UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

CIÊNCIAS ATUARIAIS

ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE DISTRIBUIÇÕES NORMAIS

Juliana Tavares

Rio de Janeiro
2015

Juliana Tavares

ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE DISTRIBUIÇÕES NORMAIS

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

Orientador: Prof. Carlos A. Abanto Valle

Co-Orientadora: Profª. Mariane Branco Alves

Rio de Janeiro
janeiro 2015

Juliana Tavares

ESTIMAÇÃO BAYESIANA EM MODELOS DE MISTURA DE DISTRIBUIÇÕES NORMAIS

Projeto de Graduação apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências Atuariais.

Examinado por:

Profª. Mariane Branco Alves

Profª. Flavia Maria Pinto Ferreira Landim

Profª. Thaís Cristina Oliveira da Fonseca

Rio de Janeiro
Janeiro 2015

Tavares, Juliana

Estimação Bayesiana em Modelos de Mistura de Distribuições Normais | Juliana Tavares – Rio de Janeiro; UFRJ| Instituto de Matemática, 2015.

IX, 66 p.: il.; 29,7cm

Orientador: Prof. Carlos A. Abanto Valle

Co-Orientadora: Mariane Branco Alves

Projeto de Graduação – UFRJ| IM| Ciências Atuariais, 2015

Referências Bibliográficas: p. 66.

1. Misturas de distribuições Normais 2. Simulação estocástica 3. Estimação bayesiana 4. Amostrador de Gibbs. I. Alves, Mariane Branco. II. Universidade Federal do Rio de Janeiro, UFRJ, Ciências Atuariais. III. Título.

DEDICATÓRIA

Dedico esta monografia e tudo que ela representa a todos os amigos e familiares que me trouxeram até ela. Em especial, dedico esta conquista aos meus pais Jorge e Wany, à minha irmã Nathália e ao meu namorado Rafael.

AGRADECIMENTOS

Agradeço a Deus pela força recebida nos momentos de dúvida, aos meus pais e irmã pela paciência tão necessária nos dias difíceis e ao meu namorado por me acompanhar nesta nova etapa.

“Não se esqueçam: a coisa que desejais, não somente já é vossa no invisível, como já partiu do coração de Deus para vós.”

Alziro Zarur

RESUMO

Misturas de distribuições são usualmente utilizadas para modelagem de dados em que as observações podem ser provenientes de diferentes grupos populacionais. Desta forma, através de uma mistura de k densidades Normais com médias e variâncias distintas, assumindo-se que o valor k é conhecido, objetivou-se modelar a variável sinistro agregado, presente em um banco de dados constituído por 1.500 pagamentos de indenizações referentes ao Seguro de Responsabilidade Civil Geral americano em dólar.

A teoria que norteou o presente estudo foi a teoria Bayesiana, e, através dela, pôde-se utilizar como ferramenta de estimação estatística dos parâmetros, o Amostrador de Gibbs, que consiste, basicamente, em uma simulação estocástica via Cadeias de Markov usualmente aplicada quando a variável de interesse apresenta uma estrutura complexa ou quando trabalha-se com problemas que têm muitas dimensões.

Ao longo desta monografia encontram-se: um pequeno resumo da teoria de mistura de distribuições e a questão da identificabilidade de misturas de distribuições; os principais conceitos, definições e metodologias necessárias para a estimação do conjunto de parâmetros; a teoria da Inferência Bayesiana; simulação estocásticas via Cadeias de Markov; e o método de Monte Carlo via Cadeias de Markov.

Conclui-se com algumas sugestões para pesquisas futuras.

Palavras-chave: Misturas de distribuições Normais, Simulação estocástica, Estimação bayesiana, Amostrador de Gibbs.

ABSTRACT

Mixture distributions are normally used for modeling data when the observations can be collected from different populations. So, by using a mixture of k Normal densities with distinct means an variances and by assuming a known value for k, it has been aimed up to model the random variable 'total claim amount' presented in a database of 1.500 claim payments of an American General Liability insurance coverage.

The theory behind the present study was the Bayesian Theory, and through this theory it was possible to use the Gibbs Sampling as a tool for statistical estimation of the parameters. The Gibbs Sampling consists basically in a stochastic simulation through Markov Chains usually applied when the variable of interest presents a complex structure or problems with many dimensions are considered.

Throughout this text can be found a short summary of the mixture distribution's theory, the issue of identificability of mixture distributions, the main concepts, definitions and methodologies needed for the parameter estimation, the Bayesian Inference theory, Markov Chain stochastic simulations and the Monte Carlo Markov Chain (MCMC) method. Some suggestions for further studies are made in the end of the text.

Keyword: Mixture of Normal Distributions, Stochastic simulation, Bayesian Estimation, Gibbs Sampling.

LISTA DE FIGURAS

SUMÁRIO

1. Introdução

Em um processo de tarifação em seguros, uma etapa preliminar e de fundamental importância é a modelagem da distribuição da variável aleatória “valor total das indenizações ocorridas em uma carteira de seguros”, também conhecida como “sinistro agregado”, em um determinado período de tempo.

Através desta variável alguns tipos de prêmios, tal como o Prêmio de Risco, podem ser precificados. Define-se Prêmio de Risco como: $P = E[S]$, em que *S* é a variável aleatória “valor total das indenizações ocorridas em uma carteira de seguros”.

Indo mais além, dentre os princípios de cálculo de prêmio, ou seja, dentre as funções $f: S \rightarrow P$ que associam a cada distribuição de sinistros agregada *S* um número real *P*, destaca-se o Princípio do Percentil que, apesar de ser considerado um dos melhores princípios, por permitir à seguradora dimensionar eficazmente o risco que ela assume, segundo FERREIRA (2010), “*nem sempre ele é utilizado em função da impossibilidade/dificuldade de se calcular a função de distribuição acumulada do sinistro agregado”*.

Percebe-se no meio atuarial a importância da variável sinistro agregado, e esta foi a principal motivação deste estudo, que objetiva modelar a variável sinistro agregado por meio de um banco de dados constituído por 1.500 pagamentos de indenizações referentes ao Seguro de Responsabilidade Civil Geral americano em dólar, que protege o segurado de eventuais reclamações ou ações na justiça em que este seja responsabilizado civilmente por ter causado danos involuntários a outras pessoas, sejam materiais ou corporais.

Devido à característica dos dados, em que as observações podem ser provenientes de diferentes grupos populacionais, optou-se pela modelagem através

de uma mistura de k densidades Normais com médias e variâncias distintas para cada um dos k subgrupos da população, assumindo-se que o valor k é conhecido.

A teoria que norteou a presente monografia foi a teoria Bayesiana, e, através dela, pôde-se utilizar como ferramenta de estimação estatística dos parâmetros, o Amostrador de Gibbs, que consiste, basicamente, em uma simulação estocástica via Cadeias de Markov usualmente aplicada quando a variável de interesse apresenta uma estrutura complexa ou quando trabalha-se com problemas que têm muitas dimensões.

Além deste primeiro capítulo introdutório, esta monografia está estruturada em outros 4 capítulos. No capítulo 2 expõe-se um pequeno resumo da teoria de mistura de distribuições e a questão da identificabilidade de misturas de distribuições é apresentada devido à sua importância no contexto da teoria da Inferência Bayesiana.

Composto por três subtítulos, o capítulo 3 abrange os principais conceitos, definições e metodologias necessárias para a estimação do conjunto de parâmetros. Inicialmente, aborda-se a teoria da Inferência Bayesiana; posteriormente, trata-se da simulação estocásticas via Cadeias de Markov; e, encerrando o capítulo, o método de Monte Carlo via Cadeias de Markov é descrito.

No capítulo 4 o leitor encontrará a descrição do banco de dados utilizado e os resultados após a modelagem dos dados através de uma mistura de 3 densidades Normais cujas médias, pesos e variâncias eram desconhecidos.

Encerra-se este estudo no seu quinto capítulo com algumas considerações finais e propostas para futuros trabalhos.

2. Teoria de Misturas

O uso de mistura de distribuições é uma forma conveniente e flexível para modelagem de dados em que as observações podem ser provenientes de diferentes grupos populacionais.

Mais formalmente, considere uma variável aleatória Y (ou vetor aleatório) de interesse presente em uma população que é formada por k subgrupos, cada qual com a proporção ω_1, ω_2, ... , ω_k em relação ao total.

Assuma que a variável Y seja identicamente distribuída dentro de cada um dos k subgrupos, e heterogeneamente distribuída entre estes.

Admite-se, geralmente, que as distribuições de probabilidade de Y em cada um dos grupos vêm de uma mesma família paramétrica $p(y|\boldsymbol{\theta})$, distinguindo-se apenas pelo vetor paramétrico $\boldsymbol{\theta}$. Além disso, tais grupos podem ser rotulados por uma variável discreta indicadora S, a qual assume valores no conjunto $\{1, \ldots, \mathrm{k}\}$.

Amostrando-se aleatoriamente da população, pode-se registrar não só o valor da variável de interesse Y como também o indicador do grupo S, que tem probabilidade ω_s de ser amostrado. Caso isso seja verdade, condicionado ao conhecimento de S, Y é uma variável aleatória que segue a distribuição $p(y|\boldsymbol{\theta}_S)$, em que $\boldsymbol{\theta}_S$ são os parâmetros dentro do grupo S. A densidade conjunta $p(y,S)$ é então dada por

$$p(y,S) = p(y|S).p(S) = p(y|\boldsymbol{\theta}_s).\omega_s. \qquad (2.1)$$

No entanto, a variável indicadora S, por vezes chamada de variável latente, tipicamente não é observável, restando apenas a observação da variável Y. Neste caso, uma mistura finita de distribuições surge através da densidade marginal $p(y)$:

$$p(y) = \sum_{S=1}^{K} p(y,S) = w_1.p(y|\boldsymbol{\theta}_1) + \cdots + w_k.p(y|\boldsymbol{\theta}_k) \qquad (2.2)$$

Em uma amostra, geralmente, ocorre heterogeneidade, que pode ser caracterizada pela diferença dos valores assumidos pela média da variável Y entres os diferentes grupos de sujeitos. Neste caso, uma maneira comum de lidar com a modelagem dos dados é assumindo que Y segue uma distribuição Normal e que cada um dos k grupos apresenta média μ_i^S.

Supondo que o desvio de μ_i^S em relação à média comum μ possa ser especificado por um fator observável z_i, que assume valores no conjunto $\{1, \dots, k\}$, consegue-se modelar μ_i^S da seguinte forma:

$$\mu_i^S = \beta_0 + \beta_1 . z_i \tag{2.3}$$

Da mesma forma, os parâmetros desconhecidos podem ser estimados pela regressão do tipo:

$$Y = \beta_0 + \beta_1 . z_i + \varepsilon_i \quad \text{, com } \varepsilon_i \sim N(0, \sigma^2) \tag{2.4}$$

Neste caso, Y assume uma distribuição $N(\mu_i^S, \sigma^2)$, com $\mu_i^S = \beta_0 + \beta_1 . z_i$.

No entanto, assim como explicitado acima, no caso da variável indicadora S, o fator z_i é não observável e, para tanto, tem-se que lidar com a heterogeneidade não observável. Na população, ω_1, ω_2, ... , ω_k denotam a distribuição discreta de z_i.

Se os dados são amostras aleatórias da população sob investigação, então a distribuição marginal da variável de interesse Y, dada por $Y = \beta_0 + \beta_1 . z_i + \varepsilon_i$, será uma mistura finita de k distribuições Normais com variâncias iguais.

$$Y \sim \omega_1 . N(\mu_1, \sigma^2) + \cdots + \omega_k . N(\mu_2, \sigma^2) \tag{2.5}$$

Estendendo o modelo acima para contemplar tanto heterogeneidade na média quanto na escala, na presente monografia o estudo está pautado na aplicação de misturas de k densidades Normais com médias e variâncias distintas, a fim de modelar uma variável de interesse Y em uma população que se admite ser heterogênea, em que o número k de componentes na mistura é conhecido.

$$f(y|\psi) = \sum_{j=1}^{K} \omega_j . N\left(\mu_j, \sigma_j^2\right) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2) + \ \ldots + \omega_k . N(\mu_k, \sigma_k^2)$$

Denota-se, na presente pesquisa, $\boldsymbol{\theta}$ como todos os distintos parâmetros presentes em cada densidade da componente $\boldsymbol{\theta_i} = (\omega_i,\ \mu_i,\ \sigma_i^2)$, e $\boldsymbol{\psi}$ como a coleção completa de todos os distintos parâmetros que aparecem no modelo de misturas $\psi = (\omega_1,\ \omega_2, \ldots, \omega_k,\ \mu_1,\ \mu_2, \ldots,\ \mu_k,\ \sigma_1^2,\ \sigma_2^2,\ \ldots,\ \sigma_k^2)$.

Modelos estatísticos baseados em misturas de distribuições apresentam características que exigem um tratamento matemático rigoroso. Isto posto, antes de se encerrar o presente capítulo, faz-se necessário abordar a questão da identificabilidade de misturas de distribuições, fundamental no contexto da teoria que norteia a presente pesquisa: Inferência Bayesiana, que será abordada no capítulo 3.

2.1. Identificabilidade

No tratamento bayesiano de distribuições de mistura, nem sempre as distribuições a posteriori e as distribuições conjuntas a posteriori levam a respostas coerentes. Em algumas circunstâncias a mistura pode ser não-identificável.

Trata-se, portanto, neste item, de um importante conceito associado a modelos de misturas: a identificabilidade; essencial para a estimação dos parâmetros de interesse.

Em geral, uma família de distribuições paramétricas é dita ser identificável se parâmetros de valores distintos determinam membros distintos desta família.

Formalmente, uma família paramétrica de distribuições, indexada pelos parâmetros $\boldsymbol{\psi}$ pertencente a um espaço paramétrico $\boldsymbol{\Theta}$, que é definida sobre o espaço de estados Υ, é dita ser identificável se, dados quaisquer dois valores para

parâmetros no espaço paramétrico $\boldsymbol{\psi}$ e $\boldsymbol{\psi}^*$ em $\boldsymbol{\Theta}$, se estes definem a mesma lei de probabilidade em Υ, então, obrigatoriamente, $\boldsymbol{\psi}$ e $\boldsymbol{\psi}^*$ devem ser idênticos.

Em relação à correspondência de densidades de probabilidades, para todo $y \in \Upsilon$, os parâmetros $\boldsymbol{\psi}$ e $\boldsymbol{\psi}^*$ também precisam ser idênticos:

$$p(y \mid \boldsymbol{\psi}) = p(y \mid \boldsymbol{\psi}^*) \text{ , para todo y} \in Y \rightarrow \boldsymbol{\psi} = \boldsymbol{\psi}^* \qquad (2.6)$$

Uma distribuição Normal única, indexada por $\boldsymbol{\theta} = (\mu, \sigma^2)$ é, claramente, identificável. No entanto, para misturas de duas ou mais Normais, a questão da indentificabilidade é mais delicada.

Segundo Teicher (1963), exceto para misturas de distribuições Uniformes, muitas outras misturas finitas de densidades contínuas são identificáveis.

Vale ainda destacar que identificabilidade é uma característica inerente aos conjuntos paramétricos.

Pode-se, no contexto de mistura de distribuições, distinguir três tipos de não-identificabilidade: (i) não-identificabilidade causada pela invariância da mistura após reordenadar os componentes, também conhecido como "*label-switching*"; (ii) não-identificabilidade causada pelo potencial de superajuste que ocorre quando um dos componente é vazio ou quando dois ou mais componentes são iguais; e (iii) não-identificabilidade devido à falta de uma propriedade genérica de certas classes de misturas de distribuição.

O primeiro tipo de não-identificabilidade é causado por uma invariância de uma mistura de distribuições após a reordenação de seus componentes.

Como exemplificado por FRUHWIRTH-SCHNATTER (2006), seja $p(y|\boldsymbol{\psi})$ uma densidade de probabilidade obtida através de uma mistura de duas distribuições Normais:

$$p(y|\boldsymbol{\psi}) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2) \text{ ,}$$

com $\boldsymbol{\psi} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \omega_1, \omega_2)$, e seja $\boldsymbol{\psi}*$ um parâmetro arbitrário que é obtido permutando a ordem dos componentes: $\boldsymbol{\psi}* = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_1, \omega_2, \omega_1)$. Verifica-se que as distribuições $p(y|\boldsymbol{\psi})$ e $p(y|\boldsymbol{\psi}*)$ são a mesma, embora estes parâmetros sejam distintos.

$$p(y|\boldsymbol{\psi}) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2)$$
$$= \omega_2 . N(\mu_2, \sigma_2^2) + \omega_1 . N(\mu_1, \sigma_1^2) = p(y|\boldsymbol{\psi}*) \quad (2.7)$$

Portanto, em uma mistura finita de distribuições com *k* componentes existem *k!* maneiras de se reordenar os componentes, cada uma delas como uma permutação dos parâmetros.

Devido a esta invariância, uma mistura de distribuições se torna não-identificável.

No entanto, segundo FRUHWIRTH-SCHNATTER (2006), *"este não é um problema de identificabilidade grave, pois todos os parâmetros são relacionados entre si e, de fato, só diferem pela forma em que os componentes são arranjados."*

A segunda causa da não-identificabilidade é devido ao potencial de superajuste proveniente da existência de componentes vazios ou iguais na mistura.

Considerando a mesma mistura de duas distribuições Normais acima, esta pode ser escrita como uma mistura de três distribuições Normais ao adicionar uma terceira componente com peso $\omega_3 = 0$:

$$p(y|\boldsymbol{\psi_3}) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2) + 0 . N(\mu_3, \sigma_3^2) \quad (2.8)$$

Neste caso, o parâmetro $\boldsymbol{\psi_3}$ corresponde a uma mistura que se encontra em um conjunto não-identificável, pois a densidade $p(y|\boldsymbol{\psi_3})$ é a mesma para quaisquer valores de μ_3 e σ_3^2.

Ou ainda, pode-se da mistura de duas distribuições Normais acima, gerar uma mistura de três distribuições Normais ao se dividir um dos componentes da mistura em outras duas distribuições Normais:

$$p(y|\boldsymbol{\psi_3}) = \omega_1.N(\mu_1, \sigma_1^2) + (\omega_2 - \omega_3).N(\mu_2, \sigma_2^2) + \omega_3.N(\mu_2, \sigma_2^2) \qquad (2.9)$$

Novamente, o parâmetro $\boldsymbol{\psi_3}$ se encontra em um conjunto não-identificável, pois a densidade $p(y|\boldsymbol{\psi_3})$ é a mesma para quaisquer valores de ω_3 com $0 < \omega_3 < \omega_2$.

Segundo LIU (2010), "*quando o número de componentes k é grande, alguns dos pesos podem se tornar tão próximos de 0 que o modelo de misturas é quase não-identificável; ou se dois componentes são bem próximos um do outro, a densidade de mistura de k componentes pode ser empiricamente indistinguível de uma mistura com menos do que k componentes"*.

Para que as duas situações acima, de invariância da mistura e de superajuste, não aconteçam, faz-se necessário restringir o espaço paramétrico **Θ** a fim de que a densidade $p(y|\boldsymbol{\psi})$ seja uma mistura de k componentes distintos e não-vazios.

A primeira restrição que se pode impor é a de que os pesos sejam positivos. Isto evita a não-identificabilidade causada pelos componentes vazios.

$$\omega_k > 0\,, \quad \forall\, k = 1, 2, \ldots, K$$

Depois, uma condição de desigualdade dos parâmetros dos componentes evita a não-identificabilidade devido à igualdade de alguns componentes:

$$\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k`}\,, \forall\, k \neq k`\,, \qquad k, k` = 1, 2, \ldots, K$$

No entanto, se pode aplicar uma desigualdade mais fraca que a anterior, requerendo para quaisquer dois parâmetros $\boldsymbol{\theta}_k \neq \boldsymbol{\theta}_{k`}$, que estes difiram em pelo menos um elemento que não precisa ser o mesmo para todos os componentes. Mais formalmente $\forall\, k \neq k`\,,\ k, k` = 1, 2, \ldots, K$:

$$\exists\, j\ (k, k`) \in \{1, \ldots, d\} : \boldsymbol{\theta}_{k,j}(k, k`) \neq \boldsymbol{\theta}_{l,j}(k, k`)\,,$$

em que *d* é o número de elementos do vetor paramétrico $\boldsymbol{\theta}_k$.

Estas duas últimas condições para a identificabilidade garantem uma rotulagem única, pois impõem uma restrição de ordem sobre qualquer um dos d elementos $\boldsymbol{\theta}_{k,j}$, $j = 1, \dots, d$:

$$\boldsymbol{\theta}_{1,j} < \cdots < \boldsymbol{\theta}_{k,j}$$

Sobre esta última desigualdade, a invariância ao se reordenar os k componentes da mistura desaparece e não existe mais o conjunto não-identificável.

Porém, mesmo atendidas as restrições acima, o conjunto paramétrico ainda pode ser não-identificável como exemplificado nos dois casos a seguir encontrados em FRUHWIRTH-SCHNATTER (2006).

Caso 1) Considere uma mistura de três distribuições Normais, em que $\mu_1 = \mu_2$, ${\sigma_1}^2 \neq {\sigma_2}^2$, $\mu_2 \neq \mu_3$ e ${\sigma_2}^2 = {\sigma_3}^2$, ou seja, nenhum dos dois parâmetros são diferentes em todos os componentes e uma restrição de ordem sob um único elemento corresponde a $k - 1$ desigualdades.

Pode-se então substituir algumas dessas desigualdades por uma restrição em um elemento diferente em $\boldsymbol{\theta}$. Neste caso, uma das desigualdades pode envolver (μ_2, μ_3) e a outra $({\sigma_1}^2, {\sigma_2}^2)$ para descrever as diferenças entre os parâmetros dos componentes.

Percebe-se que a identificação da validade da restrição em dimensões maiores se torna um desafio.

Caso 2) Seja $0 < \omega_1 < \cdots < \omega_k$ a restrição formal assumida para a identificabilidade em uma mistura de k distribuições. Esta condição exclui os componentes vazios e induz a uma rotulagem única, porém, não exclui a não-identificabilidade devido ao potencial de parâmetros iguais.

Estes dois casos mostram que misturas de distribuições finitas podem permanecer não-identificáveis mesmo que as restrições formais de identificabilidade

sejam aplicadas. A imposição de uma restrição formal genérica que exclua qualquer um dos problemas de identificabilidade exemplificados acima faz-se necessária.

Fruhwirth-Schnatter (2006) afirma que *uma família de uma mistura finita de distribuições é identificável se, e somente se, os membros $T(\boldsymbol{\theta})$ da família de distribuições subjacentes são linearmente independentes sobre o campo dos números reais*; o que pode ser verificado mostrando que algumas transformações $G(z;\boldsymbol{\theta})$ de $T(\boldsymbol{\theta})$, tal como a função característica ou a função geradora de momentos, são linearmente independentes.

Isto posto, alguns resultados já foram apresentados na literatura estatística, tais como: uma mistura de Normais univaridas e multivariadas são genericamente identificáveis. O que pode ser observado através da seguinte citação encontrada em FRUHWIRTH-SCHNATTER (2006): "*Teicher (1963) provou que várias misturas de densidades contínuas univariadas, especialmente misturas de Normais univariadas, misturas de Exponenciais e distribuições Gama são genericamente identificáveis. Estes resultados são estendidos por Yakowitz e Spragins (1968) para várias famílias multivariadas tais como mistura de Normais multivariadas.*"

Levando-se em consideração que esta monografia tem como foco principal mistura de distribuições Normais, a partir dos resultados encontrados na literatura estatística, admite-se que seja possível a identificabilidade nos conjuntos paramétricos dos modelos propostos.

3. Estimação dos Parâmetros

Tomando-se como proposta principal deste trabalho a modelagem dos dados através de mistura de k distribuições Normais, e definidos anteriormente os conceitos relativos à mistura de distribuições, este capítulo tem como objetivo apresentar os conceitos, definições e metodologias necessárias para a estimação do conjunto de parâmetros $\boldsymbol{\psi} = (\omega_1, \omega_2, \dots, \omega_k,\ \mu_1,\ \mu_2, \dots,\ \mu_k,\ \sigma_1^2,\ \sigma_2^2,\ \dots,\ \sigma_k^2)$.

3.1. Inferência Bayesiana

A abordagem inferencial adotada neste trabalho é bayesiana. Semelhante ao método clássico ou frequentista, esta metodologia também inclui a observação dos dados y que são descritos por uma função de probabilidade ou densidade de probabilidade $f(y|\theta)$, em que θ é o parâmetro que se deve conhecer para descrever completamente o processo. Tipicamente, assume-se que as observações y_i são independentes e identicamente distribuídas.

A principal característica que diferencia os métodos bayesianos dos métodos clássicos ou frequentistas é que, além da distribuição observacional $f(y|\theta)$, métodos bayesianos consideram uma distribuição que incorpora formalmente o conhecimento que se tem sobre as componentes não observáveis, mesmo que a informação não seja precisa. Esta distribuição recebe o nome de distribuição a priori $p(\theta)$.

Em alguns casos, a distribuição a priori $p(\theta)$ pode ser especificada com a ajuda de outras constantes que enriquecem o modelo, assim como especificado acima para y. Estas constantes são chamadas de hiperparâmetros, pois nada mais são do que os parâmetros das distribuições dos parâmetros.

Desta forma, a inferência bayesiana é constituída basicamente por três

elementos: (i) a distribuição das observações que, quando considerada como função dos parâmetros é chamada de função de verossimilhança $l(\theta) = f(y|\theta)$; (ii) a densidade a priori $p(\theta)$, que contém a distribuição de θ antes das observações dos dados; e (iii) a densidade a posteriori $p(\theta|y) = \pi(\theta)$, que é a distribuição de θ depois das observações dos valores de y.

Vale destacar que, além da diferença em relação à inferência clássica ou frequentista apresentada acima, esta metodologia oferece a riqueza de uma especificação completa de uma distribuição a posteriori do parâmetro de interesse θ.

A teoria da inferência bayesiana recebe este nome, pois a densidade a posteriori de θ pode ser obtida através do Teorema de Bayes que segue abaixo:

$$p(\theta|y) = \frac{f(y|\theta).p(\theta)}{f(y)}, \tag{3.1}$$

em que

$$f(y) = \int f(y|\theta).p(\theta)d\theta. \tag{3.2}$$

Como $f(y)$ é apenas uma constante, a distribuição a posteriori de θ pode ser escrita de uma forma mais compacta:

$$\pi(\theta) \propto l(\theta).p(\theta) \tag{3.3}$$

3.1.1. Distribuições a prioris, função de verossimilhança e distribuições a posteriori do modelo

Como especificado no capítulo 2, a modelagem dos dados será feita através da aplicação de misturas de k densidades Normais com médias e variâncias distintas em uma população que admite-se ser heterogênea, com um número k de componentes conhecido.

Para a modelagem dos dados, serão utilizadas as mesmas prioris sugeridas

na dissertação de LIU (2010). Quais sejam:

$$\mu_j|\sigma_j^2 \sim N\left(\lambda_j, \frac{\sigma_j^2}{\tau_j}\right), \sigma_j^2 \sim GI(\alpha_j, \beta_j), (w_1, w_2, w_3) \sim D(\gamma_1, \gamma_2, \gamma_3),$$

com N, GI e D denotando as densidades Normal, Gama inversa e Dirichlet, respectivamente, e λ_j, τ_j, α_j e β_j representam os hiperparâmetros do modelo definidos da seguinte forma: $\lambda_j = \frac{\sum_{i=1}^{n} y_i}{n}$, $\tau_j = \frac{2{,}6}{(y_{máx} - y_{mín})^2}$, $\alpha_j = 1{,}28$, $\beta_j = 0{,}36 \cdot S_j^2$, com $S_j^y = \sum_{i=1}^{n} \Pi_{z_i^{(t)}=j} x_l$.

Segundo Liu, uma das vantagens da adaptação do método bayesiano está no enriquecimento do modelo através da inclusão das distribuições a priori.

Desenvolvendo analiticamente as distribuições apresentadas acima, tem-se:

$$p\left(\mu_j|\sigma_j^2\right) = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{\tau_j}{\sigma_j^2}} \cdot \exp\{-\frac{1}{2} \cdot \frac{\tau_j}{\sigma_j^2} \cdot \left(\mu_j - \lambda_j\right)^2\}$$

$$p\left(\sigma_j^2\right) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \cdot (\sigma_j^2)^{-(\alpha_j+1)} \cdot e^{-\frac{\beta_j}{\sigma_j^2}}$$

$$p(w_1, w_2, w_3|\gamma_1, \gamma_2, \gamma_3) = \frac{\Gamma(\gamma_1 + \gamma_2 + \gamma_3)}{\Gamma(\gamma_1) \cdot \Gamma(\gamma_2) \cdot \Gamma(\gamma_3)} \cdot {w_1}^{\gamma_1 - 1} \cdot {w_2}^{\gamma_2 - 1} \cdot {w_3}^{\gamma_3 - 1}$$

Não obstante, como as observações são independentes e identicamente distribuídas, a densidade conjunta observacional pode ser obtida através do produtório das densidades marginais.

Para uma amostra de tamanho *n*, a função de verossimilhança é formada pelo produtório de *n* distribuições normais como segue:

$$f(y|\boldsymbol{\psi}) = \prod_{i=1}^{n} \omega_{z_i} \cdot \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\sigma_{z_i}^2}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_{z_i}^2} \cdot (y_i - \mu_{z_i})^2\}$$

$$f(y|\boldsymbol{\psi}) = {\omega_{z_i}}^{n_j} \cdot \frac{1}{(\sqrt{2\pi})^{n_j}} \cdot \frac{1}{(\sqrt{\sigma_{z_i}^2})^{n_j}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_{z_i}^2} \cdot \sum_{i=1}^{n} I_{z_i=j}(y_i - \mu_{z_i})^2\} \quad (3.4)$$

Pela inferência bayesiana, por sua distribuição a priori e pela função de

verossimilhança, pode-se encontrar o núcleo das densidades condicionais completas a posteriori, isto é, a densidade de cada um dos parâmetros de interesse, condicional às observações e aos demais parâmetros. Tais densidades serão denotadas por $\pi(\mu_1),\ \pi(\mu_2),\dots,\pi(\mu_k),\ \pi(\sigma_1),\ \pi(\sigma_2),\dots,\pi(\sigma_k),\ \pi(\omega_1),\ \pi(\omega_2),\dots,\pi(\omega_3)$:

$$\pi(\mu_1) \propto \frac{1}{(\sqrt{2\pi})^{n_1}} \cdot \frac{1}{(\sqrt{\sigma_1^2})^{n_1}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_1^2} \cdot \sum_{i=1}^{n} I_{z_i=1}(y_i - \mu_1)^2\} \cdot \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{\tau_1}{\sigma_1^2}} \cdot \exp\{-\frac{1}{2} \cdot \frac{\tau_1}{\sigma_1^2} \cdot (\mu_1 - \lambda_1)^2\} \tag{3.5}$$

$$\pi(\sigma_1^2) \propto \frac{1}{(\sqrt{2\pi})^{n_1}} \cdot \frac{1}{(\sqrt{\sigma_1^2})^{n_1}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_1^2} \cdot \sum_{i=1}^{n} I_{z_i=1}(y_i - \mu_1)^2\} \cdot \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \cdot (\sigma_1^2)^{-(\alpha_1+1)} \cdot e^{-\frac{\beta_1}{\sigma_1^2}} \tag{3.6}$$

$$\pi(\omega_1) \propto \frac{1}{(\sqrt{2\pi})^{n_1}} \cdot \frac{1}{(\sqrt{\sigma_1^2})^{n_1}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_1^2} \cdot \sum_{i=1}^{n} I_{z_i=1}(y_i - \mu_1)^2\} \cdot \frac{\Gamma(\gamma_1+\gamma_2+\gamma_3)}{\Gamma(\gamma_1)\cdot\Gamma(\gamma_2)\cdot\Gamma(\gamma_3)} \cdot {\omega_1}^{\gamma_1-1} \cdot {\omega_2}^{\gamma_2-1} \cdot {\omega_3}^{\gamma_3-1} \tag{3.7}$$

Note-se, entretanto, que a variável z_i não é observável. Portanto, as distribuições a posteriori acima não apresentam em seu núcleo uma forma analítica fechada.

O mesmo acontecerá para a estimação dos parâmetros $\mu_2,\dots,\mu_k,\sigma_2,\dots,\sigma_k,\omega_2,\dots,\omega_k$. Este limitante obriga à utilização de métodos computacionais de aproximação capazes de estimar os parâmetros do conjunto paramétrico $\boldsymbol{\psi}$.

Destaca-se o método Monte Carlo via Cadeias de Markov, entre eles o Amostrador de Gibbs, que será explorado nos próximos itens deste capítulo e implementado para a modelagem do banco de dados, como mostrado no capítulo 4.

3.2. Simulação estocástica via Cadeias de Markov

Quando a variável não observável de interesse apresenta distribuição a posteriori com estrutura complexa ou quando trabalha-se com problemas que têm muitas dimensões, os métodos de simulação estocástica via Cadeias de Markov exprimem grande vantagem, por serem de simples compreensão, facilmente implementados e por não exigirem a utilização de softwares sofisticados.

Amplamente utilizado em diversas áreas, estes métodos têm como ideia central *simular uma cadeia de Markov de forma que esta convirja exatamente para a distribuição da qual se deseja amostrar* (ALVES, 1999).

Para a exploração do método Monte Carlo via Cadeias de Markov, alguma revisão da teoria de Processos Estocásticos, especificamente, em Cadeias de Markov, fazem-se necessárias.

O tópico seguinte objetiva revisar, de forma sumária, alguns conceitos que fundamentam as técnicas de simulação iterativas.

Vale ressaltar que, como a teoria que alicerça a simulação estocástica iterativa aplicada sobre o banco de dados neste estudo é sobre Cadeias de Markov a tempo discreto, este resumo não contemplará os espaços de estados contínuos e, tão somente, os espaços de estados discretos.

3.2.1. Cadeias de Markov

Segundo GAMERMAN E LOPES (2006), Cadeias de Markov são *um tipo especial de processo estocástico que tratam da caracterização de variáveis aleatórias.*

Formalmente, um processo estocástico a tempo discreto é uma Cadeia de Markov tal que, condicionado ao conhecimento do estado presente e dos estados passados, a distribuição do estado futuro só depende do estado presente. Em outras palavras, *uma Cadeia de Markov é um processo estocástico no qual dados o estado presente, os estados passados e futuros são independentes* (GAMERMAN E LOPES, 2006).

Analiticamente, tem-se:

$$P_{ij} = P(i,j) = P(X_{t+1} = j \mid X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}, X_t = i)$$
$$= P(X_{t+1} = j \mid X_t = i) \quad (3.8)$$

Quando a probabilidade de transição do estado i para o estado j descrita acima não depende de t, a cadeia é dita ser homogênea.

Por sua vez, um processo estocástico nada mais é do que uma coleção de quantidades aleatórias $\{X^{(t)} : t \in T\}$, para algum conjunto de índices T, com espaço de estados S.

Limitaremo-nos à revisão de teoria para o caso em que T é um conjunto enumerável e o processo estocástico de interesse tem espaço de estados discreto $S = \{x_0, x_1, x_2, \dots\}$.

3.2.1.1. Probabilidades de transição

A probabilidade de transição a um passo P_{ij} ou $P(i,j)$, isto é, a probabilidade

de que a cadeia passe do estado i para o estado j em um passo, satisfazem as seguintes propriedades:

- $P(i,j) \geq 0, \forall\, i\,, j\ \in S$
- $\sum_{j=0}^{\infty} P(i,j) = 1, \forall\, i\ \in S$

Estas probabilidades de transição, por conseguinte, podem ser expostas por uma matriz ***P*** de transições, na qual a soma de qualquer uma das linhas tem valor 1:

$$\boldsymbol{P} = \begin{bmatrix} P_{00} & P_{01} & \cdots \\ P_{10} & P_{11} & \cdots \\ \vdots & \vdots & \vdots \\ P_{n0} & P_{n1} & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Ademais, as probabilidades de transição a n passos ${P_{ij}}^{n}$ ou $P^n(i,j)$, ou seja, a probabilidade de que a cadeia passe do estado i para o estado j em, exatamente, n passos (ou transições), podem ser expressas como:

$$P_{ij}^{n} = \ P^n(i,j) = \ P\{X_{n+t} = j \mid X_t = i\} \tag{3.9}$$

Pode-se demonstrar, e é facilmente encontrado na literatura, que através da aplicação das equações de Chapman-Kolmogorov

$${P_{ij}}^{a+b} = \sum_{k=0}^{\infty} P_{kj}^{b} \,.\, P_{ik}^{a} \quad \forall a, b \geq 0, \ \ \forall i, j \in S, \tag{3.10}$$

a matriz de transição a n passos é obtida por meio da n-ésima potência da matriz de transições a 1 passo.

3.2.1.2. Espaço de estados

Uma Cadeia de Markov é composta por classes de estados. Cada um desses estados recebe uma classificação em virtude das características da matriz de transição ***P***.

Para a classificação dos estados e das classes de uma Cadeia de Markov, a

revisão de alguns conceitos e quantidades mostra-se fundamental:

(i) A probabilidade da Cadeia atingir j, estando em i, é denotada por P_{ij}.

(ii) Diz-se que um estado i se comunica com o estado j se $P_{ij} > 0$ e $P_{ji} > 0$ e denota-se por $i \leftrightarrow j$. Neste caso, eles pertencem a uma mesma classe.

(iii) O número de visitas da cadeia a um estado i qualquer é

$$N(i) = \#\left\{n > 0 : X^{(n)} = i\right\} = \sum_{n=1}^{\infty} I(X^n = i) \tag{3.11}$$

(iv) Seja *d* o máximo divisor comum do conjunto $\{n \geq 1 : P_{ii}^n > 0\}$. Diz-se que um estado i tem periodicidade *d* se, para todo *n* não divisível por *d*, $P_{ii}^n = 0$.

Um estado i é classificado como recorrente se, começando no estado i de uma Cadeia de Markov, ele retorna ao estado i com probabilidade um (P_{ii} = 1). Caso esta probabilidade seja menor que um ($P_{ii} < 1$), o estado é dito transiente.

Se uma Cadeia de Markov começar em um estado i recorrente, então o número esperado de vistitas a este estado i é infinito. Em outras palavras: *se um estado é recorrente, então uma vez que seja atingido, o processo volta a ele com certeza.* (ALVES, 1999)

Além disso, se o tempo médio esperado de retorno a este estado recorrente i for finito, classifica-se o estado como recorrente positivo. Caso contrário, ele é dito ser recorrente nulo.

No caso de um estado transiente j, o número esperado de visitas a ele é finito.

Pode-se resumir analiticamente a diferença entre estados recorrentes e transientes da seguinte forma:

- Se um estado $i \in S$ é transiente então, para todo $j \in S$:

$$P(N(i) < \infty) = 1 \text{ e } E[N(i)|X^0 = j] = \frac{P_{ji}}{1 - P_{ii}} < \infty. \tag{3.12}$$

- Se um estado $i \in S$ é recorrente, então:

$$P(N(i) = \infty) = 1 \text{ e } E[N(i)|X^0 = j] = \infty. \quad (3.13)$$

Se todos os estados de uma Cadeia se comunicam, existe apenas uma classe e, portanto, a cadeia é dita ser irredutível.

Um estado com período 1 é dito ser aperiódico. Um estado recorrente positivo e aperiódico é chamado ergódico. Se todos os estados que constituem uma cadeia são ergódicos, então a cadeia é dita ergódica. Esta classificação é crucial na análise de convergência da cadeia como mencionado abaixo no item 3.2.2.4.

3.2.1.3. Distribuição estacionária

Um problema fundamental ao se implementar uma simulação estocástica é a necessidade da existência de uma distribuição estacionária, pois, espera-se, através de um algoritmo iterativo, construir uma Cadeia de Markov que convirja para a distribuição em que se tem interesse quando $n \rightarrow \infty$.

Uma distribuição π é dita ser uma distribuição estacionária em uma cadeia se

$$\sum_{i \in S} \pi(i).P_{ij} = \pi(j), \forall\, j \in S. \quad (3.14)$$

Esta distribuição também é conhecida como distribuição invariante ou de equilíbrio.

Uma vez que a cadeia atinge um estado com a distribuição π, então todos os estados subsequentes terão também distribuição π.

3.2.1.4. Teoremas limites

Apresentada, no item anterior, a definição de distribuição estacionária, uma

vez que ela exista e que $lim_{n\to\infty} P_{ij}^{n} = \pi(j)$, então, independentemente do valor inicial da distribuição, a cadeia atingirá π quando $n \to \infty$.

Apesar da distribuição estacionária existir, a convergência das probabilidades de transição pode não ser garantida. Isto é, pode-se ter uma distribuição estacionária para a cadeia, sem que a distribuição limite exista.

Portanto, faz-se necessário impor restrições que garantam um comportamento assintótico da cadeia quando o número de iterações n tende ao infinito. Em outras palavras, garantir que a cadeia de Markov convirja para a distribuição limite.

Estas restrições são fundamentais para a aplicação dos métodos de simulação de Monte Carlo, apresentados na seção 3.3, que, segundo ALVES (1999): *trabalham essencialmente com a ideia de que, uma vez que algumas condições sejam satisfeitas, é possível construir uma cadeia de Markov que convirja para a distribuição em que se tem interesse. Para tanto, é necessário que exista uma distribuição estacionária – ou de equilíbrio – para a cadeia.*

Sabe-se que, se a cadeia for recorrente positiva, a distribuição estacionária existe; e se a cadeia for aperiódica pode-se garantir a convergência das probabilidades de transição. Resumindo, para que as probabilidades de transição convirjam para a distribuição estacionária de interesse, faz-se necessário que a cadeia seja ergódica.

Uma vez estabelecida a ergodicidade da cadeia, pode-se enunciar dois importantes teoremas-limite: o teorema ergódico, que equivale à Lei dos Grandes Números para cadeias de Markov, e o Teorema Central do Limite para Cadeias de Markov.

(i) Teorema ergódico

De acordo com GAMERMAN E LOPES (2006), dada que a cadeia é ergódica, ou seja, recorrente positiva e aperiódica, pode-se calcular a média ergódica:

$$\bar{t_n} = \frac{1}{n} . \sum_{i=1}^{n} t(X^{(i)}) \quad (3.15)$$

Se $E_\pi[t(X)] < \infty$ para uma distribuição limite única π, então:

Como estes resultados, segundo ALVES (1999): *pode-se utilizar as médias dos valores da cadeia como estimadores consistentes dos parâmetros da distribuição limite.*

(ii) Teorema Central do limite para cadeias de Markov

Ainda segundo GAMERMAN E LOPES (2006), antes de se falar do Teorema Central do Limite para cadeias de Markov, alguns conceitos fazem-se necessários.

Diz-se que uma cadeia é geometricamente ergódica se existe uma constante real $0 \leq \lambda < 1$ e uma função integrável M(x) tal que:

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x).\lambda^n \quad \text{, para todo } x \in S. \quad (3.16)$$

Se a função M não depende de x, a ergodicidade é uniforme.

Define-se ainda: autocovariância de ordem $k \geq 0$ da cadeia $t^{(n)} = t\left(X^{(n)}\right)$ como $\gamma_k = Cov_\pi\,(t^{(n)}, t^{(n+k)})$, a variância de $t^{(n)}$ como $\sigma^2 = \gamma_0$, a autocorrelação de ordem k como $\rho_k = \frac{\gamma_k}{\sigma^2}$ e, a variância ergódica como $Var_\pi\,(\bar{t_n})$.

Pode-se mostrar, segundo GAMERMAN E LOPES (2006), que:

$$\tau_n^2 = \sigma^2 \left(1 + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} . \rho_k\right) \quad (3.17)$$

E ainda $\tau_n^2 \to \tau^2$ quando $n \to \infty$, com:

$$\tau^2 = \sigma^2\,(1 + 2.\sum_{k=1}^{\infty} \rho_k) \quad (3.18)$$

Se a cadeia é uniformemente (geometricamente) ergódica e $t^2(X)(t^{2+\epsilon}(X))$ é integrável com respeito a π, para algum $\epsilon > 0$, então:

$$\frac{\bar{t_n} - E_\pi[t(X)]}{\tau/\sqrt{n}} \to N(0,1). \quad (3.19)$$

3.2.1.5. Cadeias reversíveis

Seja $(X^{(n)})_{n\geq 0}$ uma cadeia de Markov com probabilidades de transição $P(i,j)$ e distribuição estacionária π. Pode-se mostrar, GAMERMAN E LOPES (2006), que a sequência reversa de estados $X_n,\ X_{n-1}, X_{n-2}$, ...é também uma cadeia de Markov. As probabilidades de transição são $P_n^*(i,j) = \frac{\pi^{(n)}(j).P(j,i)}{\pi^{(n+1)}(i)}$.

Se $n \rightarrow \infty$, então $P_n^*(i,j) = \frac{\pi(j).P(j,i)}{\pi(i)}$ e a cadeia se torna homogênea.

A cadeia de Markov com essas propriedades é dita reversível e a condição de reversibilidade pode ser formalmente dada por:

$$\pi(i).P(i,j) = \pi(j).P(j,i) \tag{3.20}$$

Caso seja possível encontrar números não negativos α_i cuja soma seja 1, e de forma que a equação acima seja satisfeita, então a cadeia de Markov é reversível e estes números representam exatamente a distribuição estacionária da cadeia, pois se

$$\alpha_i.P(i,j) = \alpha_j.P(j,i) \quad \forall i,j \in S \quad \text{e} \quad \sum_i \alpha_i = 1 \tag{3.21}$$

Então

$$\sum_i \alpha_i .P(i,j) = \alpha_j.\sum_i P(j,i) = \alpha_j \tag{3.22}$$

Como encontrado em ROSS (1996), como a única solução das equações acima é a distribuição estacionária, tem-se que:

$$\alpha_i = \pi_i \quad \forall_i \tag{3.23}$$

3.3. Monte Carlo via Cadeias de Markov

Os métodos de Monte Carlo via Cadeias de Markov – MCMC – são, usualmente utilizados em problemas em que as densidades de interesse

apresentam uma estrutura complexa ou naqueles em que existe um alto número de dimensões.

Amplamente utilizados, estes métodos consistem, basicamente, na geração de valores de uma distribuição π, da qual a geração direta é complexa, por simulação estocástica de uma cadeia de Markov, de forma que a cadeia convirja para a distribuição a qual se deseja amostrar π. Para tanto, a distribuição invariante deve existir e as probabilidades de transição devem convergir para ela; para que isso ocorra, a cadeia deve ser ergódica: recorrente positiva e ergódica aperiódica.

De forma geral, de acordo com GAMERMAN E LOPES (2006), um processo de simulação ocorre da seguinte forma: considerando-se uma cadeia de Markov ergódica $(X^{(n)})_{n\geq 0}$, com espaço de estados S e probabilidade de transição P_{ij} e distribuição inicial π_0, gera-se um valor inicial desta cadeia $X^{(0)}$ de π_0. O próximo valor $X^{(1)}$ é gerado da distribuição P_{X_0}. O processo se repete para um *t* grande, $X^{(t)}$ gerado da distribuição $P_{X_{t-1}}$, de forma que a distribuição dos valores gerados esteja bem próxima da distribuição limite π.

Segundo ALVES (1999): *Uma vez que se obtenha evidências de que a cadeia tenha atingido a convergência, tem-se uma amostra da densidade conjunta das variáveis básicas e pode-se, então, avaliar qualquer função de interesse nos pontos gerados.*

Percebe-se que não só os métodos para geração das cadeias é importante, como uma das maiores dificuldades quando se amostra via Cadeias de Markov é a verificação da convergência da cadeia.

Neste capítulo será abordado o algoritmo escolhido para a geração da cadeia: o Amostrador de Gibbs, e os testes para monitoração da convergência da cadeia.

3.3.1. Amostrador de Gibbs

O Amostrador de Gibbs é a técnica mais usada para simulação estocástica, usando Cadeias de Markov, quando não se consegue gerar valores diretamente da densidade conjunta das variáveis aleatórias de interesse.

Consiste em um processo alternativo de geração de uma sequência de amostras de uma distribuição de probabilidade conjunta de duas ou mais variáveis aleatórias através de iterações sucessivas de distribuições condicionais completas $\pi(x_i|x_{i-1})$.

Pode ser descrito, segundo GAMERMAN E LOPES (2006), da seguinte forma:

1. Inicialize o contador de iterações da cadeia em j=1 com um conjunto de valores iniciais $X^{(0)} = \left(X_1^{(0)}, \dots, X_d^{(0)}\right)'$;
2. Obtenha um novo valor $X^{(j)} = \left(X_1^{(j)}, \dots, X_d^{(j)}\right)'$; a partir de $X^{(j-1)}$ através da geração sucessiva dos valores:

$$X_1{}^{(j)} \sim \pi\left(X_1|X_2^{(j-1)}, \dots, X_d^{(j-1)}\right)'$$

$$X_2{}^{(j)} \sim \pi\left(X_2|X_1^{(j)}, X_3^{(j-1)}, \dots, X_d^{(j-1)}\right)'$$

$$\vdots$$

$$X_d{}^{(j)} \sim \pi\left(X_d|X_1^{(j)}, X_2^{(j)}, \dots, X_{d-1}^{(j)}\right)'$$

3. Incremente o contador de *j* para *j + 1* e retorne ao passo 2 até que a convergência seja alcançada.

Um valor da distribuição de interesse π só é obtido quando o número de iterações se aproxima do infinito. Com isso, a maior dificuldade ao se implementar o Amostrador de Gibbs é determinar quão grande deve ser o número de iterações de

forma que a cadeia atinja a convergência.

Existem duas metodologias distintas para se diagnosticar a convergência de uma cadeia: uma informal e a outra formal.

A primeira, mais prática, tem uma perspectiva estatística, isto é, analisam-se as propriedades das observações que saem da cadeia através de técnicas gráficas a fim de identificar algum comportamento assintótico.

Porém, segundo GAMERMAN E LOPES (2006), *a dificuldade com este método é que nunca se poderá garantir a convergência porque ele é baseado somente nas observações provenientes da cadeia.*

Além disso, segundo ALVES (1999), *as técnicas gráficas podem indicar uma constância não tão evidente caso uma outra escala seja escolhida para os gráficos, o que conduz ao desenvolvimento de procedimentos formais para avaliação da convergência.*

O estudo da convergência de uma cadeia também pode ser abordado de maneira mais formal, na qual se tenta medir a distância e estabelecer limites sobre a função de distribuição gerada pela cadeia. Este é um tratamento teórico do problema em oposição ao tratamento empírico apresentado anteriormente.

Embora os dois métodos para se estudar a convergência sejam válidos e complementares um ao outro, resultados teóricos são mais difíceis de serem obtidos e de serem aplicados em problemas práticos. Por este motivo, utiliza-se neste estudo o método empírico, ou informal, por meio de técnicas gráficas.

3.3.2. Mistura finita de distribuições Normais com um número k conhecido de componentes

O foco principal de estudo deste projeto está na modelagem da amostra de

uma variável aleatória Y, através uma mistura finita de distribuições Normais com um número k de componentes conhecido. Para tanto, foram revisados, nos itens anteriores, os principais conceitos e teorias que alicerçam a Estimação Bayesiana em modelos de mistura.

A partir de agora, atribui-se arbitrariamente ao número k de componentes o valor 3 (k=3).

A mistura de distribuições Normais com 3 componentes estrutura-se da seguinte forma:

$$f(y|\boldsymbol{\psi}) = \sum_{j=1}^{3} \omega_j . N\left(\mu_j, \sigma_j^2\right) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2) + \omega_3 . N(\mu_3, \sigma_3^2) \qquad (3.24)$$

Nesta mistura os vetor paramétrico $\boldsymbol{\psi} = (\omega_1, \omega_2, \omega_3, \mu_1,\ \mu_2,\ \mu_3,\ \sigma_1^2,\ \sigma_2^2,\ \sigma_3^2)$ necessário para a modelagem dos dados é desconhecido.

Para a simulação estocástica e estimação dos parâmetros, escolheu-se como metodologia o Amostrador de Gibbs, também revisado nos itens anteriores. Este foi iterado no software estatístico R Project que está disponível gratuitamente.

Para sua implementação, faz-se necessário o desenvolvimento das distribuições condicionais completas de cada um dos parâmetros ao qual se deve estimar.

3.3.3. Desenvolvimento das distribuições dos parâmetros pelas condicionais completas

Para facilitar o algoritmo de simulação dos dados via Amostador de Gibbs foram utilizadas, como proposto em Liu (2010), duas variáveis auxiliares no desenvolvimento das condicionais completas:

$$S_j^y = \sum_{i=1}^{n} I_{z_i=j}\, y_i \quad \text{e} \qquad (3.25)$$

$$S_j^v = \sum_{i=1}^{n} I_{z_i=j}\ (y_i - \mu_j)^2\ . \tag{3.26}$$

No desenvolvimento da condicional completa do parâmetro μ_j encontra-se:

$$f(\mu_j|\ \sigma_j^2, \omega_j) = f(y|\boldsymbol{\psi})\ .\ p(\mu_j|\sigma_j^2)\ .\ p(\sigma_j^2)\ .\ p(\omega_1, \omega_2, \omega_3|\gamma_1, \gamma_2, \gamma_3)$$

$$= {\omega_{z_j}}^{n_j} \frac{1}{(\sqrt{2\pi})^{n_j}} \cdot \frac{1}{(\sqrt{\sigma_j^2})^{n_j}} \cdot \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_j^2} \cdot \sum_{i=1}^{n} I_{z_i=j}(y_i$$

$$-\ \mu_j)^2\} \cdot \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{\tau_j}{\sigma_j^2}} \cdot \exp\ \{-\frac{1}{2} \cdot \frac{\tau_j}{\sigma_j^2} \cdot (\mu_j$$

$$-\ \lambda_j)^2\} \cdot p(\sigma_j^2)\ .\ p(\omega_1, \omega_2, \omega_3|\gamma_1, \gamma_2, \gamma_3)\ \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_j^2} \cdot \sum_{i=1}^{n} I_{z_i=j}(y$$

$$-\ \mu_j)^2\} \cdot \exp\ \{-\frac{1}{2} \cdot \frac{\tau_j}{\sigma_j^2} \cdot (\mu_j - \lambda_j)^2\}$$

$$\propto \exp\{-\frac{1}{2} \cdot \frac{1}{\sigma_j^2} \cdot \sum_{i=1}^{n} I_{z_i=j}\ ({y_i}^2 - 2.\mu_j.y_i + {\mu_j}^2)$$

$$-\ \frac{2}{2} \cdot \frac{\tau_j}{\sigma_j^2} \cdot\ ({\mu_j}^2 - 2.\mu_j\lambda_j + {\lambda_j}^2)\}$$

$$\propto \exp\{-\frac{\tau_j}{2.\sigma_j^2}\ [\frac{1}{\tau_j} \cdot (-2.\mu_j \cdot \sum_{i=1}^{n} I_{z_i=j}\ y_i \cdot n_j \cdot {\mu_j}^2 + {\mu_j}^2\ - 2.\mu_j\lambda_j]\}$$

$$\propto \exp\{-\frac{1}{2\sigma_j^2}\ [-2.\mu_j.S_j^y + (n_j + \tau_j) \cdot {\mu_j}^2 - 2.\mu_j.\lambda_j.\tau_j]\}$$

$$\propto \exp\{-\frac{1}{2\sigma_j^2} \cdot (n_j + \tau_j)[\ {\mu_j}^2 - 2.(\frac{S_j^y + \tau_j.\lambda_j}{n_j + \tau_j}).\mu_j]\}$$

encontra-se o núcleo de uma densidade Normal:

$$\mu_j|\ \sigma_j^2, y_i, z \sim N\left(\frac{S_j^y + \tau_j.\lambda_j}{n_j + \tau_j}, \frac{\sigma_j^2}{n_j + \tau_j}\right) \tag{3.27}$$

No desenvolvimento da condicional completa de σ_j^2 tem-se:

$$f(\sigma_j^2|\,\mu_j\,,w_j) = f(y|\psi)\,.p(\sigma_j^2)\,.\;p(\mu_j|\sigma_j^2)\;.\;p(w_1\,,w_2\,,w_3|\gamma_1\,,\gamma_2\,,\gamma_3)$$

$$= {\omega_{z_j}}^{n_j}.\frac{1}{(\sqrt{2\pi})^{n_j}}\,.\frac{1}{(\sqrt{\sigma_j^2})^{n_j}}\,.\exp\{-\frac{1}{2}\,.\frac{1}{\sigma_j^2}\,.\sum_{i=1}^{n} I_{z_i=j}(y_i$$

$$-\;\mu_j)^2\}\,.\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}\,.(\sigma_j^2)^{-(\alpha_j+1)}\,.e^{-\frac{\beta_j}{\sigma_j^2}}.\frac{1}{\sqrt{2\pi}}\,.\sqrt{\frac{\tau_j}{\sigma_j^2}}\,.\exp\{-\frac{1}{2}\,.\frac{\tau_j}{\sigma_j^2}\,.(\mu_j$$

$$-\;\lambda_j)^2\}$$

$$\propto (\sigma_j^2)^{-\frac{n_j}{2}}\,.\exp\{-\frac{1}{2}\,.\frac{1}{\sigma_j^2}\,.\sum_{i=1}^{n} I_{z_i=j}(y$$

$$-\;\mu_j)^2\}\,.(\sigma_j^2)^{-(\alpha_j+1)}\,.e^{-\frac{\beta_j}{\sigma_j^2}}.(\sigma_j^2)^{-0,5}.\exp\{-\frac{1}{2}\,.\frac{\tau_j}{\sigma_j^2}\,.(\mu_j-\;\lambda_j)^2$$

$$\propto (\sigma_j^2)^{-0,5n_j\,-\alpha_j-1-0,5}.\exp\{-\frac{1}{\sigma_j^2}\left(\frac{1}{2}\sum_{i=1}^{n} I_{z_i=j}(y-\;\mu_j)^2+\;\beta_j\right.$$

$$\left.+\,0{,}5\tau_j(\mu_j-\;\lambda_j)^2)\right\}$$

$$\propto (\sigma_j^2)^{-(\alpha_j+0,5n_j+0,5+1)}.\exp\{-\frac{1}{\sigma_j^2}\;(0{,}5S_j^{\mathrm{v}}+\;\beta_j+0{,}5\tau_j(\mu_j-\;\lambda_j)^2)\}$$

Observa-se o núcleo de uma Gama Inversa:

$$\sigma_j^2\,\Big|\mu_j,y,z\sim GI\,\left(\alpha_j+0{,}5(n_j+1)\,,0{,}5S_j^{\mathrm{v}}+\;\beta_j+0{,}5\tau_j(\mu_j-\;\lambda_j)^2\right) \qquad (3.28)$$

Para ω_i, o desenvolvimento da condicional completa conduz a:

$$\pi(\omega_i)\propto\;{\omega_{z_j}}^{n_j}.\frac{1}{(\sqrt{2\pi})^{n_1}}\,.\frac{1}{(\sqrt{\sigma_1^2})^{n_1}}\,.\exp\{-\frac{1}{2}\,.\frac{1}{\sigma_1^2}\,.\sum_{i=1}^{n} I_{z_i=1}(y_i$$

$$-\;\mu_1)^2\}\,.\frac{\Gamma(\gamma_1+\gamma_2+\;\gamma_3)}{\Gamma(\gamma_1)\,.\Gamma(\gamma_2)\,.\Gamma(\gamma_3)}\,.{\omega_1}^{\gamma_1-1}\,.{\omega_2}^{\gamma_2-1}\,.{\omega_3}^{\gamma_3-1}$$

,

$$\pi(\omega_i) \sim D\ (\gamma_1 + n_1\ , \gamma_2 + n_2\ , \gamma_3 + n_3) \tag{3.29}$$

Finalmente, ao se desenvolver a condicional completa da variável latente z_j, encontra-se:

$$\pi(z_j) = f(y_j|\psi).p(z_j)$$

$$\pi(z_j) = \frac{1}{\sqrt{2\pi}} . \frac{1}{\sqrt{\sigma_j^2}} . \exp\{ - \frac{1}{2} . \frac{1}{\sigma_j^2} . (y_i - \mu_j)^2 \} . \omega_j$$

$$\pi(z_j) \propto \frac{\omega_j}{\sigma_j} . \exp\ \{ - \frac{(y_i - \mu_j)^2}{2\sigma_j^2} \} \tag{3.30}$$

3.3.4. Amostrador de Gibbs para misturas de Normais

Para a modelagem dos dados através de uma mistura de três densidades Normais, faz-se necessária a adaptação do Amostrador de Gibbs para mistura de três componentes Normais da forma $\sum_{j=1}^{3} \omega_j . N(\mu_j, \sigma_j^2)$.

Conforme descrito em LIU (2010), tem-se o seguinte algoritmo:

1. Inicialização: escolha $\omega^{(0)}$ e $\theta^{(0)}$.
2. Etapa t. Para t = 1, 2,...

    2.1. Gerar $z_i^{(t)}$ (1 = 1,...,n) de (j = 1, 2, 3)

$$P(z_i{}^{(t)} = j) \propto \frac{\omega_j{}^{(t-1)}}{\sigma_j{}^{(t-1)}} . \exp\{ - \frac{(y_i - \mu_j{}^{(t-1)})^2}{2(\sigma_j^2)_j^{(t-1)}} \}$$

    Calcule $n_j^{(t)} = \sum_{l=1}^{n} I_{z_l^{(t)}=j}$, $(S_j^y)^{(t)} = \sum_{i=1}^{n} I_{z_i=j}\ y_i$

    2.2. Gerar $\omega^{(t)}$ de $D\ (\gamma_1 + n_1\ , \gamma_2 + n_2\ , \gamma_3 + n_3)$.

    2.3. Gerar $\mu_j^{(t)}$ de

$$N\left(\frac{(S_j^y)^{(t)} + \tau_j.\lambda_j}{n_j^{(t)} + \tau_j}, \frac{(\sigma_j^2)^{(t-1)}}{n_j^{(t)} + \tau_j}\right)$$

Calcule $(S_j^v)^{(t)} = \sum_{i=1}^{n} I_{z_i=j} \ (y_i - \mu_j{}^{(t)})^2$

2.4. Gerar $(\sigma_j^2)^{(t)}$ (j = 1, 2, 3) de:

$$GI\left(\alpha_j + 0{,}5(n_j + 1), 0{,}5(S_j^v)^{(t)} + \beta_j + 0{,}5\tau_j\left(\mu_j{}^{(t)} - \lambda_j\right)^2\right)$$

4. Modelagem dos dados

Objetivando a modelagem da variável sinistro agregado através de uma mistura de k densidades Normais com médias e variâncias distintas, assumindo k um número conhecido, foi aproveitado um banco de dados fortemente utilizado na literatura atuarial internacional.

O banco de dados consiste de 1.500 pagamentos de indenizações referentes ao seguro de Responsabilidade Civil Geral americano em dólar dos Estados Unidos.

Este banco de dados está disponível gratuitamente e pode ser encontrado no pacote do R Project *Copula and evd.*

Antes de se apresentar as características estatísticas dos dados e os resultados obtidos, faz-se necessário uma breve explanação do que seja o seguro de Responsabilidade Civil Geral, usualmente denominado, simplesmente, por seguro de Responsabilidade Civil.

Diferentemente da maioria dos tipos de seguros, no Brasil, alguns seguros de Responsabilidade Civil são obrigatórios, dentre eles, destaca-se o DPVAT – Danos Pessoais caudados por Veículos Automotores de via Terrestre, o que evidencia a sua importância na sociedade e no meio atuarial.

O principal objetivo do seguro de Responsabilidade Civil é proteger o segurado de eventuais reclamações ou ações na justiça em que este seja responsabilizado civilmente por ter causado danos involuntários a outras pessoas, sejam materiais ou corporais.

Segundo a SUSEP – Superintendência de Seguros Privados – , no Processo nº 15414.001870/2005-24, em um seguro de Responsabilidade Civil "*para cada cobertura contratada, a seguradora garante pagar quantias devidas e/ou reembolsar as despendidas, pelo Segurado, na reparação de danos materiais e/ou corporais*

causados a terceiros, e/ou na ações emergenciais empreendidas para tentar evitá-los e/ou minorá-los...”.

Apresentada a estrutura do banco de dados, e definido o tipo de seguro que o compõe, segue abaixo o histograma dos 1.500 pagamentos de indenizações referentes ao seguro de Responsabilidade Civil Geral americano:

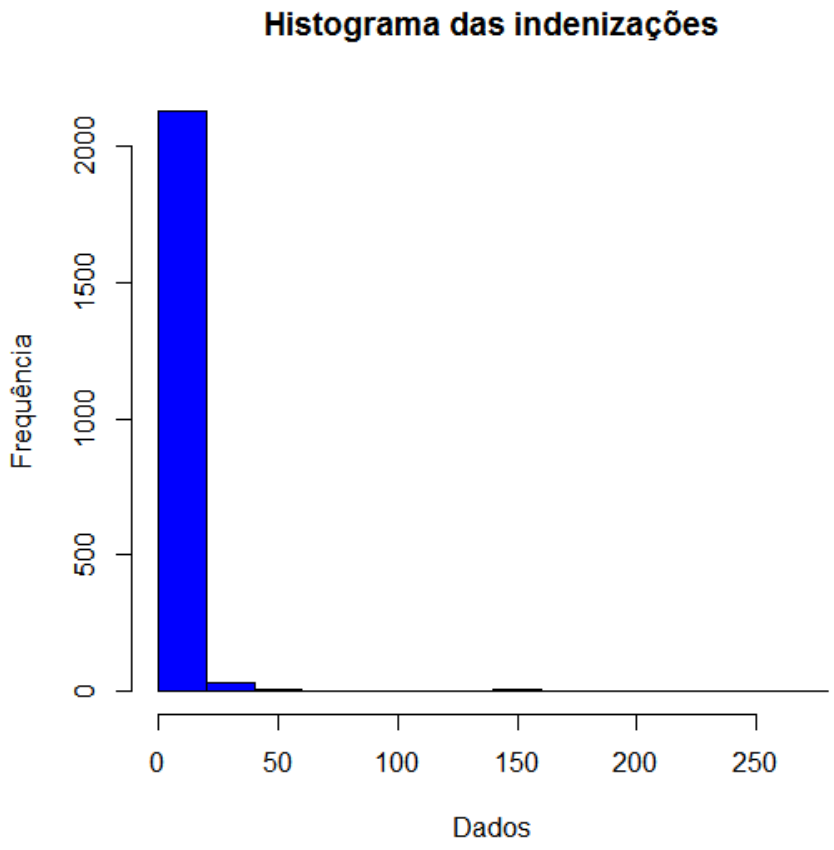

Figura 1 – Histograma dos dados

Pelo histograma, percebe-se que os dados apresentam uma forte assimetria à direita. Nota-se ainda que este comportamento é comum em dados de pagamentos de indenizações que, geralmente, apresentam um número maior de perdas pequenas e um número menor de perdas maiores.

Devido à forte assimetria dos dados, foi aplicado o logaritmo natural aos dados. Esta prática reduz os valores extremos que causam a assimetria e curtose.

Após a transformação dos dados para a escala logarítmica, os 1.500 dados de pagamentos de indenizações tomam outra forma; como pode ser observado na figura 2.

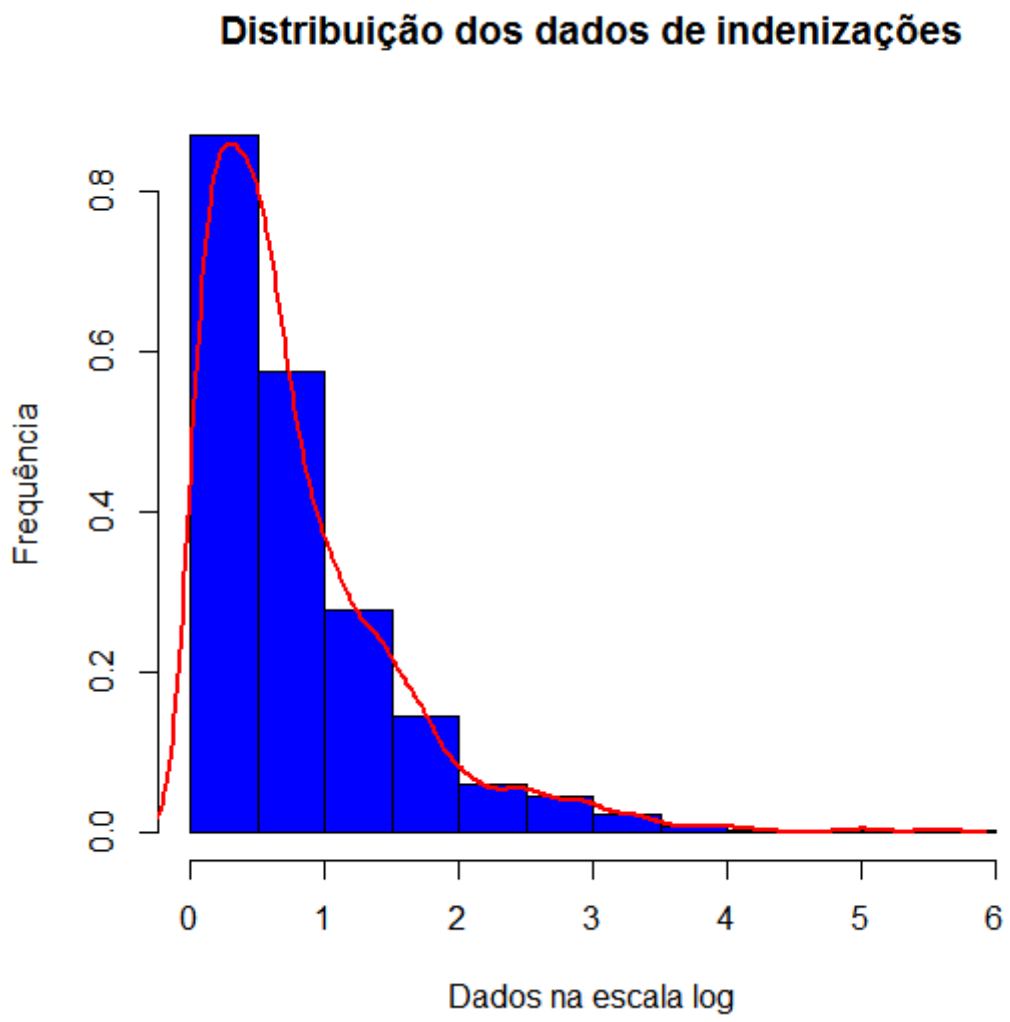

Figura 2 – Distribuição dos dados na escala log

Apesar dos dados na escala logarítmica, aparentemente, se ajustarem melhor a uma distribuição Normal Assimétrica, neste estudo os dados serão modelados por uma mistura de distribuições Normais pela sua simplicidade. Deixa-se para estudos

posteriores a modelagem dos dados por uma mistura de Normais Assimétricas.

Os dados foram modelados, através do Amostrador de Gibbs, por uma mistura de k distribuições Normais:

$$f(y|\boldsymbol{\psi}) = \sum_{j=1}^{K} \omega_j . N\left(\mu_j, \sigma_j^2\right) = \omega_1 . N(\mu_1, \sigma_1^2) + \omega_2 . N(\mu_2, \sigma_2^2) + \ \ldots + \omega_k . N(\mu_k, \sigma_k^2) \quad (4.1)$$

A fim de se estimar cada um dos parâmetros que compõem o espaço paramétrico $\boldsymbol{\psi} = (\omega_1, \omega_2, \omega_3,\ \mu_1,\ \mu_2,\ \mu_3,\ \sigma_1^2,\ \sigma_2^2,\ \sigma_3^2)$, o algoritmo foi iterado 10.000 vezes e as 2.000 primeiras iterações foram excluídas da análise por estarem no período de aquecimento da cadeia, ou, seja, período que a cadeia leva até que se comporte de forma estacionária, conhecido como *burning-in period*.

Para a primeira componente foi encontrado um peso ω_1 estimado pontualmente no valor de $\omega_1 = 0.378$, média $\mu_1 = 0.982$ e desvio padrão $\sigma_1 = 0.424$, como observado nos gráficos apresentados a seguir:

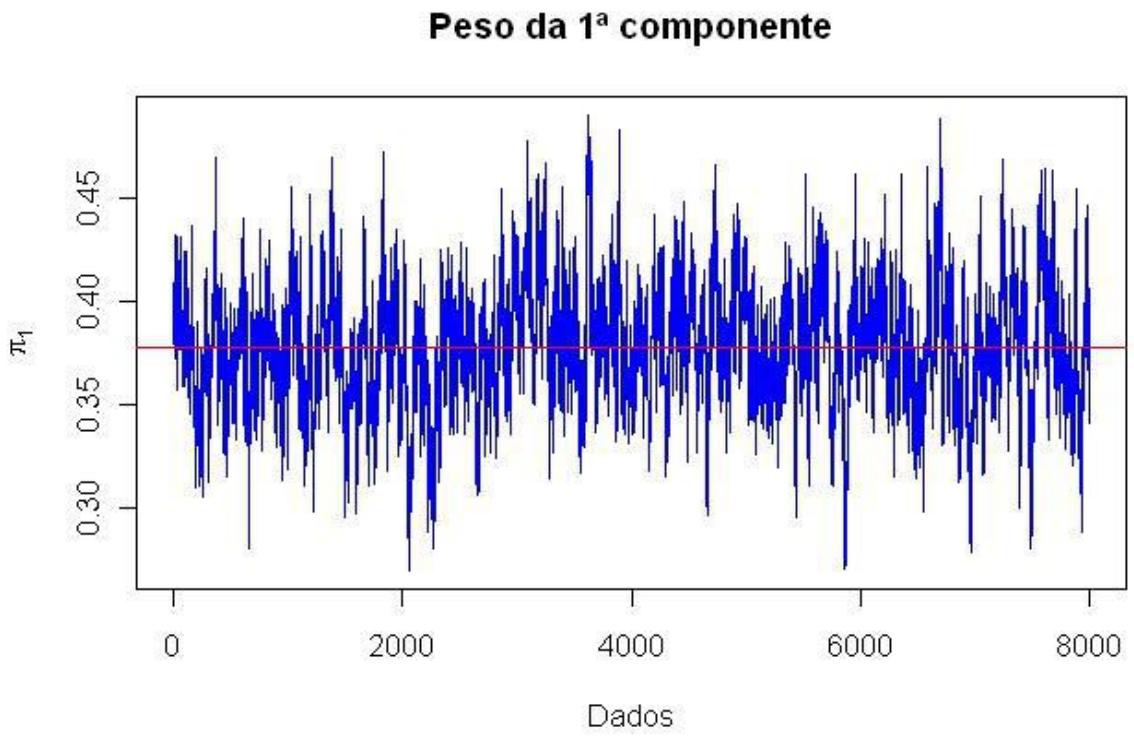

Figura 3 – Comportamento do peso da 1ª componente

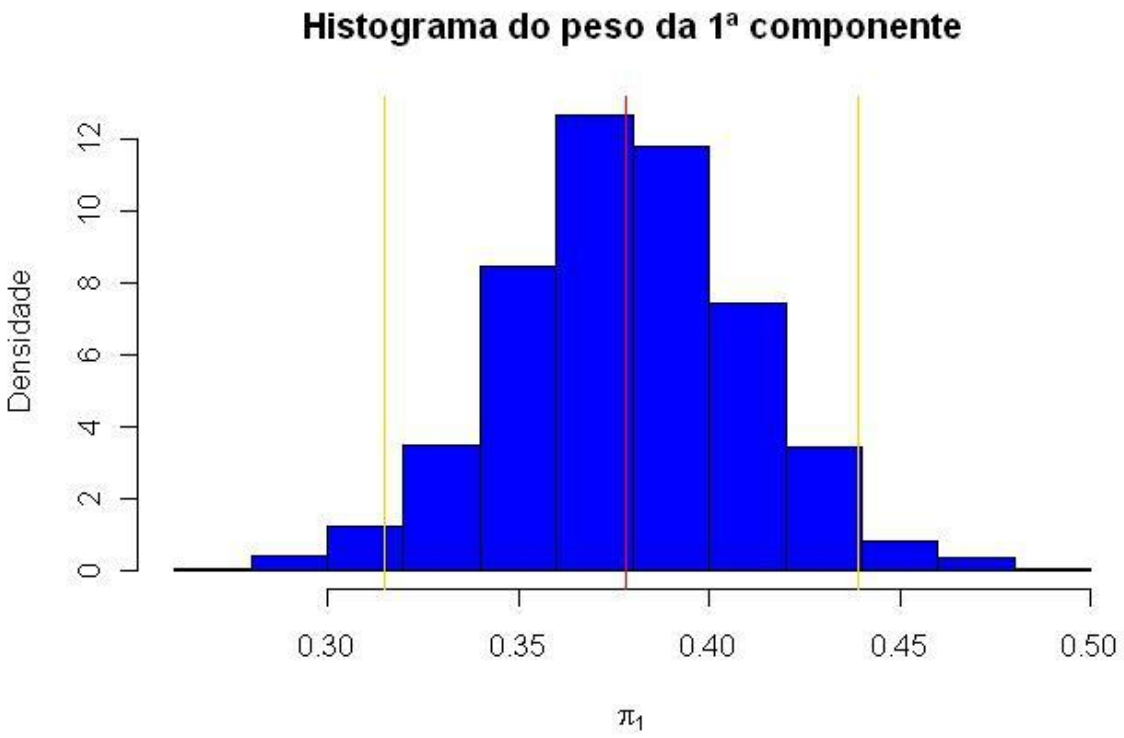

Figura 4 – Histograma dos valores gerados para o peso da 1ª componente

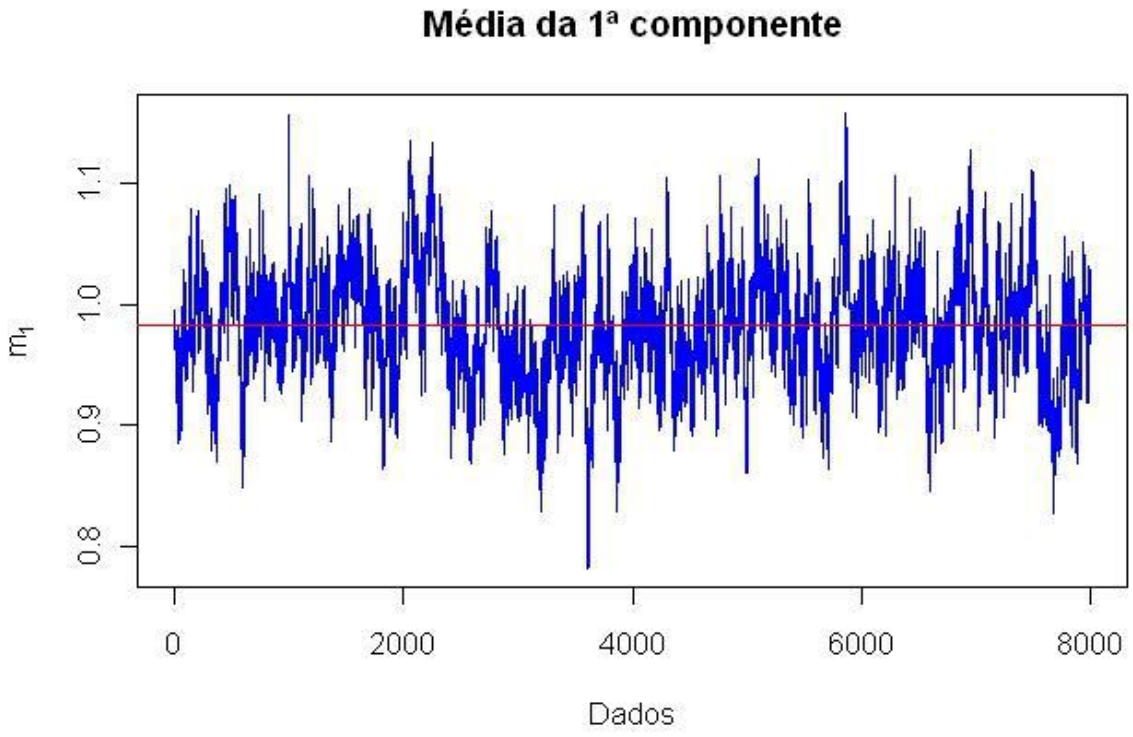

Figura 5 – Comportamento da média da 1ª componente

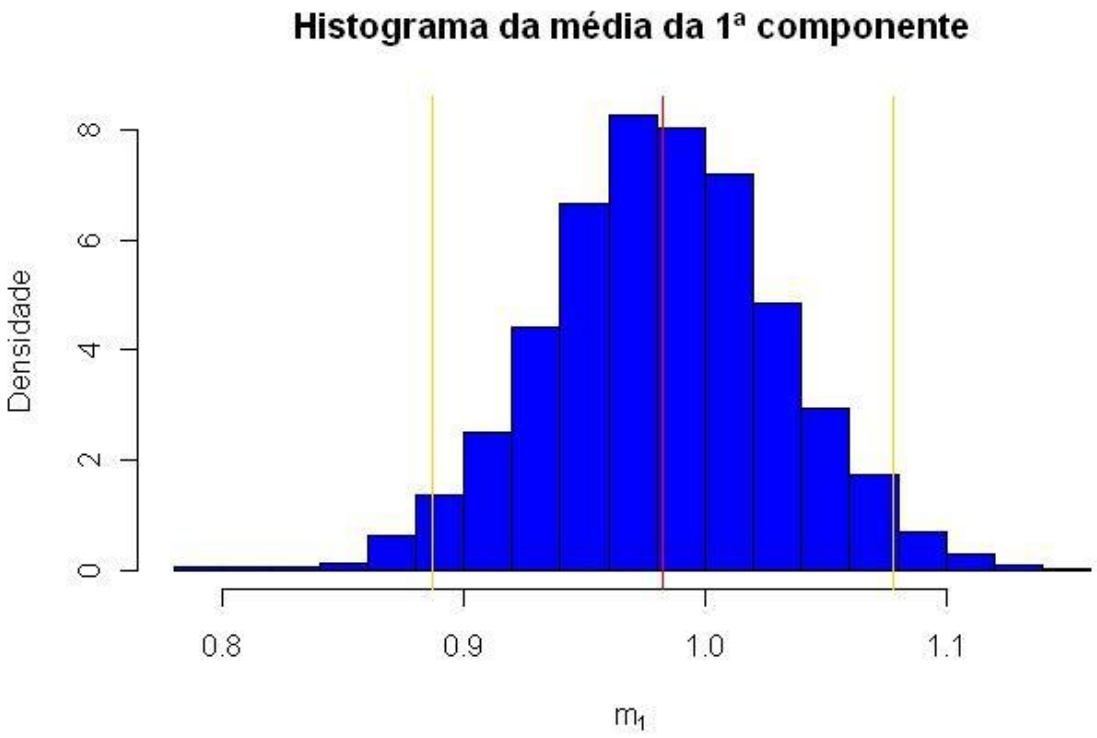

Figura 6 – Histograma dos valores gerados para a média da 1ª componente

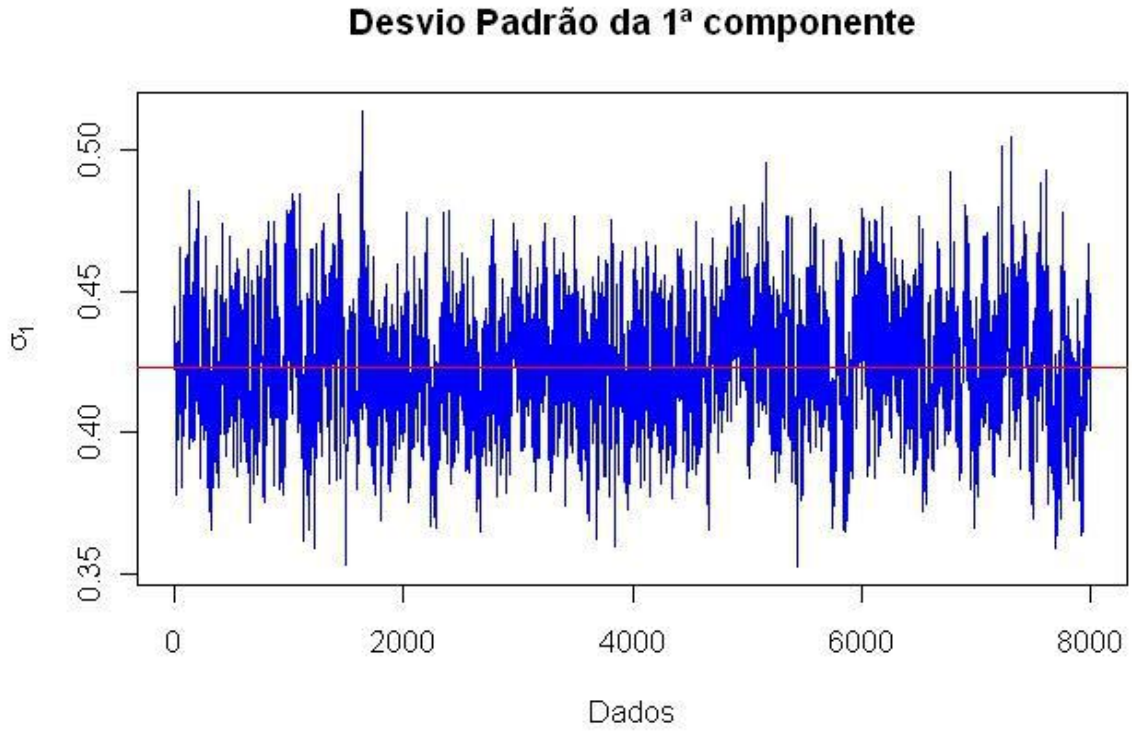

Figura 7 – Comportamento do desvio padrão da 1ª componente

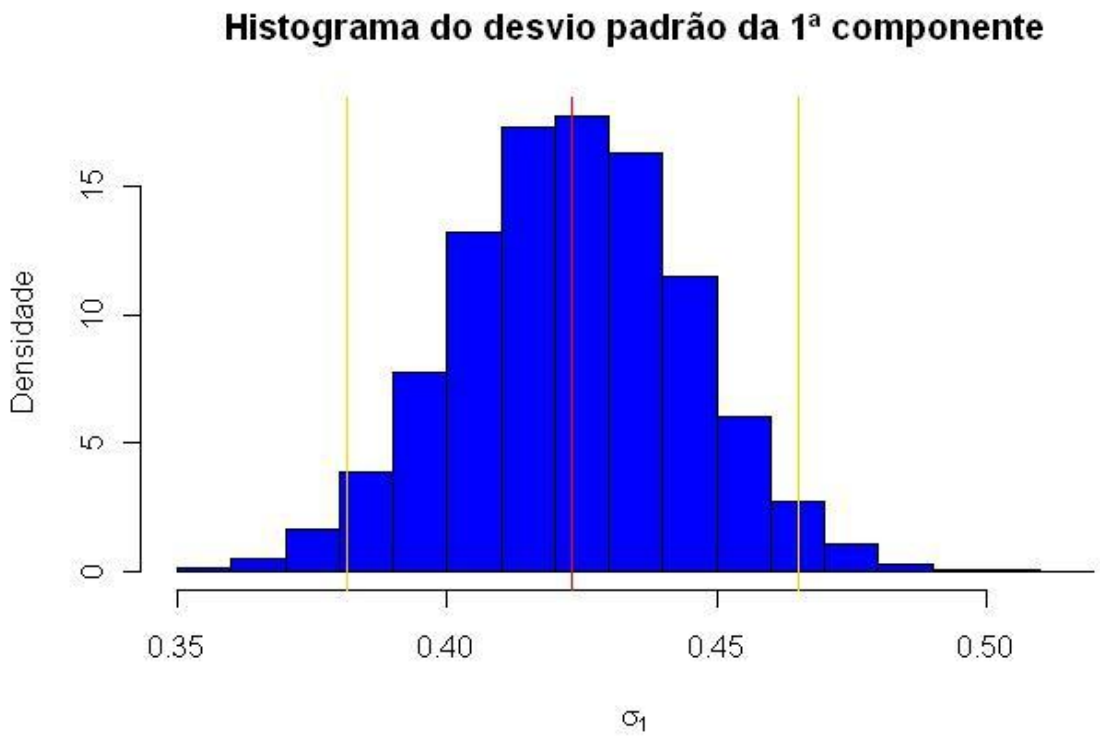

Figura 8 – Histograma dos valores gerados para o desvio padrão da 1ª componente

Já para a segunda componente foi encontrado um peso no valor de $\omega_2 = 0.502$, média $\mu_2 = 0.326$ e desvio padrão $\sigma_2 = 0.208$, que podem ser observados nos gráficos gerados que seguem:

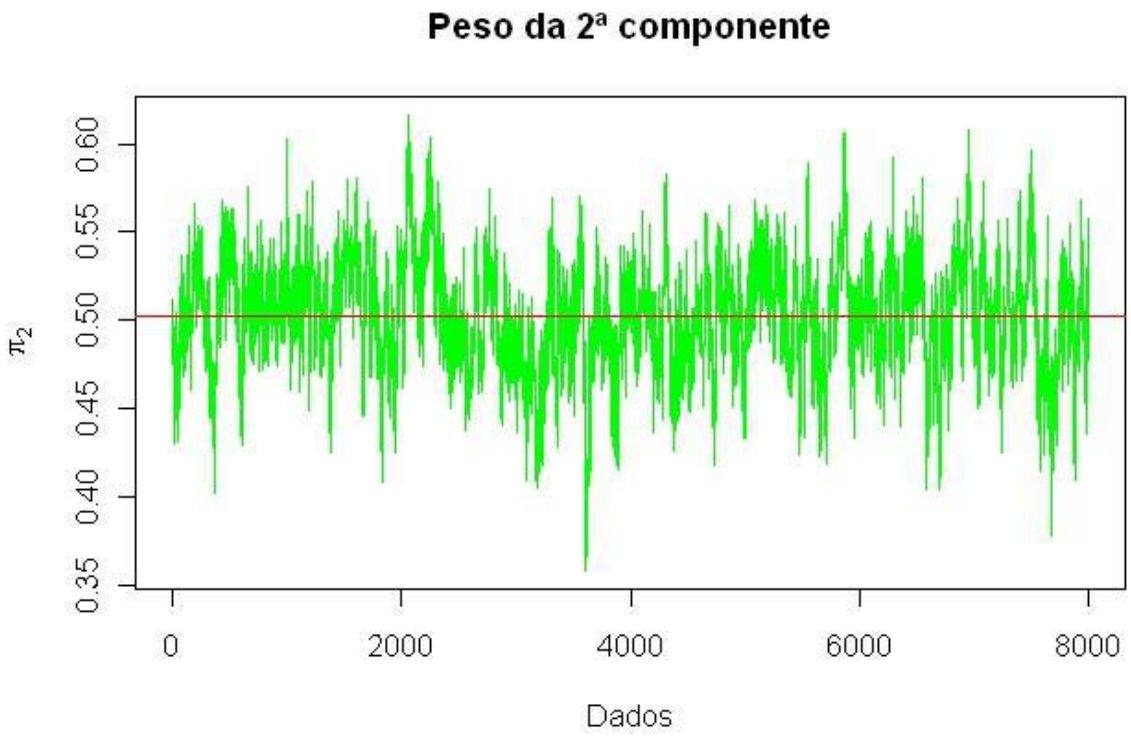

Figura 9 – Comportamento do peso da 2ª componente

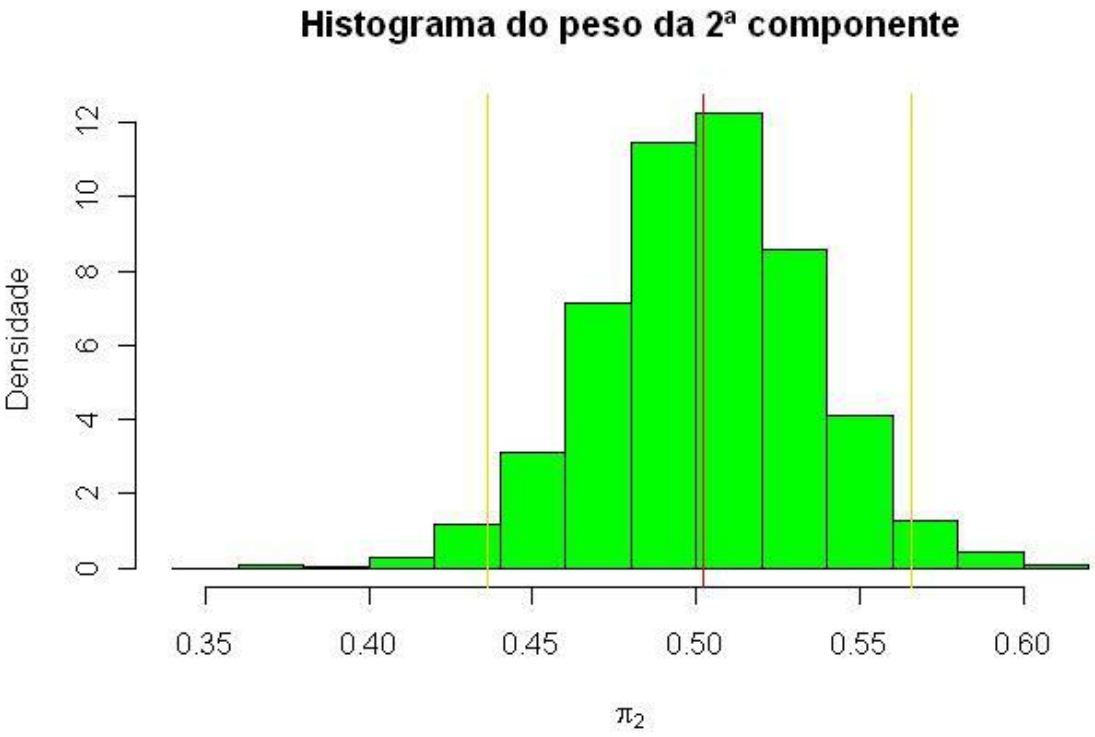

Figura 10 – Histograma dos valores gerados para o peso da 2ª componente

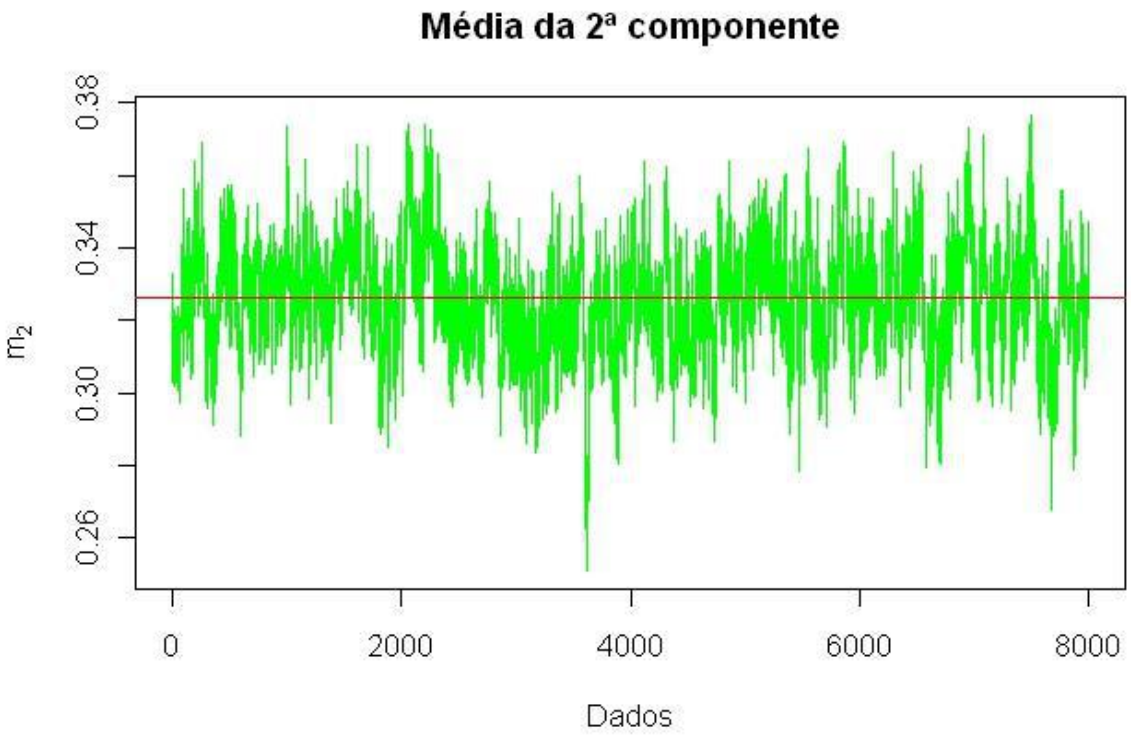

Figura 11 – Comportamento da média da 2ª componente

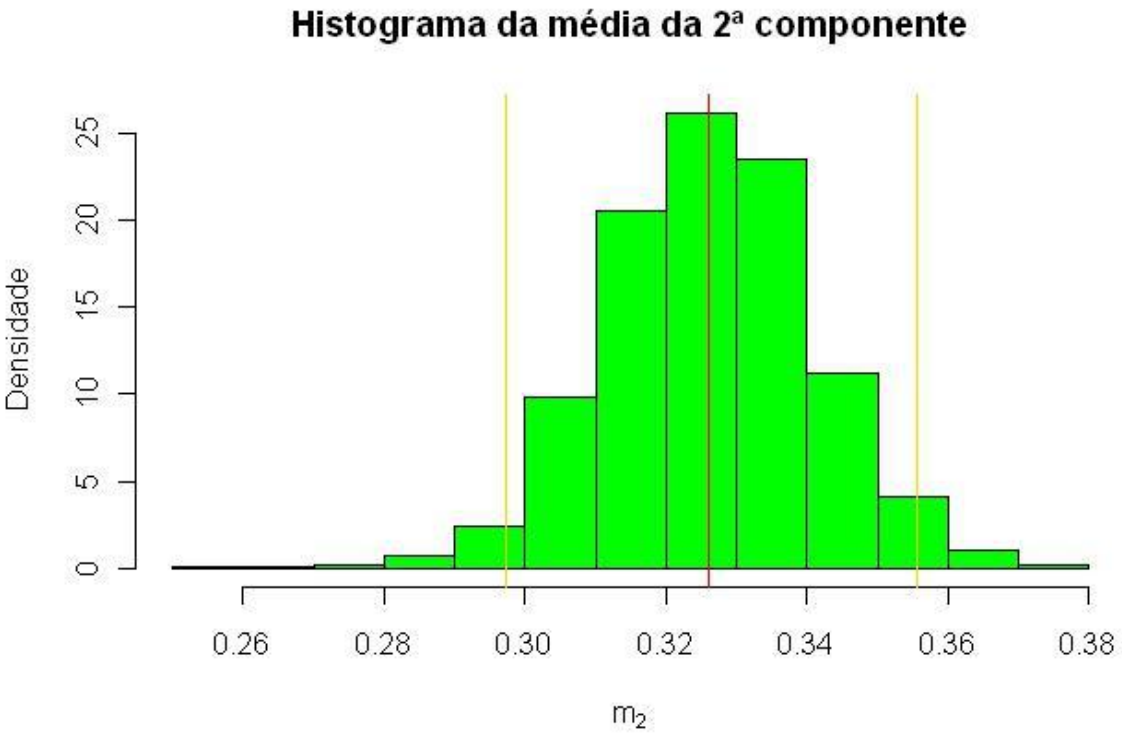

Figura 12 – Histograma dos valores gerados para a média da 2ª componente

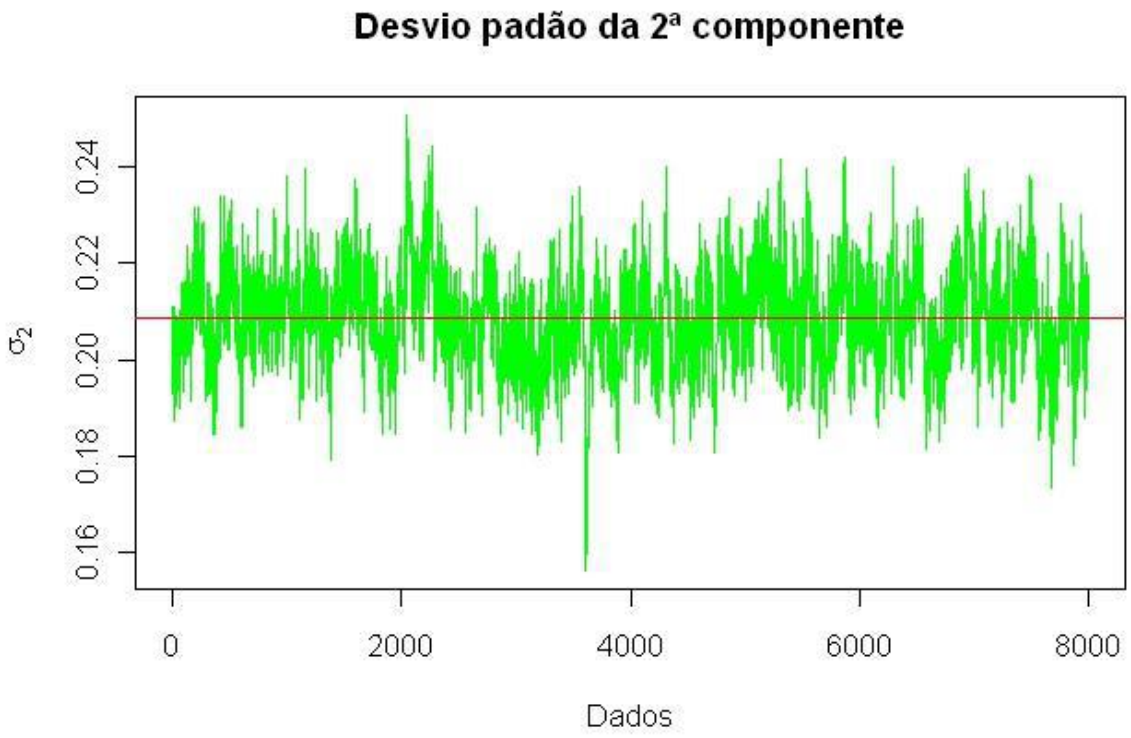

Figura 13 – Comportamento do desvio padrão da 2ª componente

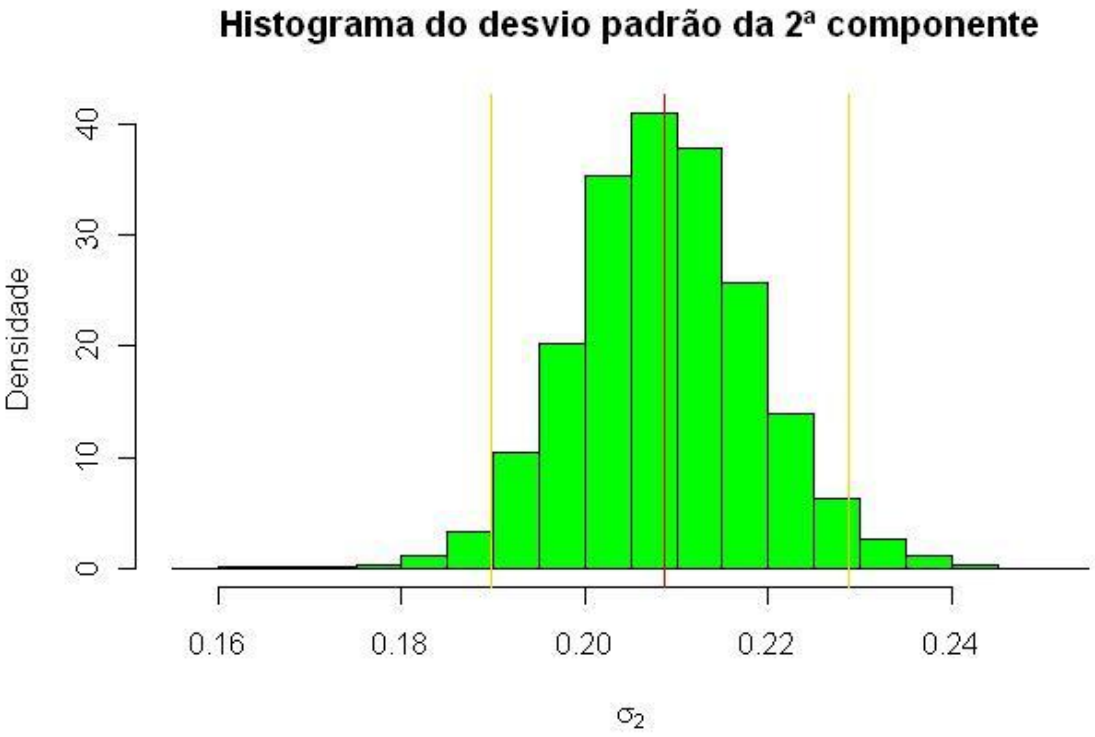

Figura 14 – Histograma dos valores gerados para o desvio padrão da 2ª componente

Finalmente, após as 10.000 iterações, foram encontradas estimativas pontuais para os três parâmetros que compõem a terceira componente, quais sejam: peso $\omega_3 = 0.119$, média $\mu_3 = 2.132$ e desvio padrão $\sigma_3 = 0.867$. Estes valores podem ser observados através dos gráficos gerados que seguem:

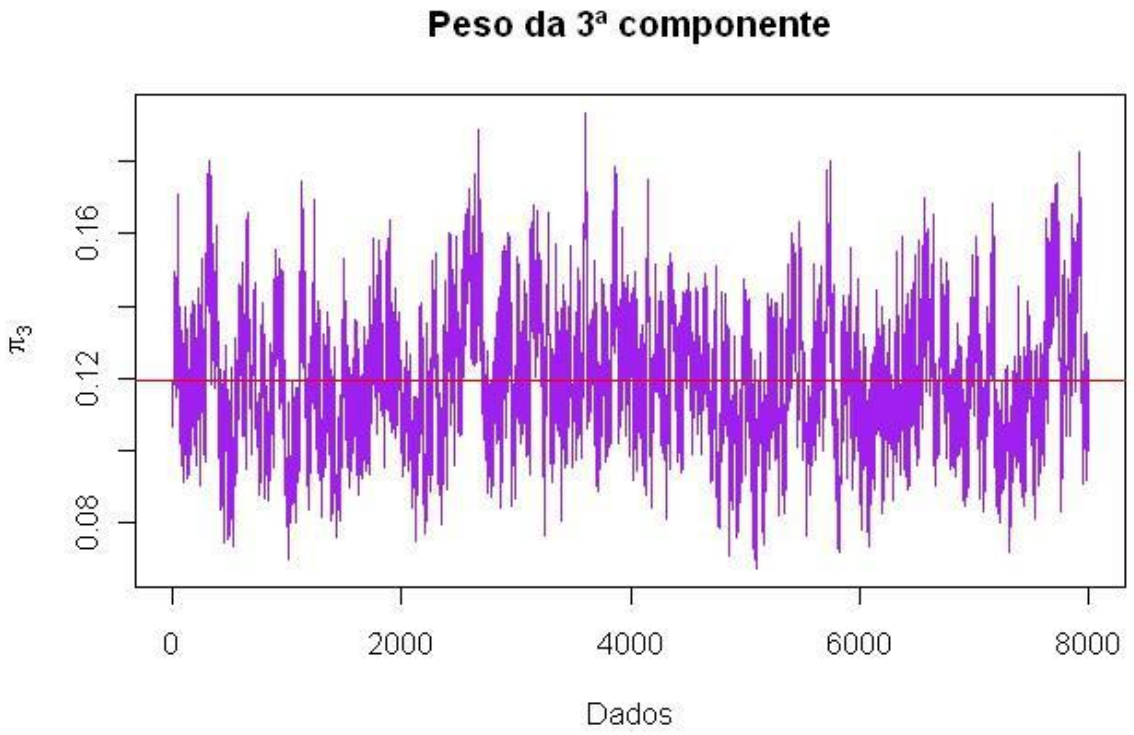

Figura 15 – Comportamento do peso da 3ª componente

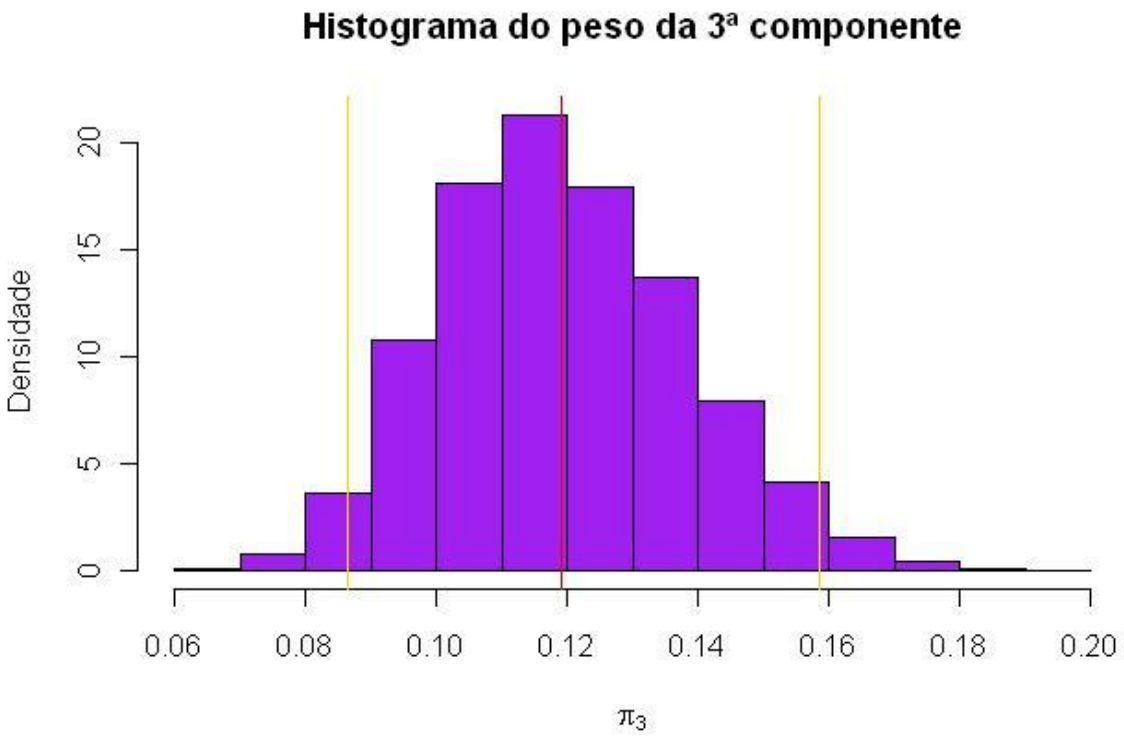

Figura 16 – Histograma dos dados gerados para o peso da 3ª componente

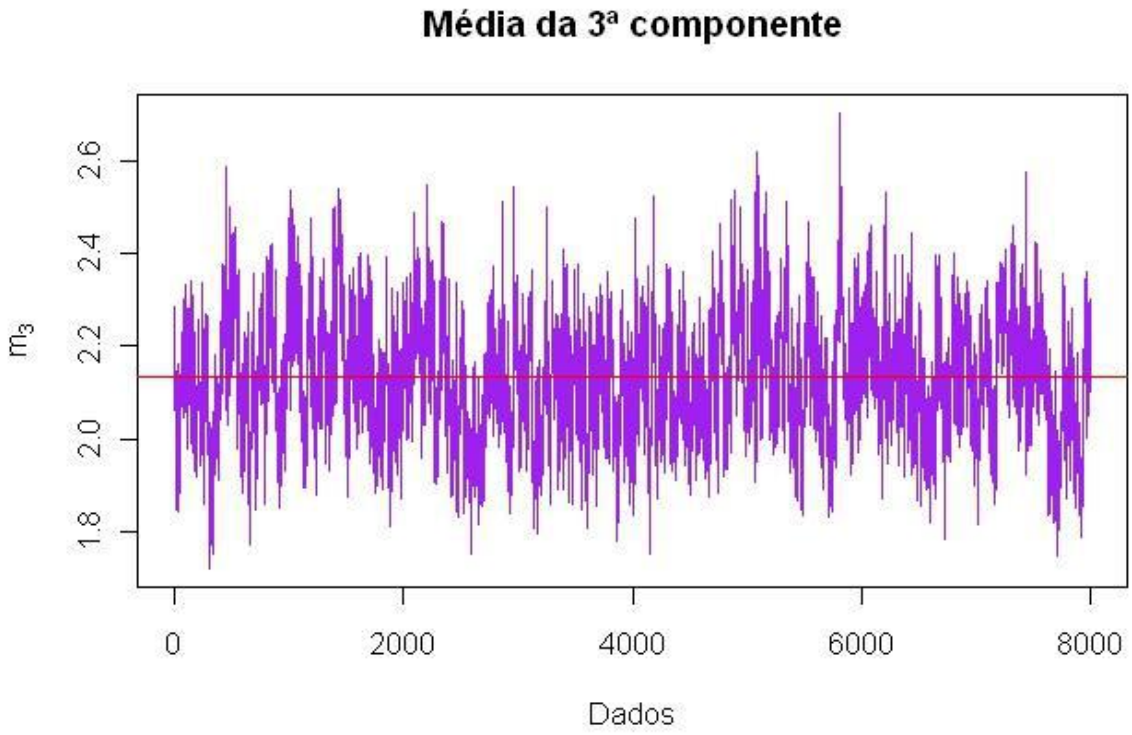

Figura 17 – Comportamento da média da 3ª componente

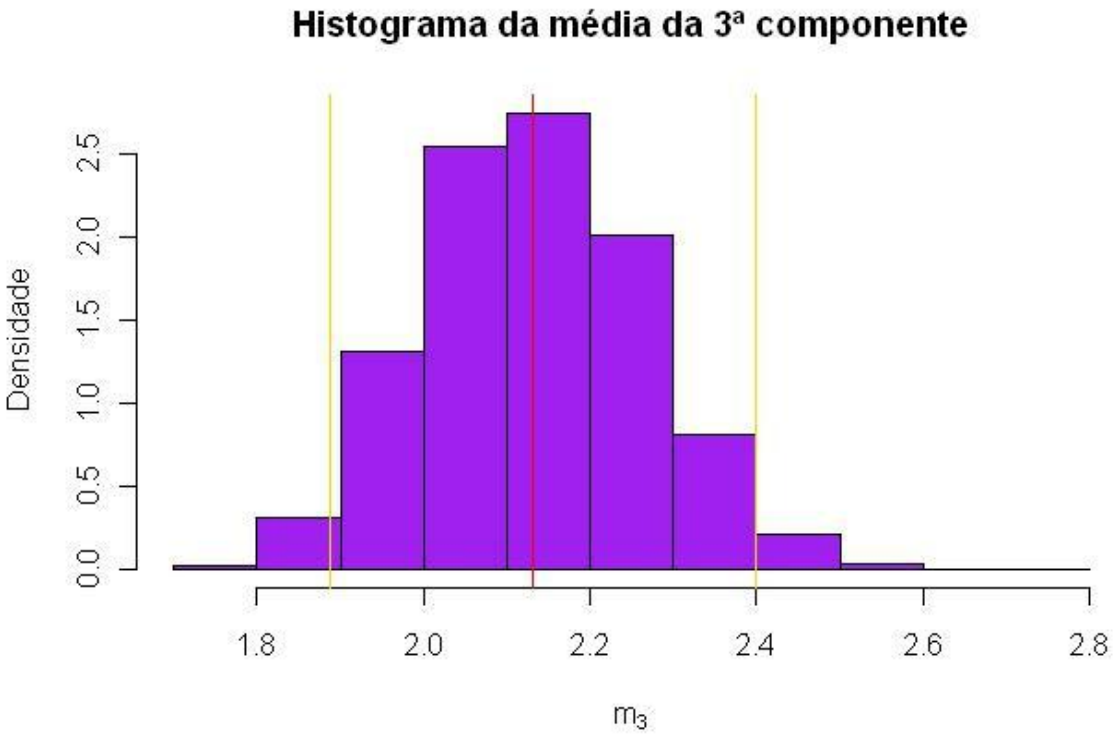

Figura 18 – Histograma dos dados gerados para a média da 3ª componente

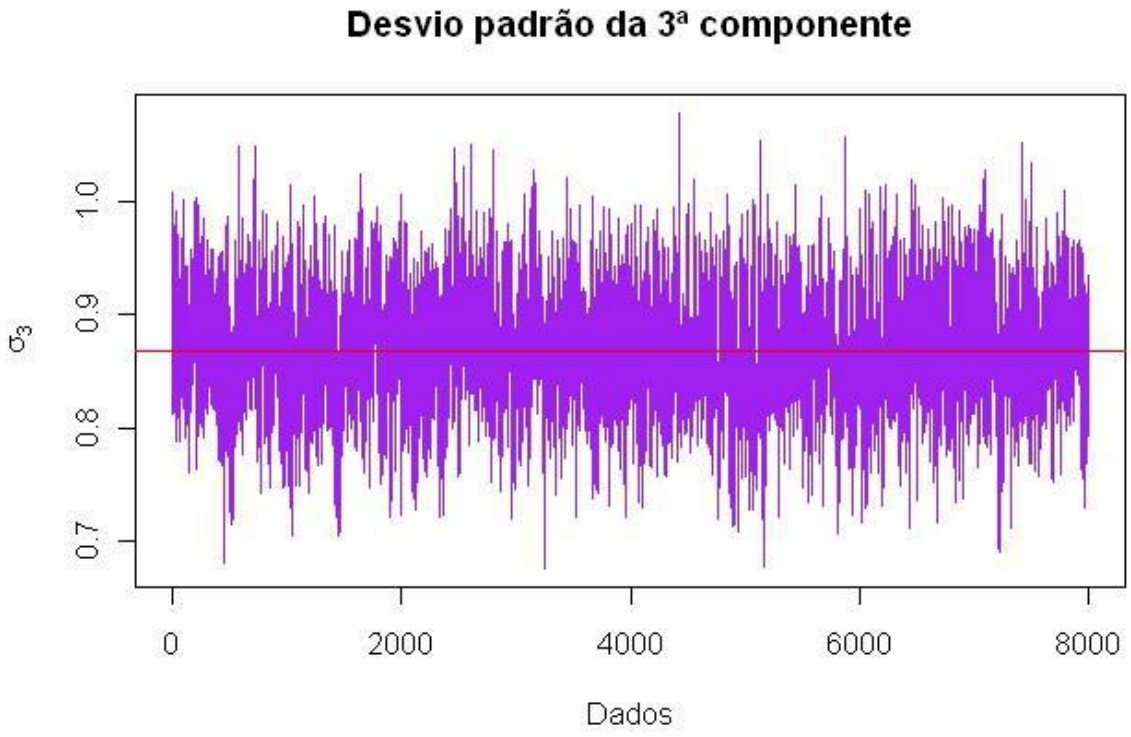

Figura 19 – Comportamento do desvio padrão da 3ª componente

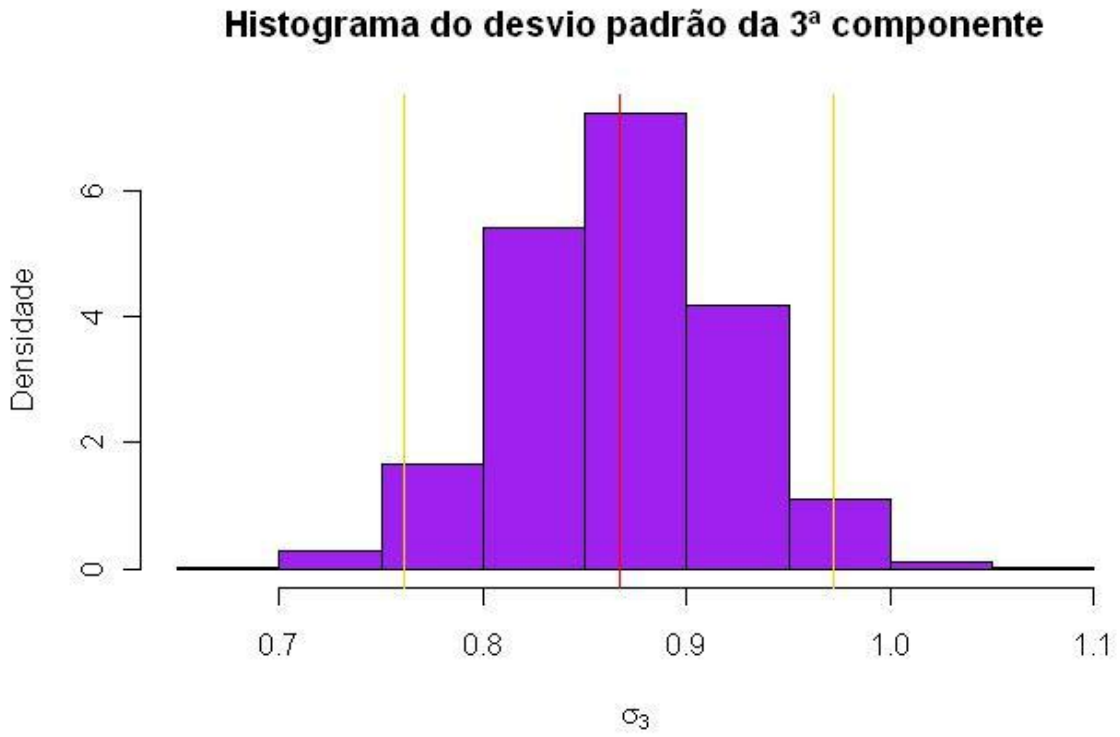

Figura 20 – Histograma dos dados gerados para o desvio padrão da 3ª componente

Por uma análise gráfica, pode-se observar que os valores gerados para cada um dos parâmetros, a partir das 10.000 iterações, apresentam um comportamento convergente. Tem-se, portanto, evidências empíricas de que a cadeia atingiu a convergência e torna-se possível amostrar da densidade conjunta.

Observa-se também, através da tabela 1 a seguir, que os parâmetros estimados apresentam valores distintos o que confirma que o conjunto paramétrico é identificável.

	1ª componente	2ª componente	3ª componente
Peso (ω)	0.378	0.502	0.119
Média (μ)	0.982	0.326	2.132

Desvio padrão (σ)	0.423	0.208	0.867

Tabela 1 – Resumo dos parâmetros estimados

Através dos resultados das iterações, pode-se formar uma matriz com os 8.000 valores gerados para cada um dos parâmetros de interesse, a partir dela, estima-se todo o conjunto paramétrico $\psi = (\omega_1, \omega_2, \omega_3, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2)$.

O processo de estimação do conjunto paramétrico ψ ocorre da seguinte forma: suponha que se tenha uma amostra de Monte Carlo de tamanho N para um parâmetro δ: $\delta^{(1)}$, $\delta^{(2)}$, ..., $\delta^{(2)}$; deseja-se obter uma amostra de Monte Carlo para $\varphi = g(\delta)$, em que g é uma função qualquer, deve-se tomar $\varphi^{(1)} = g(\delta^{(1)})$, $\varphi^{(2)} = g(\delta^{(2)})$, ..., $\varphi^{(N)} = g(\delta^{(N)})$. Uma estimativa de $E[\varphi]$ é dada por:

$$E[\varphi] = \sum_{i=1}^{N} \frac{\varphi^i}{N}. \quad (4.2)$$

Aplicando o processo de estimação acima, chega-se aos valores da densidade conjunta que pode ser representa por uma curva suave ou através de um histograma, como apresentado na figura a seguir.

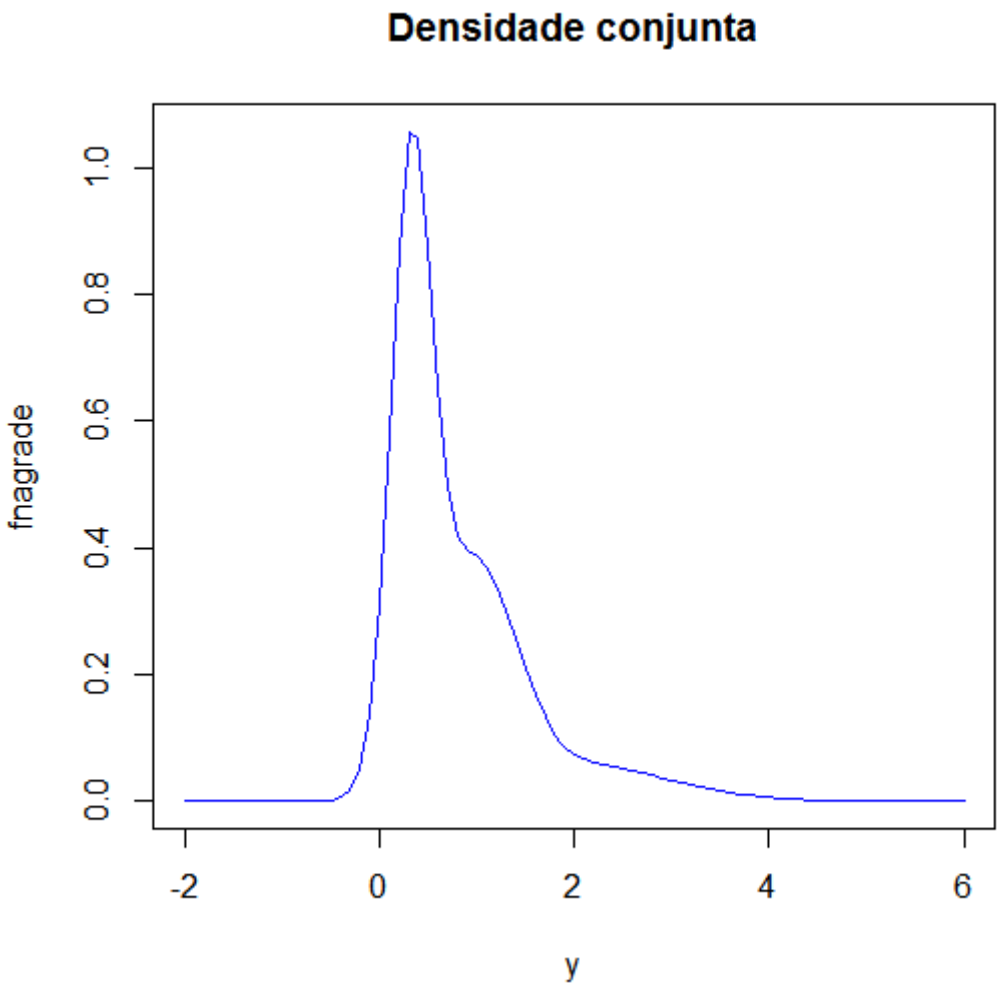

Figura 21 – Curva suave da densidade conjunta

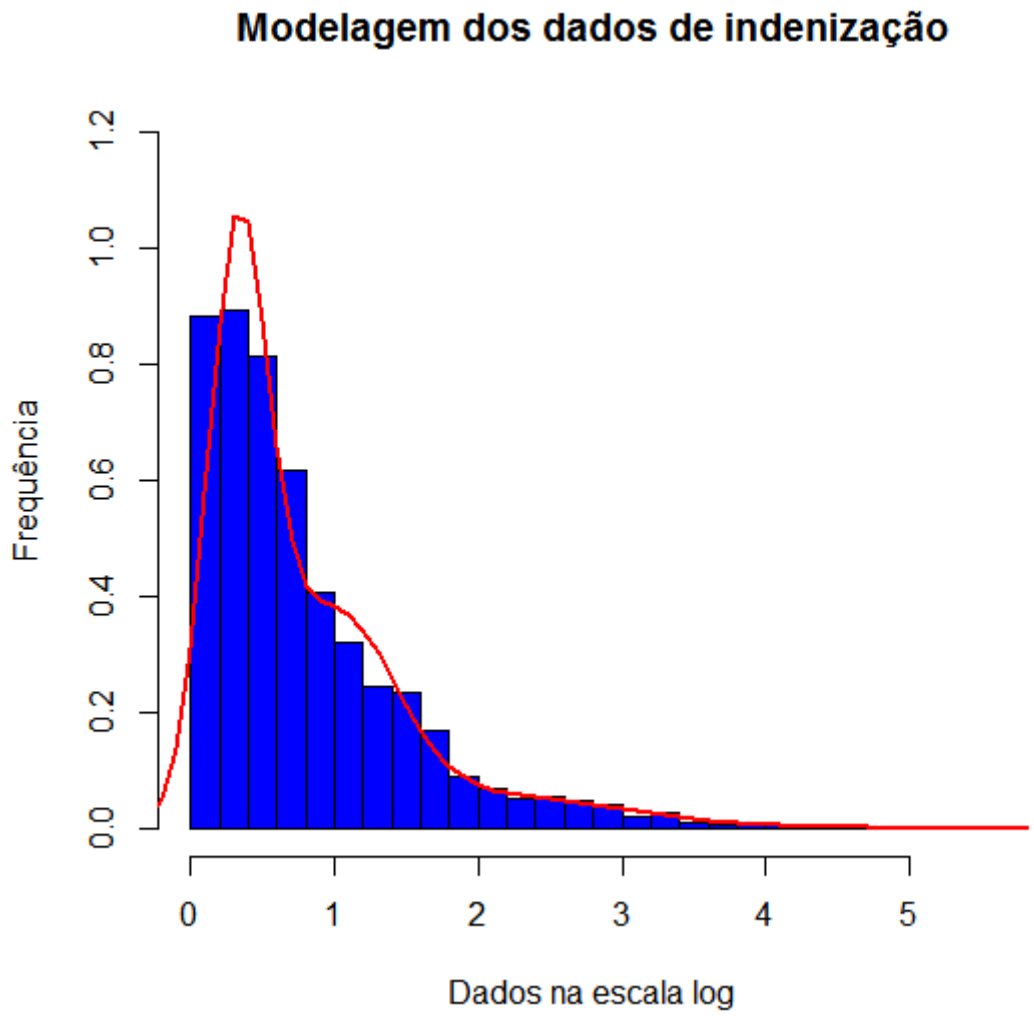

Figura 22 – Histograma dos dados reais com o modelo de mistura ajustado

Nota-se que a curva ajusta-se bem aos dados, o que indica que a modelagem através de uma mistura de três componentes Normais representa de forma eficaz o comportamento da variável sinistro agregado.

5. Conclusão e sugestões para trabalhos futuros

Estudos sobre a modelagem de dados de sinistro são essenciais no mercado atuarial e auxiliam a uma melhor performance do profissional responsável pelo setor de precificação.

A construção de modelos eficazes exige um conhecimento técnico apurado e são essenciais para uma precificação do seguro de forma justa e que agreguem os interesses da companhia.

A partir de uma motivação prática no processo de tarifação de seguros, o objetivo desta monografia foi estudar uma proposta de modelagem de dados de indenizações através de um modelo de mistura de distribuições Normais.

Nesta pesquisa, foram revisados alguns trabalhos publicados recentemente e alguns livros encontrados na literatura atuarial e estatística sobre inferência bayesiana para modelos de misturas finitas.

A teoria de misturas foi exposta e a questão da identificabilidade do conjunto paramétrico também foi estudada, pôde-se perceber que misturas de Normais univariadas e multivarias são genericamente identificáveis.

Para um número k conhecido de componentes, foi apresentado uma implementação no software estatístico R Project do Amostrador de Gibbs em uma modelagem por uma mistura de densidades Normais. Esta metodologia consiste, basicamente, em um passeio aleatório que atende as propriedades Markovianas.

A análise do banco de dados serviu de insumo para a presente pesquisa, esta consiste de 1500 pagamentos de indenizações referentes ao seguro de Responsabilidade Civil Geral americano e encontra-se disponível gratuitamente, o que possibilita a sua modelagem através de novas técnicas.

Pôde-se avaliar que apesar dos dados apresentarem uma assimetria à direita, após a implementação do Amostrador de Gibbs, diagnosticou-se empiricamente um comportamento convergente da cadeia de Markov.

Espera-se que esta revisão da literatura possa servir de base para estudos futuros e, como proposta para os outros trabalhos, sugere-se a adaptação do algoritmo exposto para mistura de outras densidades de probabilidades. Um exemplo é a modelagem através de Normais assimétricas por atenderem mais à característica de um histórico de sinistro, no qual sinistros de baixos custos tendem a acontecer com mais frequência do que sinistros de alto custo. Outro input é a utilização de dados brasileiros de sinistros agregados de outros tipos de seguros.

REFERÊNCIAS

ALVES, M. B. **Confiabilidade Estrutural por Simulação de Monte Carlo com Cadeias de Markov**. Universidade Federal do Rio de Janeiro, UFRJ, Brasil, 1999.

FERREIRA, PAULO P. **Modelos de Precificação e Ruína para Seguros de Curto Prazo**. Rio de Janeiro: Escola Nacional de Seguros – FUNENSEG, 2010.

FRUHWIRTH-SCHNATTER, S. **Finite Mixture and Markov Switching Models**. New York/Berlin/Heidelberg: Springer, 2006.

GAMERMAN E LOPES, D.; LOPES, H. F. **Markov Chain Monte Carlo**. 2nd. ed. London: Chapman & Hall, 2006.

LIU, Z. **Bayesian Mixture Models**. McMaster University. Open Access Dissertations and Theses, Paper 4499, 2010.

ROSS, SHELDON M. **Stochastic Processes**. 2nd ed. United States: Copyrighted Material, 1996.

SUSEP – Superintendência de Seguros Privados, **Processo nº 15414.001870/2005-24**. Dispões sobre o Seguro de Responsabilidade Civil Geral, Disponível em: < http://www2.susep.gov.br/bibliotecaweb/biblioteca.aspx>.

TEICHER, H. **Identifiability of finite mixtures. Annals of Mathematical Statistics** 34: 1265-1269, 1963.

YAKOWITZ, S. J. and SPRAGINS, J. D. **On the Identifiability of Finite Mixtures.** Annals of Mathematical Statistics 39: 209-214, 1968.

ANEXO – PROGRAMAÇÃO

```
#### Dados americanos de indenização do seguro de Responsabilidade Civil

dat<-read.table("loss.txt",header = TRUE,sep="", dec=".")
hist(log(dat[,2]), main='Distribuição dos dados de indenizações', xlab='Dados na
escala log', ylab='Frequência', col='blue', prob=T)
lines(density(log(dat[,2])), col="red", lwd=2)

density(log(dat[,2]))
ldat<-log(dat[,2])

#### Algoritmo - Amostrador de Gibbs

niter=10000

gibbsnorm<-function(dat, k=3, niter, alpha = 1.28, beta = 0.36*var(ldat),
                    lam=mean(ldat), tau=2.6/(diff(range(ldat)))^2,g=1)

{
  rigamma<-function(n,a,b){
    return(1/rgamma(n, shape=a, rate=b))
  }

  rdirichlet<- function(n,par){
    k=length(par)
    z=array(0, dim=c(n,k))
    s=array(0, dim=c(n,1))
    for (i in 1:k){
      z[,i]=rgamma(n, shape=par[i])
      s=s+z[,i]
    }
    for (i in 1:k){
      z[,i]=z[,i]/s
    }
    return(z)
  }

  n<-length(dat)
  mu<-rnorm(k,mean=mean(dat), sd=sd(dat))
  sig<-sd(dat)/k
  p <- rep(1/k, k)

  mixparam<-list(p=p, mu=mu, sig=sig)
  z<-rep(0,k)
  nj<-z
  sj<-z
  sj2<-z
  gibbsmu<-matrix(0,nrow=niter,ncol=k)
```

```
  gibbssig<-gibbsmu
  gibbsp<-gibbsmu

  for(i in 1:niter){
    #print(i)
    for (t in 1:n){
      prob<-mixparam$p*dnorm(dat[t], mean=mixparam$mu,
                             sd=mixparam$sig)
      z[t]<- sample(x=1:k, size=1,prob=prob)
    }
    for(j in 1:k){
      nj[j]<-sum(z==j)
      sj[j]<-sum(as.numeric(z==j)*dat)
    }
    repeat{
      gibbsmu[i,]<-rnorm(k, mean=(lam*tau+sj)/(nj+tau),
sd=sqrt(mixparam$sig^2/(tau+nj)))
      if(max(gibbsmu[i, ])<max(dat)&min(gibbsmu[i, ])>min(dat))
        break
    }
    mixparam$mu<-gibbsmu[i, ]
    for(j in 1:k){
      sj2[j]=sum(as.numeric(z==j)*(dat-mixparam$mu[j])^2)
    }
    gibbssig[i, ]<-sqrt(rigamma(k, alpha+0.5*(nj+1), beta+0.5*tau*(mixparam$mu-
lam)^2+0.5*sj2))
    mixparam$sig<-gibbssig[i,]
    gibbsp[i, ]<-rdirichlet(1,par=nj+g)
    mixparam$p<-gibbsp[i, ]
  }
  data.frame(p=gibbsp,mu=gibbsmu,sigma=gibbssig)
}

SS<-gibbsnorm(ldat,k=3, niter)
dim(SS)

#### Graficos:

## Pesos das componentes

plot(SS[2001:10000,1], lty=1, type="l", xlab='Dados', ylab=expression(pi[1]),
col='blue',main='Peso da 1ª componente')
abline(h=mean(SS[2001:10000,1]),col="red")

hist(SS[2001:10000,1], main="Histograma do peso da 1ª
componente",col='blue',xlab=expression(pi[1]),ylab='Densidade',prob=T)
abline(v=c(mean(SS[2001:10000,1]),quantile(SS[2001:10000,1],c(0.025,0.975))),col=
c("red","gold","gold"))
```

```
plot(SS[2001:10000,2], lty=1, type="l", xlab='Dados', ylab=expression(pi[2]),
col='green',main='Peso da 2ª componente')
abline(h=mean(SS[2001:10000,2]),col="red")

hist(SS[2001:10000,2],main="Histograma do peso da 2ª
componente",col='green',xlab=expression(pi[2]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,2]),quantile(SS[2001:10000,2],c(0.025,0.975))),col=
c("red","gold","gold"))


plot(SS[2001:10000,3], lty=1, type="l", xlab='Dados', ylab=expression(pi[3]),
col='purple',main='Peso da 3ª componente')
abline(h=mean(SS[2001:10000,3]),col="red")

hist(SS[2001:10000,3],main="Histograma do peso da 3ª
componente",col='purple',xlab=expression(pi[3]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,3]),quantile(SS[2001:10000,3],c(0.025,0.975))),col=
c("red","gold","gold"))


## Medias das componentes

plot(SS[2001:10000,4], lty=1, type="l", xlab='Dados', ylab=expression(m[1]),
col='blue',main='Média da 1ª componente')
abline(h=mean(SS[2001:10000,4]),col="red")

hist(SS[2001:10000,4], main="Histograma da média da 1ª
componente",col='blue',xlab=expression(m[1]),ylab='Densidade',prob=T)
abline(v=c(mean(SS[2001:10000,4]),quantile(SS[2001:10000,4],c(0.025,0.975))),col=
c("red","gold","gold"))


plot(SS[2001:10000,5], lty=1, type="l", xlab='Dados', ylab=expression(m[2]),
col='green',main='Média da 2ª componente')
abline(h=mean(SS[2001:10000,5]),col="red")

hist(SS[2001:10000,5],main="Histograma da média da 2ª
componente",col='green',xlab=expression(m[2]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,5]),quantile(SS[2001:10000,5],c(0.025,0.975))),col=
c("red","gold","gold"))


plot(SS[2001:10000,6], lty=1, type="l", xlab='Dados', ylab=expression(m[3]),
col='purple',main='Média da 3ª componente')
abline(h=mean(SS[2001:10000,6]),col="red")

hist(SS[2001:10000,6],main="Histograma da média da 3ª
componente",col='purple',xlab=expression(m[3]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,6]),quantile(SS[2001:10000,6],c(0.025,0.975))),col=
c("red","gold","gold"))
```

```
## Desvio Padrão

plot(SS[2001:10000,7], lty=1, type="l", xlab='Dados', ylab=expression(sigma[1]),
col='blue',main='Desvio Padrão da 1ª componente')
abline(h=mean(SS[2001:10000,7]),col="red")

hist(SS[2001:10000,7], main="Histograma do desvio padrão da 1ª
componente",col='blue',xlab=expression(sigma[1]),ylab='Densidade',prob=T)
abline(v=c(mean(SS[2001:10000,7]),quantile(SS[2001:10000,7],c(0.025,0.975))),col=
c("red","gold","gold"))


plot(SS[2001:10000,8], lty=1, type="l", xlab='Dados', ylab=expression(sigma[2]),
col='green',main='Desvio padão da 2ª componente')
abline(h=mean(SS[2001:10000,8]),col="red")

hist(SS[2001:10000,8],main="Histograma do desvio padrão da 2ª
componente",col='green',xlab=expression(sigma[2]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,8]),quantile(SS[2001:10000,8],c(0.025,0.975))),col=
c("red","gold","gold"))


plot(SS[2001:10000,9], lty=1, type="l", xlab='Dados', ylab=expression(sigma[3]),
col='purple',main='Desvio padrão da 3ª componente')
abline(h=mean(SS[2001:10000,9]),col="red")

hist(SS[2001:10000,9],main="Histograma do desvio padrão da 3ª
componente",col='purple',xlab=expression(sigma[3]),ylab='Densidade',prob=T,)
abline(v=c(mean(SS[2001:10000,9]),quantile(SS[2001:10000,9],c(0.025,0.975))),col=
c("red","gold","gold"))

#### Distribuição conjunta

p1=SS[2001:10000,1]
p2=SS[2001:10000,2]
p3=SS[2001:10000,3]
mu1=SS[2001:10000,4]
mu2=SS[2001:10000,5]
mu3=SS[2001:10000,6]
dp1=SS[2001:10000,7]
dp2=SS[2001:10000,8]
dp3=SS[2001:10000,9]

y<-seq(-2,6,0.1)
ngrade<-length(y)
nMC<-8000
f<-matrix(0,nrow=nMC,ncol=ngrade)
```

```
for(i in 1:nMC){
        for (j in 1:ngrade){
        f[i,j]<-
p1[i]*dnorm(y[j],mu1[i],dp1[i])+p2[i]*dnorm(y[j],mu2[i],dp2[i])+p3[i]*dnorm(y[j],mu3[i],d
p3[i])
        }
}

fnagrade<-apply(f,2,mean)
length(fnagrade)

plot(y,fnagrade, type="l", lty=1, col='blue', main='Densidade conjunta')

hist(log(dat[,2]), nclass=20, main='Modelagem dos dados de indenização',
xlab='Dados na escala log', ylab='Frequência', col='blue', prob=T,ylim=c(0,1.2))
lines(y,fnagrade, col="red", lwd=2)
```