nova condição inclui distribuições multivariadas, ao passo que os resultados de [Teicher, 1963] somente se aplicam a distribuições univariadas;

2. Aplicação do resultado de [Teicher, 1963] pode ser restritiva por causa da condição $S(F_1) \subset S(F_2)$ que é imposta à relação $F_1 \prec F_2$. Essa condição é relaxada no item (1) do teorema anterior.

3. O novo teorema proposto neste estudo pode ser aplicado a um amplo conjunto de famílias de distribuição.

Finalmente, como consequência, tem-se o corolário abaixo, que apresenta uma simplificação das hipóteses nos casos onde o ponto $t(F) = t_0$ não depende de F em $\mathcal{F}$.

Corolário 1. *Seja $\mathcal{F}$ uma família de distribuições. Seja M uma função linear que transforma qualquer $F \in \mathcal{F}$ numa função real ϕ_F com domínio $S(F) \subseteq R^d$. Seja $S_0(F) = \{t \in S(F) : \phi_F(t) \neq 0\}$ e suponha que existe um ponto t_0 tal que:*

$$t_0 \in \left[\bigcap_{1 \leq i \leq k} S_0(F_i) \right]'$$

para qualquer coleção finita de distribuições $F_1, \ldots, F_k \in \mathcal{F}$. Se a ordem

$$F_1 \prec F_2, \text{ se e somente se } \lim_{t \to t_0} \frac{\phi_{F_2}(t)}{\phi_{F_1}(t)} = 0$$

é um ordenamento total em $\mathcal{F}$, então a classe $\mathcal{H}$ de todas as misturas finitas de distribuição de $\mathcal{F}$ é identificável.

Este corolário simplifica consideravelmente a verificação da identificabilidade para algumas classes, em particular, quando $S_0(F)$ é da forma $(a(F), +\infty)$, $(-\infty, b(F))$ ou $(-\infty, +\infty)$. Nesses casos, podemos considerar t_0 como sendo $+\infty$ ou $-\infty$, respectivamente.

2.5 Monitoramento dos Modelos

A primeira análise de estabilidade populacional deve ser visual, comparando graficamente as distribuições de frequências de referência com as distribuições que serão monitoradas. Deve-se avaliar se as diferenças são significativas. Essa avaliação preliminar deverá ser ratificada com base em testes estatísticos.

Quando a análise visual sugere alterações nas distribuições de frequências, deve-se proceder a testes estatísticos para avaliar a distância entre as distribuições e concluir se diferem significantemente. Trataremos especificamente do teste KS descrito a seguir.

2.6 Teste Kolmogorov-Smirnov - KS

O teste KS será empregado neste trabalho com a finalidade de verificar se os dados simulados ou reais pertencem a uma dada distribuição de probabilidade. Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, e a função de distribuição empírica dos dados. As hipóteses consideradas são as seguintes:

$$\begin{cases} H_0 : \text{Os dados provêm da mesma distribuição de probabilidade} \\ H_1 : \text{Os dados provêm de diferentes distribuições de probabilidade} \end{cases}$$

A estatística do teste KS é dada por:

$$D_n = \sup_x |F(x) - F_n(x)|. \tag{2.19}$$

Esta função corresponde à distância máxima vertical entre os gráficos de $F(x)$ (distribuição acumulada assumida) e $F_n(x)$ (distribuição acumulada empírica dos dados) sobre a amplitude dos possíveis valores de x.

Como critério, comparamos a maior diferença absoluta entre as duas funções com um valor crítico, para um dado nível de significância. Neste trabalho, optou-se pela utilização de $\alpha = 5\%$.

A função distribuição acumulada assumida para os dados é definida por $F(x) = P(X \leq x_{(i)})$ e a função distribuição acumulada empírica é definida por uma função escada, dada pela fórmula:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{(-\infty,x]\}}(x_{(i)}), \tag{2.20}$$

onde I_A é uma função indicadora, que é definida da seguinte forma:

$$I_A = \begin{cases} 1 : \text{se } x \in A; \\ 0 : \text{caso contrário.} \end{cases}$$

Outra maneira de se escrever a função distribuição empírica, $F_n(x)$, é da seguinte maneira:

$$F_n(x) = \begin{cases} 0, \text{se } x < x_{(1)}; \\ \text{k/n}, \text{se } x_{(k)} \leq x \leq x_{(k+1)}; \\ 1, \text{se } x > x_{(n)}. \end{cases}$$

Capítulo 3

Modelagem da LGD via Mistura de Distribuições

Neste capítulo é descrita, em maiores detalhes, a técnica de mistura de distribuições para a modelagem dos dados de LGD. Mais especificamente, são mostradas as expressões do modelo considerado neste trabalho, a mistura de duas distribuições *Kumaraswamy*.

Essa mistura será descrita de forma a deixar o leitor informado sobre suas propriedades, comportamentos, expressões, e estatísticas básicas como algumas medidas de posição e dispersão, expressões para obtenção dos momentos e gráficos de densidade e da distribuição acumulada.

3.1 Mistura de Distribuições *Kumaraswamy*

A mistura de duas distribuições *Kumaraswamy* tem sua função densidade de probabilidade (fdp) dada por:

$$h(x;\Theta) \;=\; p_1 f_1(x;\Theta_1) + p_2 f_2(x;\Theta_2), \quad p_1 + p_2 = 1, \tag{3.1}$$

onde $\Theta = (a_1, a_2, b_1, b_2, p_1)$ e $f_i(x;\Theta_i)$ é a função densidade de probabilidade da $i-$ésima componente, dada por:

$$f_i(x;\Theta_i) \quad = \quad a_i b_i x^{a_i-1}(1-x^{a_i})^{b_i-1}, \quad 0 \leq x \leq 1, \quad a_i, b_i > 0, \quad i = 1, 2. \tag{3.2}$$

A fda associada ao modelo 3.1 é

$$F(x;\Theta) \quad = \quad p_1 F_1(x;\Theta_1) + p_2 F_2(x;\Theta_2), \tag{3.3}$$

onde $F_i(x;\Theta_i)$, é a i-ésima componente, dada por:

$$F_i(x;\Theta_i) \quad = \quad 1-(1-x^{a_i})^{b_i}, \quad x \in [0,1], \quad a_i, b_i > 0. \tag{3.4}$$

A figura 3.1, ilustra o comportamento da densidade da mistura de distribuições *Kumaraswamy* para diversas combinações de parâmetros.

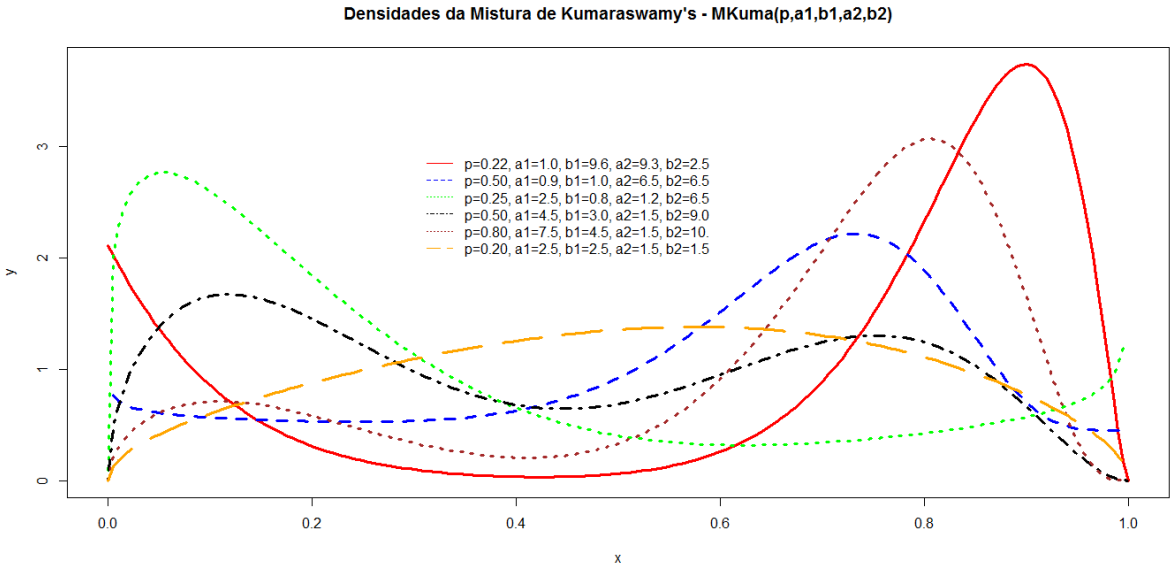


Figura 3.1: Função densidade de probabilidade da mistura de duas distribuições *Kumaraswamy* para diversas combinações de parâmetros

Com a notação utilizada em 3.2, podemos reescrever 3.1, como

$$h(x) = p_1 a_1 b_1 x^{a_1-1}(1-x^{a_1})^{b_1-1} + (1-p_1) a_2 b_2 x^{a_2-1}(1-x^{a_2})^{b_2-1} \tag{3.5}$$

As principais medidas estatísticas de uma v.a. X cuja fdp é dada por (3.5), são descritas na próxima seção.

3.1.1 Momentos

A partir da expressão dos momentos de ordem r podem ser calculadas diversas estatísticas sobre a distribuição.

$$\begin{aligned} E(X^r) &= \int_0^1 x^r f_X(x)dx \\ &= \int_0^1 x^r \left[p_1 a_1 b_1 x^{a_1-1}(1-x^{a_1})^{b_1-1} + (1-p_1)a_2 b_2 x^{a_2-1}(1-x^{a_2})^{b_2-1}\right] dx \\ &= \int_0^1 p_1 a_1 b_1 x^{a_1+r-1}(1-x^{a_1})^{b_1-1}dx + \int_0^1 p_2 a_2 b_2 x^{a_2+r-1}(1-x^{a_2})^{b_2-1}dx. \end{aligned}$$

Para facilitar a manipulação algébrica, pode-se reescrever a expressão acima da seguinte forma:

$$E(X^r) = \int_0^1 p_1 a_1 b_1 x^{(a_1+r)-1}(1-x^{a_1})^{b_1-1}dx + \int_0^1 p_2 a_2 b_2 x^{(a_2+r)-1}(1-x^{a_2})^{b_2-1}dx. \quad (3.6)$$

Dessa forma, pode-se utilizar $\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$, o resultado da função Beta, em (3.6), a fim de resolver as integrais. Para tanto, deve-se fazer as seguintes substituições: $x^{a_1} = z$ (o que implica em $x = z^{1/a_1}$ e $dx = \frac{1}{a_1}z^{\frac{1}{a_1}-1}$) e $x^{a_2} = w$ (o que implica em $x = w^{1/a_2}$ e $dx = \frac{1}{a_2}w^{\frac{1}{a_2}-1}$) que, sendo utilizadas em (3.6), deixam a expressão dos momentos de ordem r com o seguinte formato:

$$\begin{aligned} E(X^r) &= p_1 b_1 \int_0^1 z^{(\frac{r}{a_1}+1)-1}(1-z)^{b_1-1}dz + p_2 b_2 \int_0^1 w^{(\frac{r}{a_2}+1)-1}(1-w)^{b_2-1}dw \\ &= p_1 b_1 \beta(\frac{r}{a_1}+1; b_1) + p_2 b_2 \beta(\frac{r}{a_2}+1; b_2). \end{aligned} \quad (3.7)$$

Em particular, obtemos:

$$E(X) = p_1 b_1 \beta(\frac{1}{a_1}+1; b_1) + p_2 b_2 \beta(\frac{1}{a_2}+1; b_2), \quad (3.8)$$

e

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \left[p_1 b_1 \beta(\frac{2}{a_1}+1; b_1) + p_2 b_2 \beta(\frac{2}{a_2}+1; b_2)\right] - \left[p_1 b_1 \beta(\frac{1}{a_1}+1; b_1) + p_2 b_2 \beta(\frac{1}{a_2}+1; b_2)\right]^2. \end{aligned} \quad (3.9)$$

3.1.2 Taxa de Falha

A taxa de falha, denotada neste trabalho como $w(x)$ é dada pela relação $h(x)/\left(1-H(x)\right)$, onde $h(x)$ é a densidade da mistura e $H(x)$ a distribuição acumulada. A expressão $1-H(x)$ corresponde à função de sobrevivência, denotada por $R(x)$. Portanto,

$$\begin{aligned} w(x) &= \frac{h(x)}{R(x)} \\ &= \frac{p_1 a_1 b_1 x^{a_1-1}(1-x^{a_1})^{b_1-1} + (1-p_1)a_2 b_2 x^{a_2-1}(1-x^{a_2})^{b_2-1}}{1-\{p_1\left[1-(1-x^{a_1})^{b_1}\right] + p_2\left[1-(1-x^{a_2})^{b_2}\right]\}}. \end{aligned} \tag{3.10}$$

3.1.3 Função Geratriz de Momentos

A função geratriz de momentos de uma variável aleatória X com função de distribuição acumulada *Kumaraswamy* $F_K(x) = [1-(1-x^a)^b]$, $x \in [0,1]$, $a, b > 0$, e função de densidade de probabilidade $f_K(x) = abx^{a-1}(1-x^a)^{b-1}$, é dada por

$$M_X(t) = b\Gamma(b) H_{12}^{11}\left[-t\Big|_{(0,1),(-b,1/a)}^{(0,1/a)}\right], \quad t<0, \tag{3.11}$$

onde H_{12}^{11} é a função de Fox ou função H, definida em termos de uma integral de contorno de funções Gama em seu integrando. De acordo com [Mathai *et al.* , 2009], a função H é dada por

$$H_{pq}^{mn}\left[z\Big|_{(bm,Bm),(bq,Bq)}^{(an,An),(ap,Ap)}\right] = \frac{1}{2\pi_i}\int_L \frac{\prod_{j=1}^{n}\Gamma(b_j+B_j s)\prod_{j=1}^{n}\Gamma(1-a_j-A_j s)z^{-s}}{\prod_{j=m+1}^{q}\Gamma(1-b_j-B_j s)\prod_{j=n+1}^{p}\Gamma(a_j+A_j s)}ds, \tag{3.12}$$

onde A_j e B_j são números reais não-negativos e a_j e b_j são números complexos. O contorno L liga os pontos $c-i\infty$ a $c+i\infty$ tal que os polos de $\Gamma(b_j+B_j s)$, $j=1\ldots n$ ficam à esquerda de L e os polos de $\Gamma(1-a_j-A_j s)$ ficam à direita.

Para o cálculo de $M_X(t)$, basta usar a representação $e^{-z} = \frac{1}{2\pi i}\int_L \Gamma(s) z^{-s} ds$ na integral

$$\begin{aligned} M_X(t) &= b\int_a^b a x^{a-1} e^{zx}(1-x^a)^{b-1}dx \\ &= b\int_a^b e^{tu^{1/a}}(1-u)^{b-1}du, \end{aligned}$$

quando substituímos x^a por u. Assim,

$$\begin{aligned} M_X(t) &= b\frac{1}{2\pi i}\int_L \Gamma(s)(-t)^{-s}\left(\int_0^1 u^{-\frac{s}{a}}(1-u)^{b-1}du\right)ds, \quad t<0 \\ &= \frac{b}{2\pi i}\int_L \Gamma(s)(-t)^{-s}\frac{\Gamma(1-\frac{s}{a})\Gamma(b)}{\Gamma(1+b-\frac{s}{a})}ds \\ &= \frac{b\Gamma(b)}{2\pi i}\int_L \frac{\Gamma(s)\Gamma(1-\frac{s}{a})(-t)^{-s}}{\Gamma(1+b-\frac{s}{a})}ds \\ &= b\Gamma(b)H_{1\ 2}^{1\ 1}\left[-t\Big|_{(0,\frac{1}{a})}^{(0,1),(-b,\frac{1}{a})}\right], \quad t<0. \end{aligned} \tag{3.13}$$

O cálculo da função H pode ser realizado através do software maple. Em particular, quando $a = 1$ ou a e b são números inteiros essa função $H_{1\ 2}^{1\ 1}$ é uma função $G_{1\ 2}^{1\ 1}$, chamada de função Meiyer, cujo cálculo é mais simples.

3.1.4 Transformada de Laplace

Usando a transformada de Laplace, podemos obter a expressão da função geradora de momentos da mistura de *Kumaraswamy* , que também permite a obtenção dos momentos de ordem r, como segue:

$$\begin{aligned} M_X(t) &= \int_0^1 e^{-tx}f_X(x)dx \\ &= \int_0^1 \sum_{k=0}^{\infty}\frac{(-1)^k(tx)^k}{k!}\left[p_1a_1b_1x^{a_1-1}(1-x^{a_1})^{b_1-1}+(1-p_1)a_2b_2x^{a_2-1}(1-x^{a_2})^{b_2-1}\right]dx \\ &= \sum_{k=0}^{\infty}\frac{(-1)^kt^k}{k!}\left[p_1a_1b_1\int_0^1 x^{k+a_1-1}(1-x^{a_1})^{b_1-1}dx + p_2a_2b_2\int_0^1 x^{k+a_2-1}(1-x^{a_2})^{b_2-1}dx\right] \\ &= p_1b_1\sum_{k=0}^{\infty}\frac{(-1)^kt^k}{k!}\int_0^1 z^{\frac{k}{a_1}}(1-z)^{b_1-1}dz + p_2b_2\sum_{k=0}^{\infty}\frac{(-1)^kt^k}{k!}\int_0^1 z^{\frac{k}{a_2}}(1-z)^{b_2-1}dz \\ &= p_1b_1\sum_{k=0}^{\infty}\frac{(-1)^kt^k}{k!}b(\frac{k}{a_1}+1;b_1) + p_2b_2\sum_{k=0}^{\infty}\frac{(-1)^kt^k}{k!}b(\frac{k}{a_2}+1;b_2), \end{aligned} \tag{3.14}$$

definida para $t > 0$.

3.1.5 Identificabilidade para Mistura de Distribuições *Kumaraswamy*

Para provar a identificabilidade da mistura de funções de distribuição *Kumaraswamy*, primeiro definimos o conjunto

$$\mathcal{K} = \left\{F_K(x) : F_K(x) = F_K(x;a,b) = 1-(1-x^a)^b, 0 \leq x \leq 1, (a,b) > 0\right\}, \tag{3.15}$$

como a família de distribuições *Kumaraswamy* e o conjunto

$$\mathcal{H}_K = \left\{H(x) : H(x) = \sum_{i=1}^{n} p_i F_{K_i}(x;a_i,b_i), \sum_{i=1}^{n} p_i = 1, p_i \geq 0\right\} \tag{3.16}$$

como a classe de todas as misturas de distribuições *Kumaraswamy* $\mathcal{K}$.

Proposição 1:

A classe $\mathcal{H}_K$ de todas as misturas de $\mathcal{K}$ é identificável.

Prova:

Seja M uma aplicação que transforma a função de distribuição $F_K \in \mathcal{K}$ em $\Phi_{F_K}(t) = \int_{\mathbb{R}} x^t dF_K(x)$, $t \in \mathbb{R}$.

M é linear, e por [Atienza *et al.* , 2006],

$$\begin{aligned} M[F_K(.;a,b)] &= b\beta(1+\frac{t}{a},b) \\ &= \Phi_{F_K}(t), \quad t \in (0,\infty) \end{aligned} \tag{3.17}$$

Com a notação do Corolário 1, $S(F_K(.;a,b)) = (0,+\infty)$. Então, $t_0 = +\infty$ verifica a primeira condição do referido corolário. A prova será completada se provarmos que a relação é dada por

$$F_{K_1} \prec F_{K_2}$$

se e somente se

$$\lim_{t\to+\infty} \frac{\Phi_{F_{K_2}}(t)}{\Phi_{F_{K_1}}(t)} = 0, \tag{3.18}$$

é uma relação de ordem total.

Usando a fórmula de Stirling, $\Gamma(z+1) \sim \sqrt{2\pi z} z^z e^{-z}$ quando $z \to +\infty$, mostra-se a fórmula assintótica, para y fixo,

$$\beta(x,y) \sim \Gamma(y) x^{-y}, \quad x \to +\infty$$

Portanto, quando $t \to \infty$,

$$\begin{aligned} \frac{\Phi_{F_2}(t)}{\Phi_{F_1}(t)} &= \frac{b_2 \beta(1+\frac{t}{a_2}, b_2)}{b_1 \beta(1+\frac{t}{a_1}, b_1)} \\ &\sim \frac{b_2 \Gamma(b_2)(1+\frac{t}{a_2})^{-b_2}}{b_1 \Gamma(b_1)(1+\frac{t}{a_1})^{-b_1}} \to 0, \end{aligned} \tag{3.19}$$

se, e somente se, $[b_1 < b_2]$. Por outro lado, usando a aproximação de Stirling em cada expressão de $\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \beta(x,y)$, temos que, quando $t \to \infty$,

$$\frac{\Phi_{F_2}(t)}{\Phi_{F_2}(t)} \sim \left(\frac{a_2}{a_1}\right)^{1/2} \left(\frac{t/a_1+b}{t/a_1}\right)^{t/a_1} \left(\frac{t/a_1+b}{t/a_2+b}\right)^{b} \exp\left\{\frac{t}{a_2}\left[\ln\left(\frac{t}{a_2}\right) - \ln\left(\frac{t}{a_2}+b\right)\right]\right\} \to 0, \tag{3.20}$$

se, e somente se, $[b_1 = b_2 = b]$. Os limites 3.19 e 3.20 mostram que a relação de ordem $F_{K_1}(\cdot, a_1, b_1) \prec F_{K_2}(\cdot, a_2, b_2)$ é válida.

Capítulo 4

Estimação dos Parâmetros

Neste capítulo o leitor irá familiarizar-se com o *algoritmo EM*, método de estimação de parâmetros utilizado neste trabalho. Para avaliar o desempenho do algoritmo EM, utilizamos simulação Monte Carlo fixando alguns conjuntos de parâmetros.

4.1 Algoritmo EM

O algoritmo EM, do inglês *Expectation-Maximization*, é um algoritmo clássico na estatística usado para determinar estimativas por máxima verossimilhança de parâmetros de distribuições paramétricas em modelos de mistura, em modelos com dados faltantes e em modelos para os quais a estimação dos parâmetros por máxima verossimilhança apresente problemas [Otiniano & Teixeira, 2014]. Tem grande utilização prática pois apresenta boas estimativas, com propriedades assintóticas ótimas.

De acordo com [Casella & Berger, 2011], o algoritmo EM teve suas origens no trabalho de [Hartley, 1958], mas entrou em evidência na estatística somente após o trabalho de [Dempster *et al.* , 1977], que detalhou a estrutura básica do algoritmo e ilustrou seu uso em uma ampla variedade de aplicações.

Segundo [Casella & Berger, 2011], o EM é um algoritmo que seguramente converge para o EMV e tem como base a ideia de substituir uma difícil maximização da verossimilhança por uma sequência de maximizações mais fáceis, cujo limite é a resposta para o problema original. Uma de suas vantagens é que são conhecidas as condições de con-

vergência para os estimadores de máxima verossimilhança de dados incompletos, embora a prova original dessa convergência, proposta por [Dempster *et al.* , 1977], tenha apresentado uma falha, que posteriormente, nos trabalhos de [Boyles, 1983] e [Wu, 1983] foi corrigida.

Cada iteração do algoritmo EM envolve dois passos: um passo E (esperança) e um passo M (maximização). Resumindo, o passo E calcula o valor esperado do logaritmo da verossimilhança, e o passo M encontra seu máximo.

Esses passos devem ser repetidos até se atingir uma convergência. Para isto, pode ser adotado como critério de parada, por exemplo, $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$, onde θ é o parâmetro e ϵ é um valor arbitrário, maior que zero.

Em termos mais formais, seja X uma v.a. com fdp $h(x;\Theta)$, mistura de duas componentes, dada por:

$$h(x;\Theta) = p_1 f_1(x;\theta_1) + p_2 f_2(x;\theta_2), \quad p_1 + p_2 = 1. \tag{4.1}$$

onde $\Theta = (p_1, \theta_1, \theta_2)$.

O algoritmo EM é utilizado para obter estimativas de Θ, baseado nos valores $x_1, \ldots, x_N$ de uma amostra aleatória de tamanho N da v.a. X.

De acordo com [Otiniano & Teixeira, 2014], uma breve descrição do algoritmo EM é a seguinte: se assumirmos que X é observado e gerado por alguma distribuição paramétrica $f(x;\Theta)$, chamamos X ($[x_1, \ldots, x_N]$) de dados incompletos. Esses dados são completados com Y ($[y_1, \ldots, y_N]$), então $Z = (X, Y)$ são dados completos cuja densidade conjunta é

$$\begin{aligned} f(z;\Theta) &= f(x, y;\Theta) \\ &= f(y|x;\Theta) f(x;\Theta) \end{aligned} \tag{4.2}$$

Com essa nova densidade, define-se a função de verossimilhança dos dados completos

$$\begin{aligned} L(\Theta|z) &= L(\Theta|x, y) \\ &= f(X, Y;\Theta). \end{aligned} \tag{4.3}$$

O algoritmo EM alterna o passo E (da esperança), onde obtém-se

$$Q(\Theta, \Theta^{(k)}) = E\left\{\ln[f(X, Y|\Theta)]|X, \Theta^{(k)}\right\}, \tag{4.4}$$

com o passo M (da maximização), onde se calcula $\Theta^{(k+1)}$ ao maximizar $Q(\Theta, \Theta^{(k)})$, por meio da expressão

$$Q(\Theta, \Theta^{(k)}) = \sum_{l=1}^{2}\sum_{i=1}^{N} \ln\left[p_l f_l(x_i; \theta_l^{(k)})\right] f(l|x_i; \Theta^{(k)}). \tag{4.5}$$

Em síntese, na $(k+1)$-ésima iteração do passo E, a atualização da estimativa de p_1 é dada por

$$p_1^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N} f(1|x_i; \theta_1^{(k)}), \tag{4.6}$$

onde,

$$f(l|x_i; \theta_l^{(k)}) \;\; = \;\; \frac{p_l^{(k)} f_l(x_i; \theta_l^{(k)})}{\sum_{l=1}^{2} p_l^{(k)} f_l(x_i; \theta_l^{(k)})}, \quad l = 1, 2, \tag{4.7}$$

e na $(k+1)$-ésima iteração do passo M, a atualização das estimativas de $\theta_l^{(k+1)}$, $l = 1, 2$ é obtida resolvendo as equações

$$\sum_{i=1}^{N} f(l|x_i; \theta_l^{(k)}) \frac{\partial}{\partial \theta_l^{(k)}} \ln\left[f_l(x_i; \theta_l^{(k)})\right] = 0. \tag{4.8}$$

A seguir são mostradas as aplicações do algoritmo EM nos modelos de misturas de distribuições *Kumaraswamy*.

4.2 EM para Mistura de Distribuições *Kumaraswamy*

Nesta seção, são mostradas as expressões do modelo de mistura de distribuições *Kumaraswamy* que foram empregadas na implementação do algoritmo EM. Num primeiro

momento, temos que tal modelo tem a seguinte expressão de densidade:

$$h(x;\Theta) = p_1 a_1 b_1 x^{a_1-1}(1-x^{a_1})^{b_1-1} + p_2 a_2 b_2 x^{a_2-1}(1-x^{a_2})^{b_2-1}, \tag{4.9}$$

e a partir dessa expressão, fica claro que os parâmetros a serem estimados são p_1, p_2, $\theta_1 = (a_1, b_1)$ e $\theta_2 = (a_2, b_2)$, referentes às duas componentes.

O primeiro passo do algoritmo consiste no passo E, da esperança, que tem a função de estimar os pesos de cada uma das componentes da mistura. Essa etapa é realizada neste modelo pela utilização da seguinte expressão:

$$f(l|x_i;\theta_l^{(k)}) \;=\; \frac{p_l a_l b_l x^{a_l-1}(1-x^{a_l})^{b_l-1}}{\sum_{l=1}^{2} p_l a_l b_l x^{a_l-1}(1-x^{a_l})^{b_l-1}}, \quad l = 1, 2. \tag{4.10}$$

Em seguida, para a estimação dos demais parâmetros da densidade da mistura, é realizado o passo M, de maximização, por meio da seguinte expressão:

$$Q(\Theta;\Theta^{(k+1)}) \;=\; \sum_{l=1}^{2}\sum_{i=1}^{N} f(l|x_i;\theta_l^{(k)}) \frac{\partial}{\partial\theta_l} \ln\left\{p_l f_l(x_i;\theta_j^{(k)})\right\} \tag{4.11}$$

Como as duas componentes da mistura provém da mesma distribuição de probabilidade, tem-se que as expressões dos parâmetros que foram obtidas após as derivações de 4.11, para $l = \{1, 2\}$ são as mesmas para as duplas de parâmetros (a_1, a_2) e (b_1, b_2). Portanto, na $(k+1)-$ésima iteração, as estimativas de $a_l, l = 1, 2$ são obtidas ao resolver

$$\sum_{i=1}^{N} f(l|x_i;\theta^{(k)}) \left\{\frac{1}{a_l} + \ln(x_i) - \frac{(b_l-1)x_i^{a_l}\ln(x_i)}{1-x_i^{a_l}}\right\} \;=\; 0 \tag{4.12}$$

e de $b_l, l = 1, 2$, ao calcular

$$b_l \;=\; -\frac{\sum_{i=1}^{N} f(l/x_i;\theta^{(k)})}{\sum_{i=1}^{N} f(l/x_i;\theta^{(k)})\ln(1-x_i^{a_l})}. \tag{4.13}$$

As expressões mostradas acima são aplicadas de forma iterativa para a obtenção das estimativas. Para a resolução da equação 4.12 deve ser utilizado algum método numérico

como, por exemplo o *Newton Raphson*.

4.3 Simulação

O processo de simulação descrito a seguir foi utilizado para testar as estimativas dos parâmetros, realizadas por meio do algoritmo EM, descrito anteriormente. O passo-a-passo da simulação é mostrado abaixo.

1. Geração de amostras aleatórias dos valores de uma distribuição *Kumaraswamy*, por meio da inversa da distribuição acumulada, para cada escolha do vetor de parâmetros Θ_i;

2. Geração de amostra aleatória da variável aleatória X, cuja densidade é o modelo de mistura, da seguinte forma:

   (a) Geração de duas variáveis aleatórias uniformes u_1 e u_2;

   (b) Se $u_1 < p_1$, onde p_1 é o peso da primeira componente da mistura, então foi usado u_2 para gerar um valor x da v.a. X, onde $x = F_1^{-1}(u_2)$ e F_1 é a distribuição acumulada de f_1;

   (c) Se $u_1 \geq p_1$, então, foi usado u_2 para gerar um valor x, da variável aleatória X, onde $x = F_2^{-1}(u_2)$ e F_2 é a distribuição acumulada de f_2.

3. Aplicação dos passos E e M, do algoritmo EM;

4. Obtenção do vetor θ_i das estimativas;

5. Repetição de todo o processo acima descrito por 100 vezes para ao final calcular as estimativas dos parâmetros como a média dos 100 vetores de estimativas.

Foram definidos vários experimentos, valores de Θ, para testar o comportamento dos estimadores do algoritmo EM.

4.3.1 Resultados da Simulação de Dados para Mistura de Distribuições *Kumaraswamy*

Nesta seção, calculamos as estimativas dos cinco parâmetros a_1, a_2, b_1, b_2 e p, nessa ordem, que aparecem na função densidade de probabilidade da mistura de distribuições

Kumaraswamy, por meio do algoritmo EM e simulação de Monte Carlo para quatro conjuntos de parâmetros:

$$\begin{aligned}
\Theta_1 &= (4.31, 1.59, 3.14, 10.15, 0.5) \\
\Theta_2 &= (5.30, 1.00, 1.00, 9.00, 0.50) \\
\Theta_3 &= (35.0, 1.50, 1.00, 9.50, 0.75) \\
\Theta_4 &= (40.0, 0.50, 15.0, 0.80, 0.85)
\end{aligned}$$

A partir dos dados simulados, foram criadas tabelas com as estimativas dos parâmetros, o viés associado e o erro quadrático médio (EQM) para diferentes tamanhos de amostra $(250, 500, 1000)$. Por exemplo, para o conjunto de dados Θ_1, a tabela 4.1, apresenta as informações supracitadas. Vinculadas a esta tabela estão os conjunto de figuras 4.1 (n=250), 4.2 (n=500), e 4.3 (n=1000), cada um trazendo figuras como as seguintes: 4.1(a), que corresponde ao gráfico de densidades simulada e estimada; 4.1(b), que corresponde ao gráfico da distribuição acumulada empírica dos dados e distribuição acumulada estimada; e 4.1(c), que traz as curvas de distribuição acumulada estimada para cada realização do algoritmo EM, juntamente com a curva de distribuição acumulada empírica dos dados.

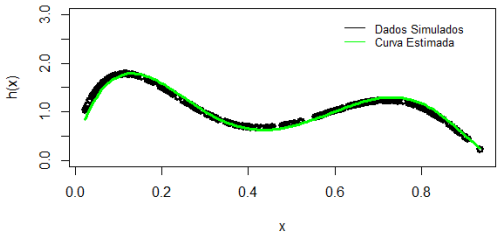
Como pode ser verificado por meio das tabelas e também dos gráficos, o algoritmo implementado para a estimação dos parâmetros apresentou resultados muito satisfatórios, o que foi corroborado por meio dos testes KS, mostrando que as estimativas apresentaram resultados próximos dos valores reais dos parâmetros.

Os gráficos como o 4.1(b) deixam claro a importância de empregar o algoritmo repetidas vezes, para então utilizar o valor médio das estimativas como valores finais do processo de estimação, pois observa-se, graficamente, a variabilidade das estimativas dos parâmetros.

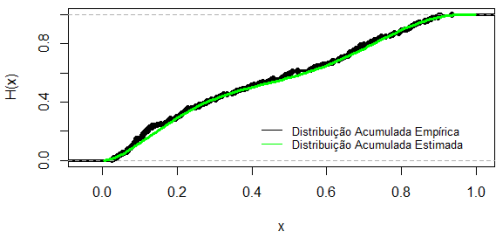
A tabela 4.5 apresenta informações referentes ao teste de Kolmorogov-Smirnov, para os três tamanhos de amostra, para cada um dos quatro conjuntos de parâmetros. A última coluna da tabela 4.5 apresenta a decisão após a aplicação do teste KS, cuja hipótese nula (H_0) declarava que os dados provêm da mesma distribuição de probabilidade.

Tabela 4.1: Estimativas dos parâmetros, viés e EQM dos parâmetros do experimento 1 $(4.31, 1.59, 3.14, 10.15, 0.5)$

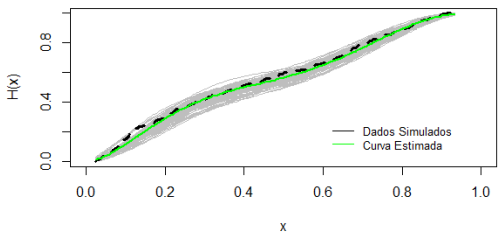
n	Parâmetros	Estimativas	Viés	EQM
250	$a_1 = 4.31$	$\hat{a}_1 = 4.369$	0.05949	0.52490
	$a_2 = 1.59$	$\hat{a}_2 = 1.618$	0.02757	0.03032
	$b_1 = 3.14$	$\hat{b}_1 = 3.246$	0.10603	0.44055
	$b_2 = 10.15$	$\hat{b}_2 = 11.205$	1.05504	24.10611
	$p = 0.5$	$\hat{p} = 0.501$	0.00056	0.00134
500	$a_1 = 4.31$	$\hat{a}_1 = 4.396$	0.08558	0.33247
	$a_2 = 1.59$	$\hat{a}_2 = 1.606$	0.01615	0.01460
	$b_1 = 3.14$	$\hat{b}_1 = 3.193$	0.05337	0.20077
	$b_2 = 10.15$	$\hat{b}_2 = 10.747$	0.59679	9.19681
	$p = 0.5$	$\hat{p} = 0.492$	0.00828	0.00067
1000	$a_1 = 4.31$	$\hat{a}_1 = 4.352$	0.04210	0.15767
	$a_2 = 1.59$	$\hat{a}_2 = 1.594$	0.00423	0.00795
	$b_1 = 3.14$	$\hat{b}_1 = 3.174$	0.03387	0.10443
	$b_2 = 10.15$	$\hat{b}_2 = 10.446$	0.29569	4.66183
	$p = 0.5$	$\hat{p} = 0.498$	0.00203	0.00033



(a) Dados Simulados e Curva Estimada

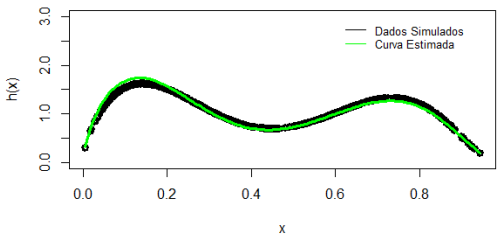


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

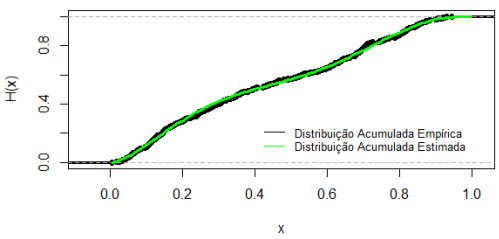


(c) Simulação Monte Carlo para o Algoritmo EM

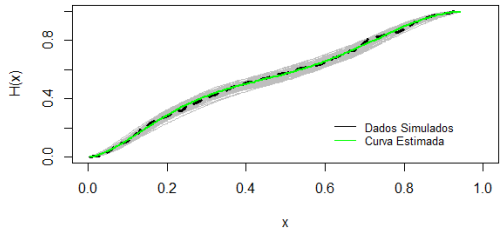
Figura 4.1: Amostra de tamanho 250 e gráficos para o conjunto de parâmetros Θ_1



(a) Dados Simulados e Curva Estimada

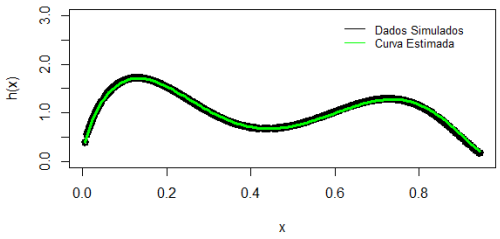


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

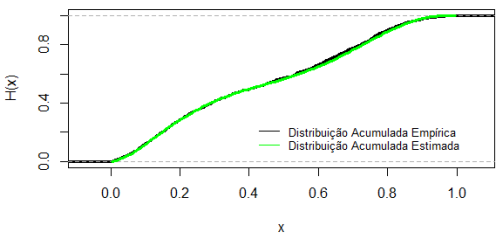


(c) Simulação Monte Carlo para o Algoritmo EM

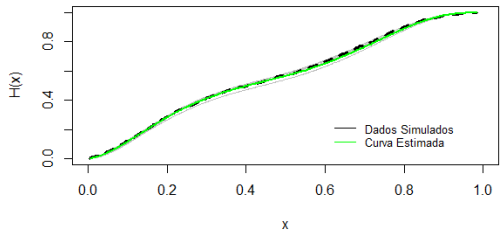
Figura 4.2: Amostra de tamanho 500 e gráficos para o conjunto de parâmetros Θ_1



(a) Dados Simulados e Curva Estimada



(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

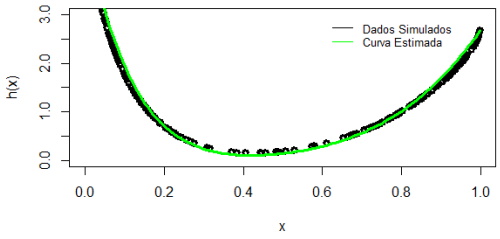


(c) Simulação Monte Carlo para o Algoritmo EM

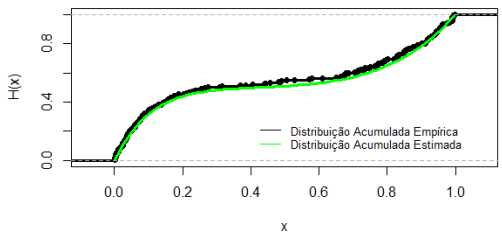
Figura 4.3: Amostra de tamanho 1000 e gráficos para o conjunto de parâmetros Θ_1

Tabela 4.2: Estimativas dos parâmetros, viés e EQM dos parâmetros do experimento 2 $(5.30, 1.00, 1.00, 9.00, 0.50)$

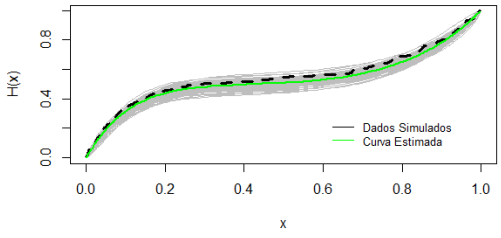
n	Parâmetros	Estimativas	Viés	EQM
	$a_1 = 5.3$	$\hat{a}_1 = 5.351$	0.05139	0.51968
	$a_2 = 1$	$\hat{a}_2 = 1.004$	0.00369	0.00895
250	$b_1 = 1$	$\hat{b}_1 = 0.999$	0.00073	0.01437
	$b_2 = 9$	$\hat{b}_2 = 9.627$	0.62692	6.16252
	$p = 0.5$	$\hat{p} = 0.502$	0.00186	0.00077
	$a_1 = 5.3$	$\hat{a}_1 = 5.310$	0.00982	0.29794
	$a_2 = 1$	$\hat{a}_2 = 1.009$	0.00896	0.00426
500	$b_1 = 1$	$\hat{b}_1 = 1.011$	0.01110	0.00764
	$b_2 = 9$	$\hat{b}_2 = 9.362$	0.36193	2.50299
	$p = 0.5$	$\hat{p} = 0.499$	0.00137	0.00046
	$a_1 = 5.3$	$\hat{a}_1 = 5.323$	0.02260	0.14015
	$a_2 = 1$	$\hat{a}_2 = 1.009$	0.00904	0.00236
1000	$b_1 = 1$	$\hat{b}_1 = 1.012$	0.01155	0.00397
	$b_2 = 9$	$\hat{b}_2 = 9.255$	0.25488	1.44694
	$p = 0.5$	$\hat{p} = 0.503$	0.00277	0.00023



(a) Dados Simulados e Curva Estimada

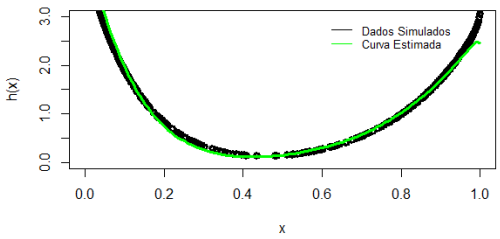


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

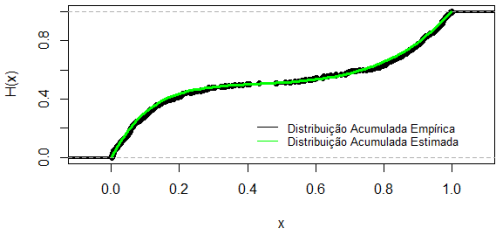


(c) Simulação Monte Carlo para o Algoritmo EM

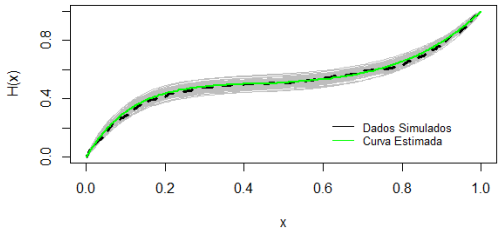
Figura 4.4: Amostra de tamanho 250 e gráficos para o conjunto de parâmetros Θ_2



(a) Dados Simulados e Curva Estimada

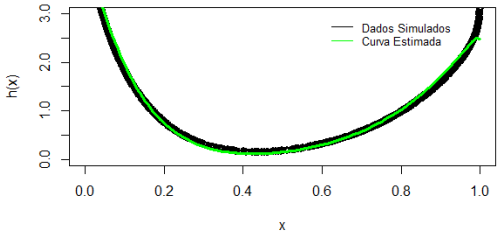


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

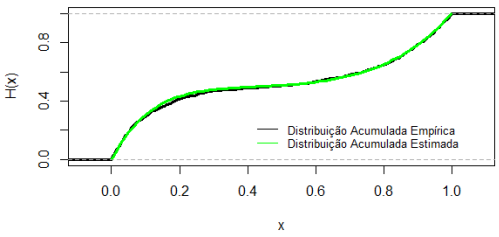


(c) Simulação Monte Carlo para o Algoritmo EM

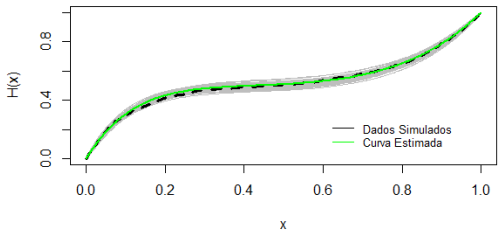
Figura 4.5: Amostra de tamanho 500 e gráficos para o conjunto de parâmetros Θ_2



(a) Dados Simulados e Curva Estimada



(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

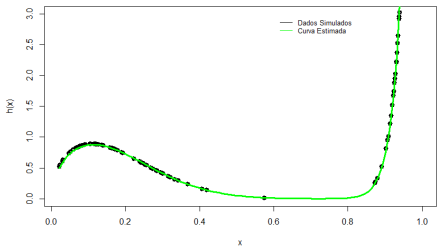


(c) Simulação Monte Carlo para o Algoritmo EM

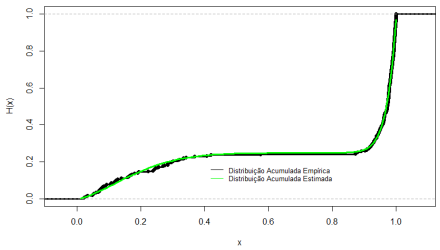
Figura 4.6: Amostra de tamanho 1000 e gráficos para o conjunto de parâmetros Θ_2

Tabela 4.3: Estimativas dos parâmetros, viés e EQM dos parâmetros do experimento 3 $(35.0, 1.50, 1.00, 9.50, 0.75)$

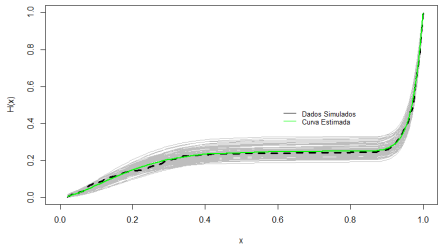
n	Parâmetros	Estimativas	Viés	EQM
	$a_1 = 35$	$\hat{a}_1 = 35.688$	0.68771	10.70159
	$a_2 = 1.5$	$\hat{a}_2 = 1.527$	0.02655	0.02734
250	$b_1 = 1$	$\hat{b}_1 = 1.021$	0.02090	0.01202
	$b_2 = 9.5$	$\hat{b}_2 = 10.288$	0.78804	7.17127
	$p = 0.75$	$\hat{p} = 0.752$	0.00229	0.00086
	$a_1 = 35$	$\hat{a}_1 = 35.436$	0.43598	6.97361
	$a_2 = 1.5$	$\hat{a}_2 = 1.505$	0.00450	0.01782
500	$b_1 = 1$	$\hat{b}_1 = 1.004$	0.00361	0.00465
	$b_2 = 9.5$	$\hat{b}_2 = 9.957$	0.45707	4.68879
	$p = 0.75$	$\hat{p} = 0.748$	0.00189	0.00041
	$a_1 = 35$	$\hat{a}_1 = 35.144$	0.14394	2.67736
	$a_2 = 1.5$	$\hat{a}_2 = 1.505$	0.00466	0.00852
1000	$b_1 = 1$	$\hat{b}_1 = 1.000$	0.00015	0.00210
	$b_2 = 9.5$	$\hat{b}_2 = 9.713$	0.21262	2.28757
	$p = 0.75$	$\hat{p} = 0.751$	0.00087	0.00015



(a) Dados Simulados e Curva Estimada

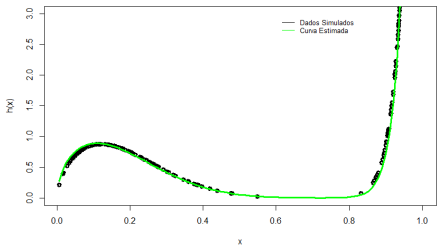


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

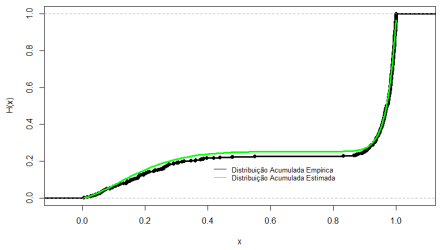


(c) Simulação Monte Carlo para o Algoritmo EM

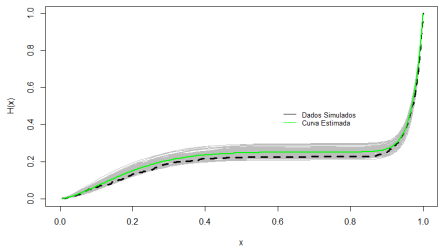
Figura 4.7: Amostra de tamanho 250 e gráficos para o conjunto de parâmetros Θ_3



(a) Dados Simulados e Curva Estimada

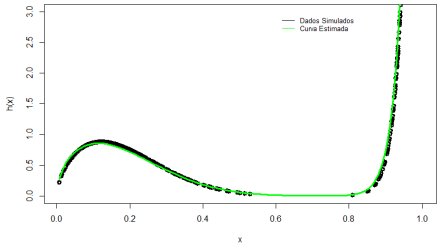


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

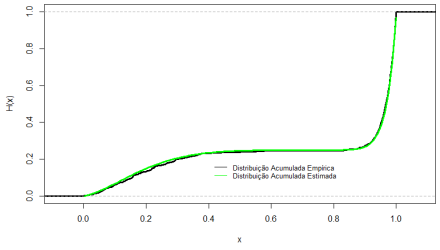


(c) Simulação Monte Carlo para o Algoritmo EM

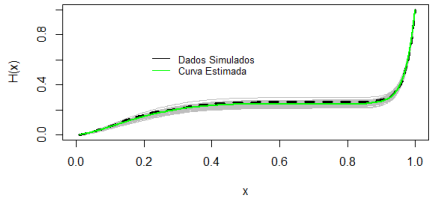
Figura 4.8: Amostra de tamanho 500 e gráficos para o conjunto de parâmetros Θ_3



(a) Dados Simulados e Curva Estimada



(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

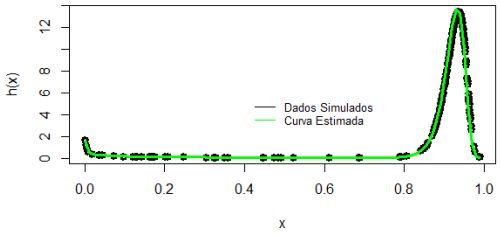


(c) Simulação Monte Carlo para o Algoritmo EM

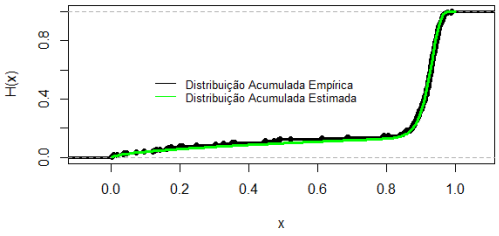
Figura 4.9: Amostra de tamanho 1000 e gráficos para o conjunto de parâmetros Θ_3

Tabela 4.4: Estimativas dos parâmetros, viés e EQM dos parâmetros do experimento 4 $(40.0, 0.50, 15.0, 0.80, 0.85)$

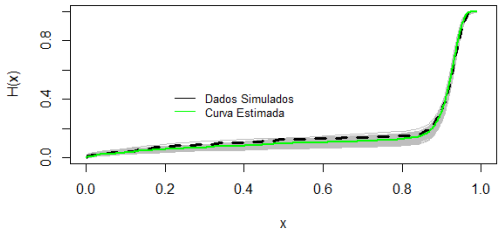
n	Parâmetros	Estimativas	Viés	EQM
	$a_1 = 40$	$\hat{a}_1 = 39.659$	0.34103	6.02221
	$a_2 = 0.5$	$\hat{a}_2 = 0.558$	0.05836	0.01852
250	$b_1 = 15$	$\hat{b}_1 = 14.958$	0.04219	7.84781
	$b_2 = 0.8$	$\hat{b}_2 = 0.966$	0.16560	0.13741
	$p = 0.85$	$\hat{p} = 0.853$	0.00345	0.00060
	$a_1 = 40$	$\hat{a}_1 = 40.049$	0.04902	4.99311
	$a_2 = 0.5$	$\hat{a}_2 = 0.512$	0.01193	0.00863
500	$b_1 = 15$	$\hat{b}_1 = 15.184$	0.18422	5.16629
	$b_2 = 0.8$	$\hat{b}_2 = 0.879$	0.07928	0.05175
	$p = 0.85$	$\hat{p} = 0.852$	0.01700	0.00030
	$a_1 = 40$	$\hat{a}_1 = 40.051$	0.05136	1.61733
	$a_2 = 0.5$	$\hat{a}_2 = 0.518$	0.01766	0.00354
1000	$b_1 = 15$	$\hat{b}_1 = 15.082$	0.08192	1.69508
	$b_2 = 0.8$	$\hat{b}_2 = 0.831$	0.03119	0.01443
	$p = 0.85$	$\hat{p} = 0.851$	0.00092	0.00013



(a) Dados Simulados e Curva Estimada

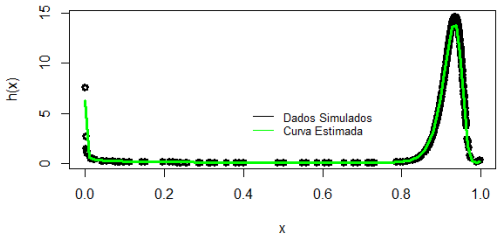


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

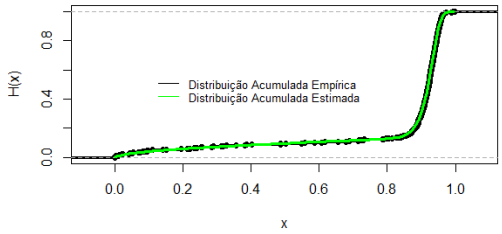


(c) Simulação Monte Carlo para o Algoritmo EM

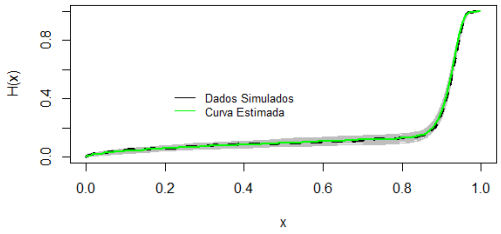
Figura 4.10: Amostra de tamanho 250 e gráficos para o conjunto de parâmetros Θ_4



(a) Dados Simulados e Curva Estimada

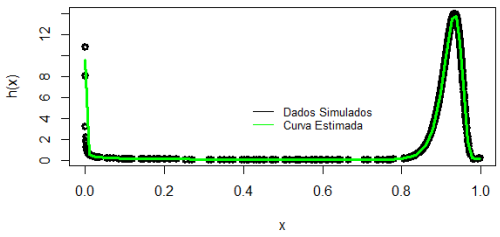


(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada

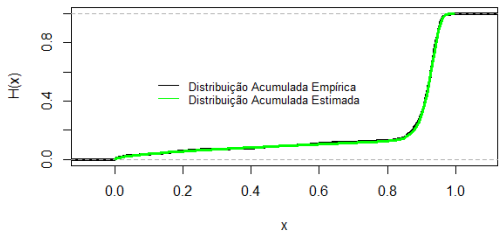


(c) Simulação Monte Carlo para o Algoritmo EM

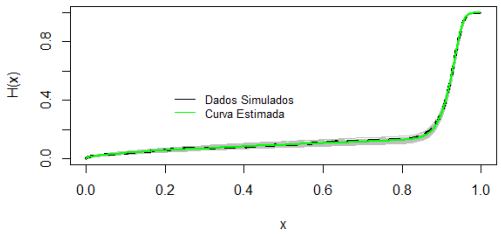
Figura 4.11: Amostra de tamanho 500 e gráficos para o conjunto de parâmetros Θ_4



(a) Dados Simulados e Curva Estimada



(b) Distribuição Acumulada Empírica e Distribuição Acumulada Estimada



(c) Simulação Monte Carlo para o Algoritmo EM

Figura 4.12: Amostra de tamanho 1000 e gráficos para o conjunto de parâmetros Θ_4

Tabela 4.5: KS, p-valor e decisão

Experimento	n	KS	p-Valor	Decisão
$\theta_1(4.31, 1.59, 3.14, 10.15, 0.50)$	250	0.044	0.9689	Não Rejeita H_0
	500	0.028	0.9895	Não Rejeita H_0
	1000	0.018	0.9969	Não Rejeita H_0
$\theta_2(5.30, 1.00, 1.00, 9.00, 0.50)$	250	0.044	0.9689	Não Rejeita H_0
	500	0.030	0.9780	Não Rejeita H_0
	1000	0.027	0.8593	Não Rejeita H_0
$\theta_2(35.0, 1.50, 1.00, 9.50, 0.75)$	250	0.036	0.9969	Não Rejeita H_0
	500	0.036	0.9022	Não Rejeita H_0
	1000	0.024	0.9356	Não Rejeita H_0
$\theta_2(40.0, 0.50, 15.0, 0.80, 0.85)$	250	0.064	0.6852	Não Rejeita H_0
	500	0.040	0.8186	Não Rejeita H_0
	1000	0.024	0.9356	Não Rejeita H_0

Capítulo 5

Aplicação em Dados Reais

Neste capítulo, o modelo de mistura de distribuições *Kumaraswamy* desenvolvido neste trabalho foi aplicado em dados reais de perda (LGD), de uma instituição financeira (IF) do Brasil, relacionados a dois produtos em específico. Entretanto, não serão dados maiores detalhes da instituição nem dos produtos aos quais os dados de perda estão associados por uma questão de sigilo. Essa restrição em nada prejudica o trabalho.

É mostrado também o uso que este tipo de modelagem pode ter dentro de uma instituição financeira que produz seus próprios modelos de parâmetros de risco, em especial, para os modelos de LGD. Especificamente, é mostrada a importância deste trabalho no contexto do monitoramento de modelos.

Os dados de LGD apresentam algumas características próprias, como por exemplo, a bimodalidade da distribuição, ocasionada pela concentração de valores nos extremos do intervalo de valores [0,1]. Acontece que, em se tratando de perda, os clientes são direcionados a um processo de recuperação de crédito, onde por algum tempo a IF acredita que conseguirá reaver parte dos valores emprestados. Dessa forma, tem-se que uma parte dos clientes irá pagar o que deve, gerando uma concentração de valores próximos do valor 0 de LGD; e uma outra parte dos clientes não irá honrar seu compromisso com nenhum pagamento. Estes, que são maioria, se concentram ao redor do valor 1 de LGD, quando a IF efetivamente declara perdido o valor emprestado. Entre esses dois extremos há a presença de uma pequena massa de clientes que restitui a IF em valores intermediários entre a quantia total devida e nenhum valor.

A seguir, são descritos os conjuntos de dados utilizados na aplicação do modelo de mistura de distribuições *Kumaraswamy*, por meio de medidas descritivas e gráficos. Também são mostrados os resultados das estimativas dos parâmetros após a aplicação do algoritmo EM; os gráficos histograma com as curvas de densidade e os gráficos das distribuições acumuladas estimada e empírica, para cada caso.

Foi realizado o teste KS para as distribuições acumuladas e estes também são apresentados em tabelas juntamente com as estimativas dos parâmetros e a decisão sob a hipótese H_0 de que os conjuntos de dados apresentam mesma distribuição da curva estimada. Entretanto, cabe destacar que ambos conjuntos de dados são compostos por um grande número de observações, e que essa quantidade elevada de elementos da população alvo do estudo tem influência na aplicação do teste estatístico de verificação do ajuste do modelo. Uma grande quantidade de elementos pode tornar o teste KS sensível demais à pequenos valores da estatística de distância entre as distribuições, atribuindo de forma imprópria (na prática) um p-valor excessivamente baixo, e fazendo com que a decisão a ser tomada no teste de hipóteses seja de rejeição de H_0, mesmo quando, na prática, a diferença apresentada não seja relevante.

Para contornar essa situação, a análise do teste KS foi realizada por meio do cálculo de um *p-valor* empírico. Para isso, foi realizado um processo de reamostragem dos dados das populações utilizadas, e para cada uma das amostras foi calculada a maior distância entre as distribuições empírica (com os dados originais) e estimada (com os dados da reamostragem). Após 1000 repetições deste processo, foi construído o gráfico com a distribuição das estatísticas KS das 1000 repetições e foi calculado o número de vezes em que os valores observados foram maiores ou iguais ao valor original de KS, a referência. Dessa forma, foram obtidos os p-valores empíricos de cada um dos conjuntos de dados, e foi possível concluir através do teste o que estava sendo percebido pela análise gráfica.

A seguir, na tabela 5.3 o leitor também encontrará os valores dos p-valores empíricos de cada conjunto de dados, bem como as respectivas decisões quanto à hipótese nula, e os histogramas com o comportamento dos valores apurados de KS no processo de reamostragem.

5.1 Conjuntos de Dados e Aplicação do Algoritmo para Estimação dos Parâmetros do Modelo de Mistura de Distribuições *Kumaraswamy*

Os dados reais utilizados neste estudo referem-se a dois produtos diferentes, mas pela tabela 5.1 é possível verificar que apresentam comportamentos semelhantes quanto à assimetria, curtose e distribuição. São dados compreendidos no período de 30 de junho de 2006 a 30 de setembro de 2008.

Tabela 5.1: Medidas descritivas dos conjuntos de dados reais de LGD

Prod.	Qtd Registros	Mín	Máx	Média	Mediana	Variância	Curtose	Assimetria
1	117,409	0	1.0	0.57306	0.73342	0.14527	1.32807	-0.27254
2	159,410	0	1.0	0.72527	0.89922	0.11534	2.77389	-1.22189

A partir dos resultados apresentados na tabela 5.1 é possível concluir pela assimetria negativa de ambos os conjuntos de dados, indicando uma concentração de valores na parte direita do gráfico. Na prática, isso corresponde a dizer que a maior parte dos clientes que entrou em situação de *default* não respondeu de maneira eficaz aos procedimentos de recuperação de crédito desta IF. Dessa forma, representaram perdas significativas diante dos respectivos valores de dívida quando da caracterização do inadimplemento.

Observa-se também que a distribuição destes dados, por causa da grande concentração de valores nas proximidades da LGD igual a 1, é classificada, quanto à curtose, como leptocúrtica, pois se apresenta de forma afilada, com pouco achatamento. Além disso, os valores de ambos conjuntos de dados se apresentam no intervalo $[0, 1]$.

A tabela 5.2 apresenta as estimativas dos parâmetros para os conjuntos de dados reais, realizadas por meio da aplicação do algoritmo EM. Logo depois é mostrada a utilização dessas estimativas como insumo para a construção da curva de densidade estimada, feita juntamente com o histograma dos dados originais (figura 5.1), e da curva de distribuição acumulada estimada plotada juntamente com a curva de distribuição acumulada empírica dos dados (figura 5.2). A partir daí foi realizado também o teste de Kolmogorov-Smirnov para avaliação da hipótese de que ambas as curvas de distribuição acumulada pertenciam à mesma distribuição de probabilidade (tabela 5.3).

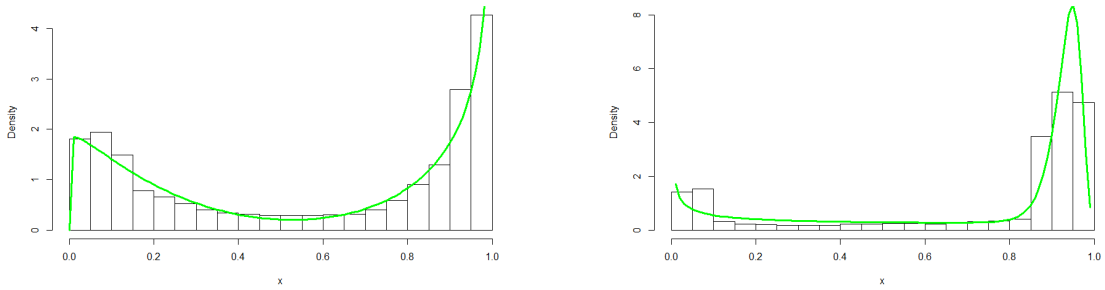
A partir dos gráficos e dos valores observados pela aplicação do teste KS, na tabela 5.2, é possível observar o comportamento dos dados de perda, e verificar que o ajuste das estimativas por meio das curvas estimadas em contraste com as curvas empíricas dos dados foi satisfatório, ou seja, a hipótese nula de igualdade das distribuições para os dados e a curva estimada não foi rejeitada.

Tabela 5.2: Estimativas dos parâmetros dos conjuntos de dados reais

Prod.	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}_1$	$\hat{b}_2$	$\hat{p}$
1	1.04349	5.90483	5.27797	0.65380	0.42395
2	32.148	0.506	5.020	0.780	0.637

Tabela 5.3: Resultados da aplicação do teste KS com o p-valor original, a respectiva decisão, o p-valor* (empírico) e a respectiva decisão quanto ao ajuste nas curvas de distribuição acumulada

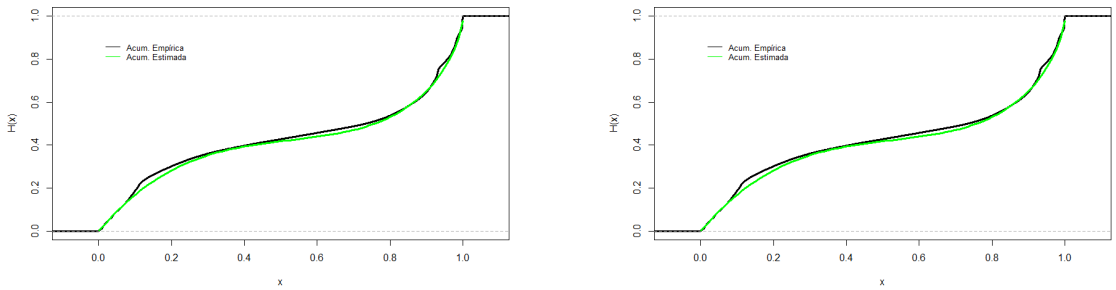
Prod.	KS	p-Valor	Decisão	p-Valor*	Decisão*
1	0.0367	< 0.001	Rejeita H_0	0.627	Não Rej. H_0
2	0.0819	< 0.001	Rejeita H_0	0.523	Não Rej. H_0



(a) Histograma e curva de densidade estimada

(b) Histograma e curva de densidade estimada

Figura 5.1: Histogramas dos dados reais e densidades estimadas.

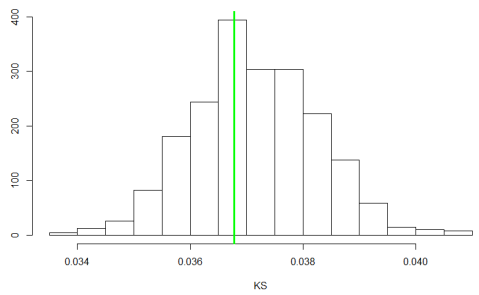


(a) Distribuição acumulada empírica e distribuição acumulada estimada do conjunto 1 de dados reais

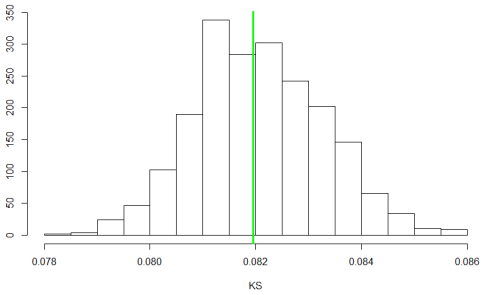
(b) Distribuição acumulada empírica e distribuição acumulada estimada do conjunto 2 de dados reais

Figura 5.2: Distribuição acumulada empírica e distribuição acumulada estimada para os dois conjuntos de dados

O gráfico da figura 5.3 permite verificar o comportamento da estatística KS para cada conjunto de dados.



(a) Distribuição do KS para o conjunto 1 de dados reais



(b) Distribuição do KS para o conjunto 2 de dados reais

Figura 5.3: Distribuição do KS para os dois conjuntos de dados

5.2 Aplicação num Cenário de Monitoramento

A aplicação dessa metodologia no contexto de monitoramento de modelos de parâmetros de risco de crédito se dá de forma acessória na tomada de decisão por meio da aferição de distâncias entre a distribuição dos dados padrão (base de dados utilizada na construção do modelo) e a distribuição dos dados nos meses seguintes, admitindo uma escala de classificação do modelo de acordo com os resultados das métricas avaliadas. No caso deste trabalho, a métrica proposta para avaliação do modelo ao longo dos meses após sua implementação é o teste de Kolmogorov-Smirnov.

Trata-se de uma avaliação de estabilidade populacional, pois uma das premissas dos modelos é que sua população não se altere, em termos de características, ao longo do tempo. Segundo [Sicsú, 2010], quando detectadas alterações na estabilidade populacional de um modelo, ações corretivas devem ser delineadas e postas rapidamente em prática para ajustar o modelo às novas condições da população.

Capítulo 6

Conclusão

Após o desenvolvimento do modelo de mistura de distribuições *Kumaraswamy*, a simulação de amostras para a validação do processo de estimação dos seus parâmetros e sua aplicação em dados reais de perda, verificou-se que o objetivo inicial, de utilizar tal modelo para modelar o comportamento de dados referentes à perda, no caso de modelos de LGD, teve resultado satisfatório, balizado pela análise gráfica e, principalmente, pela análise do teste de Kolmogorov-Smirnov, que se resume na proposta de técnica a ser implementada no processo de monitoramento dos modelos de parâmetros de risco de crédito de LGD.

Na parte referente à simulação dos dados e estimação dos parâmetros, o modelo apresentou bons resultados, e isso pôde ser visto por meio da análise do viés e do erro quadrático médio associado às estimativas, bem como por meio do teste KS, que corroborou a hipótese nula de que os dados simulados e a curva estimada pertenciam à mesma distribuição.

Em seguida, na aplicação do modelo em dados reais de perda, também foi possível observar resultados consistentes no sentido de o modelo conseguir se ajustar de forma satisfatória aos dados reais.

Assim, a proposta inicial de modelar o comportamento dos dados de LGD de uma forma nova, por meio de uma mistura de distribuições *Kumaraswamy* e utilizar tal distribuição para a propositura de um método de auxílio no processo de monitoramento de modelos de LGD, utilizando o KS, foi cumprida.

Entretanto, o processo de monitoramento de modelos de parâmetros de risco de crédito, em especial os modelos referentes ao parâmetro LGD, ainda está em aberto, podendo receber diversas contribuições e/ou adaptações de outros testes já utilizados em outros tipos de modelagem.

Para proposta de trabalhos futuros, poderia ser levantada a questão de uma avaliação quanto às variáveis utilizadas na modelagem da LGD, pois também é necessário verificar a estabilidade destas em consonância com a estabilidade da distribuição final dos dados, mostrada aqui por meio do teste KS.

Bibliografia

[Ahmad, 1988] Ahmad, Khalaf E. 1988. Identifiability of finite mixtures using a new transform. *Annals of the Institute of Statistical Mathematics*, **40**(2), 261–265.

[Atienza *et al.* , 2006] Atienza, N, Garcia-Heras, J, & Munoz-Pichardo, JM. 2006. A new condition for identifiability of finite mixture distributions. *Metrika*, **63**(2), 215–221.

[Barndorff-Nielsen, 1965] Barndorff-Nielsen, O. 1965. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, **12**(1), 115–121.

[Bennett *et al.* , 2005] Bennett, Rosalind L, Catarineu, Eva, & Moral, Gregorio. 2005. Loss Given Default Validation. *Basel Committee on Banking Supervision, Studies on the Validation of Internal Rating Systems, Working Paper*, 60–93.

[Boyles, 1983] Boyles, Russell A. 1983. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47–50.

[Calabrese, 2014] Calabrese, Raffaella. 2014. Downturn Loss Given Default: Mixture distribution estimation. *European Journal of Operational Research*, **237**(1), 271–277.

[Casella & Berger, 2011] Casella, G, & Berger, RL. 2011. *Inferência estatística-tradução da 2ª edição norteamericana.*

[Chandra, 1977] Chandra, Satish. 1977. On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, 105–112.

[Dempster *et al.* , 1977] Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

[Hartley, 1958] Hartley, HO. 1958. Maximum likelihood estimation from incomplete data. *Biometrics*, **14**(2), 174–194.

[Horta, 2009] Horta, Michelle Matos. 2009. *Modelos de mistura de distribuições na segmentação de imagens SAR polarimétricas multi-look.* Ph.D. thesis, Universidade de São Paulo.

[Jacobs Jr & Karagozoglu, 2007] Jacobs Jr, Michael, & Karagozoglu, A. 2007. *Understanding and predicting ultimate loss-given-default on bonds and loans*. Tech. rept. Working Paper.

[Jones, 2009] Jones, MC. 2009. Kumaraswamyâs distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, **6**(1), 70–81.

[Kumaraswamy, 1980] Kumaraswamy, Ponnambalam. 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46**(1), 79–88.

[Mathai *et al.* , 2009] Mathai, Arakaparampil M, Saxena, Ram Kishore, & Haubold, Hans J. 2009. *The H-function: theory and applications*. Springer Science & Business Media.

[McLachlan & Peel, 2000] McLachlan, Geoffrey, & Peel, David. 2000. *Finite mixture models*. John Wiley & Sons.

[Miu & Ozdemir, 2006] Miu, Peter, & Ozdemir, Bogie. 2006. Basel requirement of downturn LGD: Modeling and estimating PD and LGD correlations. *Journal of Credit Risk*, **2**(2), 43–68.

[Otiniano & Teixeira, 2014] Otiniano, Cira Etheowalda Guevara, & Teixeira, ECM. 2014. Estimação dos parâmetros da mistura de duas componentes GEV via Algoritmo EM. *TEMA (São Carlos)*, **15**(1), 59–71.

[Sicsú, 2010] Sicsú, Abraham Laredo. 2010. *Credit Scoring: desenvolvimento, implantação, acompanhamento*. Blucher.

[Teicher, 1963] Teicher, Henry. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 1265–1269.

[Wu, 1983] Wu, CF Jeff. 1983. On the convergence properties of the EM algorithm. *The Annals of statistics*, 95–103.

Apêndice A

Programação em R

Mistura de Distribuições *Kumaraswamy*

```
1
# Quantidade de amostras
3 m <- 100

5 # Inicializando a matriz q vai armazenar os valores dos parâmetros para cada amostra
  res <- matrix(rep(NA,m*7),m,7)
7
# Loop das amostras
9 for (k in 1:m) {
  ## Tamanho da Amostra
11  n=?

13  ## Parâmetros
  a <- c(?,?)
15  b <- c(?,?)
  p <- c(?,?)
17
  ## Função Acumulada Inversa para KUMARASWAMY
19  fmenos1_kuma = function(x,a,b){
    (1-(1-x)^(1/b))^(1/a)
21  }

23  ## Simulação da amostra a partir de Uniformes
  x <- numeric(n)
25  for (i in 1:n){
    u1 <- runif(n)
27    u2 <- runif(n)
    if(u1[i] <  p[1]) x[i]= fmenos1_kuma(u2[i],a[1],b[1])
29    if(u1[i] >= p[1]) x[i]= fmenos1_kuma(u2[i],a[2],b[2])
  }
31
  ## Mistura de duas distribuições KUMARASWAMY (expressão em termos da densidade)
33  mkuma = function(x){
    (a[1]*b[1]*(x^(a[1]-1))*(1-x^a[1])^(b[1]-1))*p[1] +
35      (a[2]*b[2]*(x^(a[2]-1))*(1-x^a[2])^(b[2]-1))*p[2]
  }
37
  # Verificando a geração dos valores
39  hist(x,prob=T)
  y = seq(min(x),max(x),0.05)
```

```
41    lines(y,mkuma(y),type="l",col="red",lwd=2)

43    ## Estimação
      erro <- Inf
45    iter <- 0

47    while(erro>0.000001){

49      ## Passo E (Expressão decorrente da Regra de Bayes)
        fpost = function(x,j){
51        fp1 = p[j]*(a[j]*b[j]*(x^(a[j]-1))*(1-x^a[j])^(b[j]-1))
          fp2 = mkuma(x)
53        fp  = fp1/fp2
          return(fp)
55      }

57      ## Maximizar Proporção
        for (j in 1:2){
59        pi[j] = mean(fpost(x,j))
        }
61
        veros <- function(param){
63        a = c(param[1],param[2])
          b = c(param[3],param[4])
65        if ((a[1]>0)&&(a[2]>0)&&(b[1]>0)&&(b[2]>0))
            return(-sum(log((a[1]*b[1]*(x^(a[1]-1))*(1-x^a[1])^(b[1]-1))*pi[1] +
67          (a[2]*b[2]*(x^(a[2]-1))*(1-x^a[2])^(b[2]-1))*pi[2])))
          else return(-Inf)
69      }

71      vero_ant <- veros(c(a,b))

73      # Valor Inicial
        vi <- c(a,b)
75
        # Estimação com o ConstrOptim [Nelder-Mead]
77      estima <- constrOptim(vi, veros, NULL, method="Nelder-Mead",
                              ui=rbind(c(1,0,0,0),c(0,1,0,0),c(0,0,1,0),c(0,0,0,1)),
79                            ci=c(0,0,0,0),hessian=FALSE, outer.iterations=500)

81      a <- c(estima$par[1], estima$par[2])
        b <- c(estima$par[3], estima$par[4])
83      vero_atu <- veros(c(a,b))
        erro=abs(vero_atu-vero_ant)/vero_ant
85      erro=abs(erro)
        iter=iter+1
87
        res[k,] <- c(iter,a,b,pi)
89    }
  }
91
  # Fazendo a média dos parâmetros das amostras para a aferição das estimativas finais
93 par_med <- numeric(6)
  for (i in 1:6){
95   par_med[i] <- mean(res[,i+1])
  }
97
  ## Mistura de duas distribuições KUMARASWAMY (expressão em termos da densidade)
99 mkuma_est <- function(x){
```

```
  (par_med[1]*par_med[3]*(x^(par_med[1]-1))*(1-x^par_med[1])^(par_med[3]-1))*par_med[5] +
    (par_med[2]*par_med[4]*(x^(par_med[2]-1))*(1-x^par_med[2])^(par_med[4]-1))*par_med[6]
}

# Verificando o ajustamento
plot(sort(x),mkuma(sort(x)),type="p",col="black",lwd=3,
     ylab="h(x)",xlab="x") # Curva dos dados simulados
curve(mkuma_est,lwd=3,add=T,col="green") # Curva dos dados estimados
legend(.4,10,lty=c(1,1),
       c("Dados Simulados","Curva Estimada"),
       bty="n",cex=0.8,col=c("black","green"))

# Histograma e curva estimada
hist(x,prob=T, nclass=15,main="",xlab="x",ylab="h(x)")
curve(mkuma_est,lwd=3,add=T,col="green") # Curva dos dados estimados
legend(.6,3,lty=c(1,1),
       c("Curva Estimada"),
       bty="n",cex=0.8,col=c("green"))

acmkuma <- function(x){
  (1-(1-sort(x)^par_med[1])^par_med[3])*par_med[5]+
  (1-(1-sort(x)^par_med[2])^par_med[4])*par_med[6]
}

# Distribuição Acumulada
plot(ecdf(sort(x)),main="",ylab='H(x)',xlab='x',lwd=3,col='black')
curve(acmkuma,lwd=3,add=T,col="green") # Curva dos dados estimados
legend(.1,0.6,lty=c(1,1),
       c('Distribuição Acumulada Empírica',"Distribuição Acumulada Estimada"),
       bty="n",cex=0.8,col=c('black',"green"))

# Construção das tabelas com as estimativas, viés e EQM
round(par_med,3)

# EQM das estimativas
a2 <- c(?,?)
b2 <- c(?,?)
p2 <- c(?,?)

eqma1<-sum((res[,2]-a2[1])^2)/m
eqma1
eqma2<-sum((res[,3]-a2[2])^2)/m
eqma2
eqmb1<-sum((res[,4]-b2[1])^2)/m
eqmb1
eqmb2<-sum((res[,5]-b2[2])^2)/m
eqmb2
eqmp <-sum((res[,6]-p2[1])^2)/m
eqmp

# KS das curvas acumuladas
acum=ecdf(x)
y1 = sort(acum(x))
y2 = acmkuma(x)
ks.test(y1,y2)

# Viés das estimativas
estimati <- par_med[-6]
original <- c(?,?,?,?,?)
```

```
vies <- abs(estimati-original)
round(vies,5)

# Gráficos com as estimativas e a média
acmkuma <- function(x,a1,a2,b1,b2,p1,p2){
  (1-(1-sort(x)^a1)^b1)*p1+
    (1-(1-sort(x)^a2)^b2)*p2
}

plot(sort(x),acmkuma(x,res[1,2],res[1,3],res[1,4],res[1,5],res[1,6],res[1,7]),
     type='l',col="gray",xlab='x',ylab='H(x)',main='',xlim=c(0,1))
for (i in 2:100) {
  lines(sort(x),acmkuma(x,res[i,2],res[i,3],res[i,4],res[i,5],res[i,6],res[i,7]),
        type='l',col="gray")
}
lines(sort(x),y1,col='black',lwd=3,lty=2)
lines(sort(x),acmkuma(x,par_med[1],par_med[2],par_med[3],par_med[4],par_med[5],
              par_med[6]),type='l',col="green",lwd=2)
legend(.2,.7,lty=c(1,1),
       c("Dados Simulados","Curva Estimada"),
       bty="n",cex=0.8,col=c("black","green"))
```

Análise com Dados Reais

```
# Importação dos dados reais
pf<-read.table("F:\\01 - Mestrado\\00 - Final\\dados\\?.txt",h=T,dec=",")
head(pf)
class(pf$vl_lgd)
x <- pf$vl_lgd
hist(x,prob=T)
mean(x)
var(x)
median(x)
kurtosis(x)
skewness(x)
length(x)

res <- matrix(rep(NA,1*7),1,7)

## Mistura de duas distribuições KUMARASWAMY (expressão em termos da densidade)
mkuma = function(x){
  (a[1]*b[1]*(x^(a[1]-1))*(1-x^a[1])^(b[1]-1))*p[1] +
    (a[2]*b[2]*(x^(a[2]-1))*(1-x^a[2])^(b[2]-1))*p[2]
}

## Estimação - Valores inicias
erro <- Inf
iter <- 0
a <- c(?,?)
b <- c(?,?)
p <- c(?,?)

while(erro>0.000001){

  ## Passo E (Expressão decorrente da Regra de Bayes)
  fpost = function(x,j){
    fp1 = p[j]*(a[j]*b[j]*(x^(a[j]-1))*(1-x^a[j])^(b[j]-1))
    fp2 = mkuma(x)
    fp  = fp1/fp2
```

```
    return(fp)
  }

  ## Maximizar Proporção
  for (j in 1:2){
    pi[j] = mean(fpost(x,j))
  }

  veros <- function(param){
    a = c(param[1],param[2])
    b = c(param[3],param[4])
    if ((a[1]>0)&&(a[2]>0)&&(b[1]>0)&&(b[2]>0))
      return(-sum(log((a[1]*b[1]*(x^(a[1]-1))*(1-x^a[1])^(b[1]-1))*pi[1] +
                      (a[2]*b[2]*(x^(a[2]-1))*(1-x^a[2])^(b[2]-1))*pi[2])))
    else return(-Inf)
  }

  vero_ant <- veros(c(a,b))

  # Valor Inicial
  vi <- c(a,b)

  # Estimação com o ConstrOptim [Nelder-Mead]
  estima <- constrOptim(vi, veros, NULL, method="Nelder-Mead",
                        ui=rbind(c(1,0,0,0),c(0,1,0,0),c(0,0,1,0),c(0,0,0,1)),
                        ci=c(0,0,0,0),hessian=FALSE, outer.iterations=500)

  a <- c(estima$par[1], estima$par[2])
  b <- c(estima$par[3], estima$par[4])
  vero_atu <- veros(c(a,b))
  erro=abs(vero_atu-vero_ant)/vero_ant
  erro=abs(erro)
  iter=iter+1

  res[1,] <- c(iter,a,b,pi)
}

round(res[1,],5)

acmkuma <- function(x){
  (1-(1-sort(x)^res[1,2])^res[1,4])*res[1,6]+
    (1-(1-sort(x)^res[1,3])^res[1,5])*res[1,7]
}

# Distribuição Acumulada
plot(ecdf(sort(x)),main="",ylab='H(x)',xlab='x',lwd=3,col='black')
curve(acmkuma,lwd=3,add=T,col="green") # Curva dos dados estimados
legend(.1,0.8,lty=c(1,1),
       c('Distribuição Acumulada Empírica',"Distribuição Acumulada Estimada"),
       bty="n",cex=0.8,col=c('black',"green"))

# KS das curvas acumuladas
acum=ecdf(x)
y1 = sort(acum(x))
y2 = acmkuma(x)
ksprim <- ks.test(y1,y2)$statistic

ks <- NULL
for (i in 1:1000){
```

```
  x1 <- sample(x,replace=TRUE)
  acum=ecdf(x1)
  y1 = sort(acum(x1))
  y2 = acmkuma(x1)
  ks[i] <- ks.test(y1,y2)$statistic
}

# p-valor empírico
pvalor <- length(ks[ks>=ksprim])/1000
pvalor

# Histograma dos p-valores
hist(ks,prob=TRUE,main='',xlab="KS",ylab="")
abline(v=ksprim,lwd=3,col="green")

plot(ecdf(x),main="",ylab='H(x)',xlab='x',lwd=4,col='black')
curve(acmkuma,lwd=4,add=T,col="green") # Curva dos dados estimados
legend(.1,0.8,lty=c(1,1),
       c('Distribuição Acumulada Empírica',"Distribuição Acumulada Estimada"),
       bty="n",cex=0.8,col=c('black',"green"))

mkuma_est <- function(x){
  (res[1,2]*res[1,4]*(x^(res[1,2]-1))*(1-x^res[1,2])^(res[1,4]-1))*res[1,6] +
    (res[1,3]*res[1,5]*(x^(res[1,3]-1))*(1-x^res[1,3])^(res[1,5]-1))*res[1,7]
}

# Verificando o ajustamento
plot(sort(x),mkuma(sort(x)),type="p",col="black",lwd=3,
     ylab="h(x)",xlab="x") # Curva dos dados simulados
curve(mkuma_est,lwd=3,add=T,col="green") # Curva dos dados estimados
legend(.4,6,lty=c(1,1),
       c("Dados Simulados","Curva Estimada"),
       bty="n",cex=0.8,col=c("black","green"))
```