UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

Márcio Augusto da Cruz Almeida

Uso do FBST no Teste de Homogeneidade em Misturas Finitas: Caso Normal e Poisson

Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2011

Márcio Augusto da Cruz Almeida

Uso do FBST no Teste de Homogeneidade em Misturas Finitas: Caso Normal e Poisson

Dissertação apresentada ao Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará como requisito parcial para a obtenção do grau de Mestre em Estatística.

Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2011

Márcio Augusto da Cruz Almeida

Uso do FBST no Teste de Homogeneidade em Misturas Finitas: Caso Normal e Poisson

Esta Dissertação foi julgada e aprovada, para a obtenção do grau de Mestre em Estatística, no Programa de de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará.

Belém, 29 de Março de 2011.

Prof. Dr. Giovany de Jesus Malcher Figueiredo,
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística da UFPA)

Banca Examinadora

Profa. Dra. Maria Regina Madruga Tavares Universidade Federal do Pará **Orientadora**	Prof. Dr. Valcir João da Cunha Farias Universidade Federal do Pará **Membro**
Profa. Dr. Héliton Ribeiro Tavares Universidade Federal do Pará **Membro**	Prof. Dr. Celso Rômulo Barbosa Cabral Universidade Federal do Amazonas **Membro**

À Deus pelo dom da minha vida.

Agradecimentos

⋆ À Deus, por ter me amado tanto que entregou seu Filho, Jesus Cristo, para morrer por mim. Hoje eu vivo e pude fazer este trabalho por causa de seu sacrifício na Cruz do Calvário.

⋆ Aos meus pais, Eremita Pantoja e Raimundo Elpídio, que me criaram com carinho, amor e educação, dando força e sempre me apoiando nos meus estudos.

⋆ À Universidade Federal do Pará (UFPA) pela oportunidade de concluir um curso de pós-graduação.

⋆ A Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro destinado à este trabalho.

⋆ Ao Programa de Pós-Graduação em Matemática e Estatística (PPGME), representado pelo Prof. Dr. Giovany Figueiredo;.

⋆ À minha orientadora, Profa. Dra. Regina Tavares, não somente pela sua competência e paciência para comigo, mas também pela sua amizade e ampla capacidade de entender diversos problemas e situações que me fizeram crescer profissionalmente ao final deste trabalho.

⋆ Aos meus amigos Rafael e Vanessa, pela suas amizades e extrema ajuda nos momentos mais difíceis que passei durante a elaboração deste trabalho.

⋆ À minha namorada Lilian Cravo, pelo seu amor, compreensão e orações que me fortaleceram e me fizeram perseverar nos momentos difíceis.

⋆ Aos meus professores durante as disciplinas lecionadas, que me capacitaram e ampliaram a minha mente com as mais complicadas demonstrações já encontradas.

⋆ Aos meus familiares, que sempre acreditaram no meu potencial, dando força e ânimo para sempre seguir em frente.

⋆ Às amizades de meus amigos Corecha, Fabricío e Francisco, que ganhei durante o mestrado, assim como Paulinho, Daniel, Gil, Alex, Jaciane, Gleycy, Juliana, Mariane e Samara.

⋆ Aos meus amigos do Alfa e Ômega, pelas reuniões de quarta-feira, orações e apoio que abençoaram a minha.

⋆ À todos que, de alguma forma, contribuíram para a conclusão deste trabalho .

Resumo

ALMEIDA, Márcio Augusto da Cruz. Uso do FBST no Teste de Homogeneidade em Misturas Finitas: Caso Normal e Poisson. 2011. Dissertação. (Mestrado em Estatística e Matemática), PPGME, UFPA, Belém, Pará, Brasil.

Neste trabalho a medida de evidência proposta por Pereira e Stern (1999) foi usada para testar a hipótese de homogeneidade em modelos de misturas finitas com número de componentes fixado. O estudo foi realizado em misturas finitas de distribuições Poisson e de normais. Através de dados simulados de modelos de misturas, para os casos Poisson e Normal, uma medida de evidência em favor da hipótese nula de homogeneidade foi obtida, com base no procedimento FBST (*Full Bayesian Significance Test*) proposto em Pereira e Stern (1999). Os resultados obtidos indicaram um bom desempenho do FBST para os dois casos estudados, fornecendo valores pequenos para a medida de evidência em favor da hipótese nula. Os resultados foram validados a partir do cálculo do poder empírico do teste, proposto em Madruga et al. (2005), que indicou um aumento no poder quando a diferença entre os parâmetros estudados era maior, ou seja, quando as componentes da mistura eram mais distintas.

Palavras-chave: Poder empírico, mistura finita, medida de evidência.

Abstract

ALMEIDA, Márcio Augusto da Cruz. Use of FBST in the Homogeneity Test in Finite Misture: Case Normal and Poisson. 2011. Dissertation. (Master's Degree in Mathematics and Statistics), PPGME, UFPA, Belém, Pará, Brazil.

In this work the evidence of measure proposed by Pereira and Stern (1999) was used to test the hypothesis of homogeneity in finite mixture models with fixed number of components. The study was performed in finite mixtures of normal and Poisson distributions. Through simulated data of mixture models for normal and Poisson cases, a measure of evidence in favor of the null hypothesis of homogeneity was obtained using the procedure FBST (Full Bayesian Significance Test) proposed by Pereira and Stern (1999). The results indicated a good performance of FBST in each cases studied, giving small values to the extent of evidence in favor of the null hypothesis. The results were validated by calculating the power of empirical test, proposed in Madruga *et al.* (2005), which indicated an increase on the power when the difference between the parameters studied was higher, i.e, when the components of the mixture were more distinct.

Words Keys: Empirical power, finite mixture, measure of evidence.

Índice

Resumo **vi**

Abstract **vii**

Lista de Tabelas **x**

Lista de Figuras **xi**

1 Introdução **1**
1.1 Aspectos Gerais 1
1.2 Objetivos 4
1.3 Estrutura do Trabalho 4

2 Modelos de Mistura **5**
2.1 Introdução 5
2.2 Modelos de Mistura sob a Abordagem Clássica 9
2.3 Inferência Bayesiana 10
2.4 Abordagem Bayesiana para Modelos de Mistura Finita de Poisson 12
2.5 Abordagem Bayesiana para Modelos de Misturas Finita de Normais 15
2.6 Aspectos Computacionais 17

3 Full Bayesian Siginificance Test **19**
3.1 Introdução 19
3.2 Definição 20
3.3 Regra de Decisão e Validação no FBST 21

4 Aplicações e Resultados **23**
4.1 Teste de homogeneidade na mistura de duas distribuições Poisson 23
4.1.1 Dados Simulados 25
4.1.2 Aplicação a Dados Reais 30
4.2 Teste de Homogeneidade na Mistura de Duas Distribuições Normais 31
4.2.1 Dados Simulados 33
4.2.2 Aplicação a Dados Reais 37

5 Considerações Finais **39**
5.1 Considerações Finais 39
5.2 Trabalhos Futuros 40

Referências Bibliográficas **41**

Lista de Tabelas

4.1 Valores da evidência e dos parâmetros estimados segundo o tamanho amostral (n) e o número de amostras geradas (m). 26
4.2 Poder Empírico do Teste FBST, Segundo a Diferença $\lambda_2 - \lambda_1$ 27
4.3 Quantidade de Clientes por Número de Parcelas em Atraso em uma Instituição Financeira da Espanha em 1990. 30
4.4 Estimativas dos Parâmetros Obtidas no Trabalho Anterior e no Atual, e a Medida de Evidência em Favor da Hipótese de Homogeneidade. 31
4.5 Valores da Evidência e dos Parâmetros Estimados Segundo o Tamanho Amostral (n) e o Número de Amostras Geradas (m). 34
4.6 Poder Empírico do Teste FBST, Segundo a Diferença $\mu_2 - \mu_1$. 35
4.7 Estimativas dos Parâmetros Obtidas no Trabalho Anterior e no Atual, e a Medida de Evidência em Favor da Hipótese de Homogeneidade. 38

Lista de Figuras

2.1 Representação de uma Mistura de Distribuiçoes Normais descrito em Linsday(1995). 7

4.1 Evidência Média, Segundo a Diferença $\lambda_2 - \lambda_1$. 28
4.2 Convergência da Evidência para $\lambda_2 - \lambda_1 = 2$. 28
4.3 Convergência da Evidência para $\lambda_2 - \lambda_1 = 4$. 28
4.4 Convergência da Evidência para $\lambda_2 - \lambda_1 = 5$. 29
4.5 Convergência da Evidência para $\lambda_2 - \lambda_1 = 6$. 29
4.6 Convergência da Evidência para $\lambda_2 - \lambda_1 = 8$. 29
4.7 Convergência da Evidência para $\lambda_2 - \lambda_1 = 10$. 30
4.8 Evidência Média, Segundo a Diferença $\mu_2 - \mu_1$. 35
4.9 Convergência da Evidência para $\mu_2 - \mu_1 = 2$. 35
4.10 Convergência da Evidência para $\mu_2 - \mu_1 = 4$. 36
4.11 Convergência da Evidência para $\mu_2 - \mu_1 = 6$. 36
4.12 Convergência da Evidência para $\mu_2 - \mu_1 = 8$. 36

Capítulo 1

Introdução

1.1 Aspectos Gerais

Estudos sobre modelos de mistura e a estimação dos parâmetros envolvidos no problema constituem uma abordagem amplamente utilizada, devido a sua vasta área de aplicação. Para Lindsay (1995), o modelo de mistura surge quando uma amostra aleatória é gerada por várias subpopulações, com um modelo probabilístico representado por uma combinação linear dos modelos associados a cada componente. Isto ocorre comumente em situações onde a variável aleatória de interesse é observada em várias condições distintas, por exemplo: quando os dados estão sendo observados em animais de diferentes espécies, ou em regiões geográficas distintas, ou em diferentes gêneros, etc. Nestes casos, geralmente observa-se uma grande variabilidade nos dados observados, indicando a não adequação de um único modelo de probabilidade para ajustar os dados.

Assim, um modelo compreendendo uma mistura de distribuições de probabilidades surge como uma alternativa eficiente para modelar os dados. Segundo Chen e Chen (1998) a idéia básica é determinar se os dados provenientes de uma amostra são de uma população homogênea ou heterogênea. Para isso, diversos métodos estatísticos têm sido utilizados para estimar os parâmetros das distribuições envolvidas no problema, bem como o número de componentes envolvidas.

Embora os estudos relacionados com distribuições provenientes de misturas estejam concentrados em diferenciar as propriedades das subpopulações envolvidas, os "modelos de mistura" são também usados para classificação de novas observações, a partir das inferências resultantes sobre a população estudada.

Do ponto de vista estatístico clássico, há problemas associados aos estimadores de máxima verossimilhança como apontado em Chen et al. (2001), que estão relacionados a:

1. Distribuição assintótica destes estimadores;

2. Condições de regularidade não inclusas no modelo de mistura;

3. Estimadores são viesados para os parâmetros do modelo.

Além disso, Dellaportas et al. (1997) considera a dificuldade em encontrar analiticamente os estimadores dos parâmetros do modelo, uma vez que a função de verossimilhança representa uma soma de vários componentes, o que dificulta o problema para casos de modelos com várias subpopulações.

A abordagem Bayesiana para a estimação dos parâmetros no modelo de mistura não havia sido bem desenvolvida até o início da década de 90, devido principalmente a obstáculos computacionais. Diebolt e Robert (1994) conseguiram contornar o problema no estudo de misturas finitas, sob o ponto de vista bayesiano, considerando uma distribuição a priori para os parâmetros desconhecidos do modelo e introduzindo no estudo variáveis latentes para identificar a subpopulação da variável aleatória de interesse. Com isso, eles conseguiram obter uma distribuição posterior conjunta para os parâmetros, a partir da informação a priori e da verossimilhança dos dados. Com a inserção das variáveis latentes foi possível gerar amostras da distribuição posterior conjunta dos parâmetros do modelo de mistura utilizando o Amostrador de Gibbs, algoritmo de simulação integrante dos métodos *Monte Carlo via Cadeia de Markov* (MCMC).

No estudo de misturas finitas, ou seja, quando o número de componentes na mistura é finito, um dos interesses é verificar se os dados provêm de uma população homogênea ou heterogênea. Entretanto, do ponto de vista clássico, o teste da razão de verossimilhanças generalizada (TRVG) não pode ser utilizado para testar homogeneidade em um modelo de mistura, uma vez que apresenta problemas na distribuição assintótica da estatística do teste por conta do não-cumprimento das condições de regularidade (Chen et al. (2001)).

Lauretto (2007) afirma que em modelos de mistura o estimador do número de componentes na mistura, via máxima verossimilhança ou máxima posterior tendem a privilegiar modelos desnecessariamente complexos, isto é, com muitas componentes. Em outro ponto de vista, o modelo de mistura é visto como uma representação conveniente da função densidade de probabilidade, quando não se encontram distribuições de probabilidade capazes de se ajustar suficientemente bem aos dados, assumindo-se assim que o problema é basicamente estimar os parâmetros associados à mistura que melhor se ajusta aos dados.

Neste trabalho, o teste de homogeneidade para misturas finitas baseia-se na medida de evidência desenvolvida por Pereira e Stern (1999). Eles obtiveram a evidência em favor de uma hipótese nula precisa (hipótese com dimensão menor que a dimensão do espaço paramétrico), baseada no cálculo da probabilidade posterior da região HPD (*Highest Posterior Density*) do espaço paramétrico tangente ao conjunto que define a hipótese nula. A proposta é um teste genuinamente bayesiano e tem a vantagem de ser realizado sem a necessidade de introduzir uma massa de probabilidade a priori positiva no espaço definido pela hipótese precisa, apresentando-se assim como uma alternativa aos testes de significâncias usuais, e denominado FBST (*Full Bayesian Significance Test*).

Madruga (2002), através do FBST, conseguiu obter bons resultados para alguns problemas clássicos de teste de hipóteses e comparou com os resultados de testes de significância alternativos. Lauretto (2007) desenvolveu duas propostas para resolução de problemas distintos, o problema de classificação não supervisionada e o problema de modelos separados, ambos baseados na aplicação do teste de significância FBST em modelos de misturas, onde os resultados convergiram mais rapidamente para a decisão correta com o FBST.

Neste contexto, este trabalho surge com a proposta de usar o FBST para testar homogeneidade no modelo de mistura finita, nos casos de dados gerados da distribuição Poisson e da distribuição Normal.

1.2 Objetivos

Os objetivos deste trabalho são relacionados a seguir:

1. Apresentar o processo de estimação bayesiana dos parâmetros no modelo de mistura finita de populações para o caso Poisson e Normal;

2. Usar o FBST, proposto por Pereira e Stern (1999), para testar homogeneidade no modelo de mistura finita de populações para o caso Poisson e Normal;

3. Validar o uso do FBST através do cálculo do poder empírico do teste nos modelos de mistura finita estudados.

1.3 Estrutura do Trabalho

Essa monografia encontra-se dividida em 5 capítulos:

1. Capítulo 1: Refere-se à introdução do trabalho, onde são abordados os aspectos gerais, justificativa e importância do trabalho, os objetivos, bem como a estrutura do trabalho;

2. Capítulo 2: Apresenta o modelo de mistura finita, principais conceitos e aspectos inferências associados nas abordagens clássica e Bayesiana, e sua especificação para o caso Poisson e o caso normal;

3. Capítulo 3: Apresenta uma abordagem do FBST e sua utilização e para o problema de modelos de mistura finita;

4. Capítulo 4: Utiliza o FBST na análise de dados simulados e reais, para modelos de mistura das distribuições Poisson e normal;

5. Capítulo 5: Apresenta as considerações finais e recomendações para trabalhos futuros.

Capítulo 2

Modelos de Mistura

2.1 Introdução

O estudo do comportamento e característica de uma amostra é fundamental para se fazer inferências adequadas sobre a população de interesse. Uma etapa fundamental na análise estatística é a Estatística Indutiva, que tem por objetivo obter e generalizar conclusões para a população a partir de uma amostra, e a partir do cálculo de probabilidades tirar conclusões que, muitas das vezes, está sempre associada a um grau de incerteza e, consequentemente, a uma probabilidade de erro.

Magalhães(2008) mostra a importância da amostragem, onde o conhecimento do comportamento é fundamental para se coletar a amostra:

***i*)** Para populações homogêneas, a amostragem pode ser casual simples ou sistemática;

***ii*)** Para populações heterogêneas, a amostragem pode ser proporcional estratificada ou por conglomerados.

Em muitas situações práticas, a amostra é coletada a partir da falsa suposição de homogeneidade da população alvo no estudo, gerando dados com alta dispersão e dificultando o processo de inferência nos parâmetros populacionais. Neste trabalho, o interesse fundamental é determinar o comportamento da população, verificando sua homogeneidade ou não a partir da amostra estudada, a fim de obter resultados inferenciais mais adequados.

Estimativas obtidas a partir de modelos que envolvem misturas de distribuições são de grande utilidade, e trazem importante contribuição da análise estatística em problemas de reconhecimento de padrões, mineração de dados e outras aplicações. Lindsay (1995) diz que o modo mais simples e natural de entender os modelos de mistura é quando se percebe que uma amostra de uma população é gerada, de fato, por várias subpopulações, o que ele chama de componentes da

população. Nestes casos, o número de componentes é representado por uma variável aleatória K, que pode ter seu valor conhecido ou não. Se K for finita tem-se o problema de misturas finitas. Neste trabalho será considerado apenas o caso em que K é conhecido..

Para ilustrar algumas caracteristicas fundamentais do problema é apresentado um exemplo simples e examinar as densidades de mistura que possam surgir. Suponha que se tenha uma população de animais composta por dois tipos de individuos, componente 1 (gênero masculino) e componente 2 (gênero feminino), e a característica de interesse X é o comprimento do animal, considerada normalmente distribuida em ambas as componentes, quando consideradas isoladamente. Considere, por simplicidade, que os dois grupos possuem a mesma variancia σ^2 para a variável X, mas que eles possuem diferentes médias, μ_1 e μ_2, respectivamente. Assume-se, também, que os animais do sexo masculino possuem média μ_1 maior que aqueles do sexo feminino μ_2. Seja π a proporção dos componentes 1 na população e $(1-\pi)$ a proporção da componente 2. Se as amostras das componentes é retirada sem rótulo de gênero, então a distribuição resultante do comprimento será uma mistura de duas normais, que será representada por:

$$\pi N(\mu_1, \sigma^2) + (1-\pi)N(\mu_2, \sigma^2) \tag{2.1}$$

aqui usa-se o símbolo $N(\mu, \sigma^2)$ para representar a distribuição normal com média μ e variância σ^2.

Se as componentes têm a mesma proporção de animais $\pi = 0,5$ e as médias representam quatro vezes mais o desvio padrão, então o modelo pode ser representado pela Figura 2.1. Neste caso, a distribuição da mistura das densidades é conhecida como bimodal. No entanto, apesar de se conhecer o valor da variável comprimento, não se pode saber se a informacão vem da população do sexo masculino ou do sexo feminino.

Outra situação mais complicada acontece quando os parâmetros da população são desconhecidos e diferentes, e a quantidade de componentes envolvidas também se torna grande e em proporções diferentes. Com isso a quantidade de parâmetros a serem estimados torna-se bastante complicado, e muitas das vezes sem solução analítica, onde muitos autores têm buscado métodos de integração, otimização e maximização numérica.

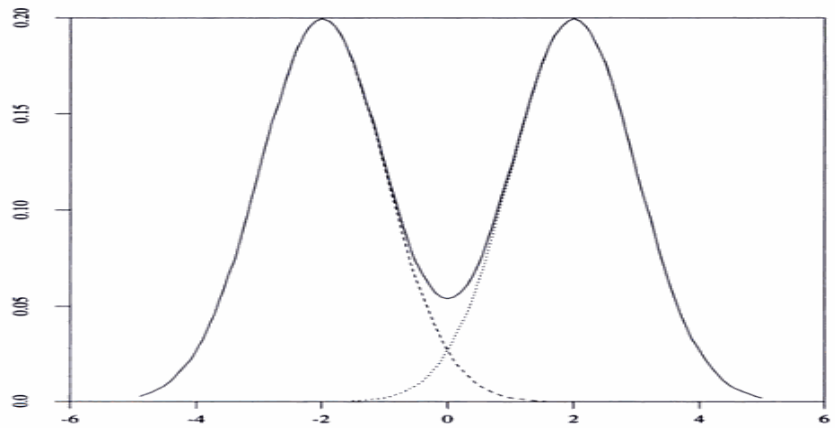

Figura 2.1 *Representação de uma Mistura de Distribuiçoes Normais descrito em Linsday(1995).*

De acordo com Chen *et al.* (2001), considere $\{f(x,\theta) : \theta \epsilon \Theta\}$ uma família paramétrica de funções densidade de probabilidade (FDPs). Então, coleta-se uma amostra aleatória $X_1, ..., X_n$ de uma população que é formada pela seguinte fdp

$$(1-\pi)f(x,\theta_1) + \pi f(x,\theta_2) \tag{2.2}$$

e considere também que $\theta_1 \leq \theta_2 \in \Theta$ e $0 \leq \pi \leq 1$. Neste caso tem-se um modelo de mistura de distribuições, onde:

i) $f(x,\theta_1)$ representa a fdp da subpopulação X com parâmetro θ_1;

ii) $f(x,\theta_2)$ representa a fdp da subpopulação X com parâmetro θ_2;

iii) π e $(1-\pi)$ representam as proporções das subpopulações.

Para Dellaportas *et al.* (1997), os dois primeiros termos representados por θ_1 e θ_2 são os parâmetros da mistura, enquanto que π é chamada de proporção da mistura. A fdp de um modelo de mistura, de uma maneira geral, pode ser representado da seguinte forma:

$$f(x) \quad = \quad \sum_{j=1}^{k} \pi_j f(x|\theta_j) \tag{2.3}$$

em que $\sum_{j=1}^{k} \pi_j = 1$. Neste caso, X é dita ter uma densidade de k misturas finitas.

Considerando a situação em que têm-se duas populações envolvidas, as hipóteses para o teste de homogeneidade são:

$$H_0 : \theta_1 = \theta_2 \quad \text{(ou, equivalententemente, } \pi = 0 \text{ ou } \pi = 1) \text{ e}$$
$$H_1 : \theta_1 \neq \theta_2 \quad \text{(ou, equivalententemente, } \pi \neq 0 \text{ ou } \pi \neq 1).$$

Muitos trabalhos têm sido desenvolvidos para solucionar este tipo de situação, e um dos testes mais realizados na literatura é o da razão de verossimilhanças generalizada (Bolfarine e Sandoval, 2002). O Teste da razão de verossimilhanças generalizada (TRVG) segue a sua forma usual de comparar os valores maximizados do logaritmo da função de verossimilhança $L(\theta, x)$ sem a restrição da hipótese nula H_0, e sob H_0, se a diferença é grande, então H_0 é rejeitada.

Chen e Kalbfleisch (2004) afirmam que obter resultados satisfatórios usando o TRVG é frequentemente mais difícil do que se poderia esperar, isso porque os modelos de mistura finita pertencem a uma classe de modelos, dita não-regular e, como consequência, muitos resultados assintóticos clássicos não são aplicáveis. Muitos pesquisadores têm tentado entender as propriedades de convergência para grandes amostras com relação à análise de modelos de mistura finita. Chen e Kalbfleisch (2001), Chen (1998) e Chen *et al.* (2001, 2002) propuseram uma modificação nas funcão de verossimilhança, obtendo resultados expressivos para os casos mais simples. Já McLachlan e Basford (1988) associaram o problema com a análise de cluster para determinar as populações envolvidas no estudo. Qin *et al.* (2009) investigaram as propriedades assintóticas da estatística do teste da razão de verossimilhanças para testar homogeneidade em modelos de mistura normal, com médias e variâncias desconhecidas. Min (1998) analisou a distribuição da razão de verossimilhanças da hipótese nula adotando uma única amostra gama com formato e escala desconhecidos contra a hipótese alternativa de uma amostra de duas gamas, cada uma com variâncias desconhecidas e iguais.

2.2 Modelos de Mistura sob a Abordagem Clássica

Seja $X_1, ..., X_n$ uma amostra aleatória de uma mistura de distribuições, conforme apresentado a Equação (2.3). A função de verossimilhança para os dados observados $\mathbf{X} = (x_1, ..., x_n)$, com $\theta = (\theta_1, ..., \theta_n)$ e $\pi = (\pi_1, ..., \pi_n)$, é escrita como

$$L(\pi, \theta, \mathbf{x}) \quad = \quad \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(x_i|\theta_j). \tag{2.4}$$

Muitos trabalhos foram desenvolvidos de modo a encontrar estimadores de máxima verossimilhança para π_j e θ_j, $j = 1, 2, ..., k$, como pode ser visto em Chen e Kalbfleisch (2004) e Chen (1998). Qin *et al* (2009) encontraram estimadores de máxima verossimilhança para π_j e θ_j, a partir da função em (2.4), tomando-se o logaritmo da verossimilhança:

$$l(\pi, \theta, \mathbf{x}) \quad = \quad log \left[\prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(x_i|\theta_j) \right] = \tag{2.5}$$

$$= \quad \sum_{i=1}^{n} log \left[\sum_{j=1}^{k} \pi_j f(x_i|\theta_j) \right] = \sum_{i=1}^{n} log \left[\pi_1 f(x_i|\theta_1) + \ldots + \pi_k f(x_n|\theta_k) \right].$$

Muitos trabalhos foram desenvolvidos de modo a encontrar estimadores para k, π_j e θ_j, $j = 1, 2, ..., k$ (CHEN e KALBFLEISCH, 2004: CHEN, 1998; QIN *et al.*, 2009; TITTERINGTON *et al.*, 1985; MCLACHLAN e BASFORD, 1988; LINDSAY, 1985), eles incluem métodos de máxima verossimilhança, momentos e distância mínima. Quando o valor de k é conhecido, uma abordagem via algoritmo EM pode ser usada para obter estimadores de máxima verossimilhança (HASSELBLAD, 1969).

Suponha, no caso $k = 2$, que $\hat{\pi}$, $\hat{\theta_1}$ e $\hat{\theta_2}$ sejam os estimadores de máxima verossimilhança para π, θ_1 e θ_2 sob a hipótese alternativa, considerando a Equação (2.2), e que $\hat{\theta}$ seja o estimador de máxima verossimilhança sob a hipótese nula. Logo, a estatística do TRVG pode ser escrita como:

$$R_n \quad = \quad r_n(\hat{\pi}, \hat{\theta_1}, \hat{\theta_2}, \hat{\theta}) = 2\{l(\hat{\pi}, \hat{\theta_1}, \hat{\theta_2}) - l(0, \hat{\theta}, \hat{\theta})\}. \tag{2.6}$$

Assim, o teste rejeita H_0 se R_n é muito grande. A distribuição assintótica nula de R_n é usada para determinar um valor crítico para o teste ou um nível descritivo (p). Um problema encontrado por muitos autores foi obter estatísticas consistentes para os parametros desconhecidos.

Hartigan (1985) mostra que a estatística do TRVG para verificar a homogeneidade no modelo de mistura tende para o infinito. Chen e Chen (1998) verificaram um comportamento divergente da distribuição assintótica sob a hipotese H_0 quando se aplicava o TRVG. Chen *et al.* (2001) apresentaram uma proposta de modificação na log-verossimilhança e consequentemente na estatística R_n, acrescentando mais um parâmetro no modelo de modo que o resultado aproxima-se mais do resultado esperado. Chen e Kalbfleisch (2004) aplicaram a proposta de modificação para o caso da hipótese alternativa apresentar uma mistura de distribuições normais com variâncias iguais e desconhecidas.

Segundo Li (2007), em muitos casos estes estimadores não existem devido à função de verossimilhança tornar-se analiticamente difícil de ser desenvolvida de modo a calcular os estimadores de máxima verossimilhança para os modelos de mistura finita. Uma série de algoritmos numéricos têm sido desenvolvidos para maximizar a função log - verossimilhança.

Nos últimos anos, uma abordagem bayesiana para modelos de mistura finita tem sido utilizada. Diebolt e Robert (1994), Richardson e Green (1997) foram os primeiros a desenvolver a metodologia sob o ponto de vista bayesiano. Eles fizeram o uso do método de Monte Carlo via Cadeia de Markov (MCMC) para estimar os parâmetros desconhecidos a partir das distribuições posteriores condicionais completas. Para se introduzir a abordagem bayesiana é necessário uma breve aprsentação da metodologia bayesiana.

2.3 Inferência Bayesiana

Os métodos bayesianos constituem uma alternativa aos métodos clássicos de inferência. Uma das principais diferenças em relação à metodologia clássica é que na inferência bayesiana é permitida a incorporação da informação *a priori* sobre os parâmetros desconhecidos do modelo e, ao contrário dos métodos clássicos, os métodos bayesianos consideram estes parâmetros como variáveis aleatórias, associando a eles uma distribuição de probabilidade.

Suponha que se deseja estimar o valor do parâmetro desconhecido . Segundo Ehlers (2003), a informação a *priori* que se tem sobre este parâmetro é de fundamental importância. Do ponto de vista bayesiano a incerteza sobre o valor desconhecido pode ser quantificada a partir de uma distribuição de probabilidade definida em um espaço paramétrico, representado por Θ. Esta

informação a *priori* é representada em termos de uma distribuição de probabilidade chamada de distribuição a *priori*.

Para Paulino *et al.* (2003) a informação inicial depende apenas do conhecimento do pesquisador sobre o problema estudado, logo, pode-se ter diferentes distribuições (ou modelos) a *priori* associadas ao parâmetro θ, representadas por $p(\theta)$, e definidas em Θ. Assim, coleta-se uma amostra aleatória da variável de interesse, onde a distribuição amostral pode ser representada por $p(X|\theta)$ ou $L(\theta; X)$, denominada de função de verossimilhança, e combina-se as informações a *priori* e a verossimilhança a partir do Teorema de Bayes, a fim de encontrar uma distribuição a *posteriori* para θ, dada por

$$p(\theta|X) \quad = \quad \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)}, \tag{2.7}$$

em que $\dfrac{1}{p(x)}$ é uma constante normalizadora da distribuição $p(\theta|X)$.

Assim, a distribuição *posteriori* representa um modelo com informação sobre θ, atualizado pelos dados. A distribuição *posteriori* é de suma importância na abordagem bayesiana, pois é ela que será utilizada para a realização de todas as inferências, como estimação pontual, construção de regiões de credibilidade ou intervalos HPD (*Highest Posterior Density*) e teste de hipóteses. Diferentemente da abordagem inferencial clássica, na abordagem bayesiana considera-se que θ é uma variável aleatória e a observação amostral é fixa.

Com o conhecimento prévio sobre o parâmetro θ, pode-se construir uma família paramétrica de densidades, de modo que as distribuições a *priori* e a *posteriori* pertençam à mesma classe de distribuições, isto é, sejam conjugadas. Neste caso, a distribuição posterior tem a mesma distribuição da *priori*, com os hiperparâmetros da distribuição sendo atualizados com a informação que a amostra (função de verossimilhança) contém. Dentre algumas distribuições que apresentam conjugação, tem-se o exemplo da distribuição Beta, que é conjugada natural da família Bernoulli. Para maiores detalhes sobre esta e outras classes de distribuições a *priori* ver Paulino *et al.* (2003).

Em algumas situações o pesquisador tem pouca informação sobre o parâmetro desconhecido, ou até mesmo nenhuma informação, tanto para o parâmetro θ quanto para a forma da distribuição. Nestes casos, é comum utilizar as chamadas distribuições a *priori* nulas ou não-informativas. Neste caso, espera-se que a informação contida na amostra seja dominante no

processo de inferência. Para esses casos, representa-se a distribuição a *priori* por uma constante. Alguns métodos foram desenvolvidos de modo a representar o máximo possível a informação a *priori*, por mais que ela seja vaga ou nula, dentre eles pode-se citar o método de Bayes-Laplace, Box-Tiao e a *priori* não informativa de Jeffreys, ver Paulino *et al.* (2003).

2.4 Abordagem Bayesiana para Modelos de Mistura Finita de Poisson

Conforme foi visto anteriormente, Diebolt e Robert (1994) e Richardson e Green (1997) foram alguns autores que iniciaram o estudo de modelos de mistura sob uma abordagem bayesiana, obtendo estimadores dos parâmetros a paritr das distribuições condicionais completas e do amostrador de Gibbs. Desde então, diversos trabalhos surgiram, dentre eles pode-se citar Dellaportas *et al.* (1997), em que analisaram sistemas de crédito financeiros (Credit Scoring) através da mistura de modelos Poisson; Pissini (2006) aplicou a abordagem em meta-análise a partir dos estimadores obtidos da mistura de distribuições normais; Lauretto (2007) utilizou misturas normais multivariada e a distribuição de Dirichlet *posteriori* para selecionar modelos.

De acordo com a Equação (2.3) e considerando o caso de uma mistura finita de modelos Poisson (Dellaportas *et al.*, 1997), a função de probabilidade pode ser escrita para k componentes na mistura como

$$f(x|\lambda,\pi) \quad = \quad \sum_{j=1}^{k} \pi_j \frac{e^{-\lambda_j}\lambda_j^x}{x!}, \tag{2.8}$$

com $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_k$ para garantir a identificabilidade do modelo. Sejam os dados $\mathbf{x}$ distribuídos conforme a Equação (2.7). Logo, o vetor de parâmetros a serem estimados é representado por $\theta = (\pi, \lambda)$ com $\pi = (\pi_1, \ldots, \pi_k)$ e $\lambda = (\lambda_1, \ldots, \lambda_k)$. Assim, tem-se que a distribuição condicional $f(x_i|\lambda_j)$ é Poisson com parâmetro λ_j.

Sob o ponto de vista bayesiano, o vetor de parâmetros desconhecido requer uma distribuição de probabilidade a *priori*. A função de verossimilhança é a mesma vista na Equação (2.4). Com isso, a distribuição a *posteriori* para o vetor de parâmetros é dado por:

$$p(\theta|x) \propto L(\theta, x) \times p(\theta). \tag{2.9}$$

Buscando solucionar os problemas apontados com relação à função de verossimilhança, Dempst

et al. (1997) iniciaram os estudos considerando dados incompletos na função de verossimilhança. Posteriormente, Diebolt e Robert (1994) introduziram nestes estudos uma abordagem bayesiana, introduzindo uma variável latente *Z*. Com isso, conseguiu-se simplificar os cálculos da distribuição a *posteriori* e obter distribuições condicionais completas, necessárias para o algoritmo de Gibbs. As variáveis latentes são usadas para classificar as observações em relação aos componentes, além de simplificar as condicionais a partir da eliminação do somatório da função de verossimilhança em (2.4).

A variável latente Z é usada para representar modelo de mistura em termos de falta de dados (Dempst *et al.*, 1997), de modo que cada Z_{ij} pode assumir apenas dois valores ($Z_{ij} = 0$ ou $Z_{ij} = 1$), onde $Z_{ij} = 1$ indica que a observação x_i pertence à componente j da mistura, e $Z_{ij} = 0$ caso contrário.

Neste trabalho, será considerado o caso particular em que *k=2*, ou seja, uma mistura com dois componentes. Neste caso, tem-se um vetor $Z_{ij} = (Z_{i1}, Z_{i2})$, $i = 1, 2, \ldots, n$, para cada observação com $Z_{i1} \sim$ Bernoulli (p_{i1}), onde cada p_{i1} representa a probabilidade da i-ésima observação pertencer a componente *1* da mistura, definida por Diebolt e Robert (1994) da seguinte forma:

$$p_{i1} = \frac{\pi_1 f_1(x_i|\lambda_1)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)}. \tag{2.10}$$

Assim, assume-se que a distribuição condicional para Z_{i1} é Bernoulli com probabilidade de sucesso p_{i1}, com função de probabilidade

$$p(z_{i1}) \propto p_{i1}^{z_{i1}}(1 - p_{i1})^{z_{i2}}, \tag{2.11}$$

onde $Z_{i1} + Z_{i2} = 1$ e Z_{i1} faz com que a variável Z em cada observação i seja atribuída a somente uma das componentes da mistura. Assim, a partir da Equação (2.11), tem-se a distribuição conjunta para $(z_{11}, ..., z_{n1})$ dada por:

$$p(Z_{11}, \ldots, Z_{n1}) \propto \prod_{i=1}^{n} p_{i1}^{Z_{i1}}(1 - p_{i1})^{Z_{i2}},$$

$$\propto \prod_{i=1}^{n} \left(\frac{\pi_1 f_1(x_i|\lambda_1)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)} \right)^{Z_{i1}} \times \left(1 - \frac{\pi_1 f_1(x_i|\lambda_1)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)} \right)^{Z_{i2}},$$

$$\propto \prod_{i=1}^{n} \left(\frac{\pi_1 f_1(x_i|\lambda_1)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)} \right)^{Z_{i1}} \times \left(\frac{\pi_2 f_2(x_i|\lambda_2)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)} \right)^{Z_{i2}},$$

$$\propto \prod_{i=1}^{n} \frac{\left[\pi_1 f_1(x_i|\lambda_1]^{Z_{i1}} \times [\pi_2 f_2(x_i|\lambda_2)\right]^{Z_{i2}}}{\left(\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j) \right)^{Z_{i1}+Z_{i2}}},$$

$$p(Z_{i1}, \ldots, Z_{n1}) \propto \frac{\prod_{i=1}^{n} \prod_{j=1}^{2} [\pi_j f_j(x_i|\lambda_j)]^{Z_{ij}}}{\prod_{i=1}^{n} \left[\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j) \right]}. \tag{2.12}$$

Combinando as Equações (2.8) e (2.12), tem-se a distribuição a *posteriori* para (θ, Z), dada por

$$p(\theta, Z|x) \propto \frac{\prod_{i=1}^{n} \prod_{j=1}^{2} [\pi_j f_j(x_i|\lambda_j)]^{Z_{ij}}}{\prod_{i=1}^{n} \left[\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j) \right]} \times \prod_{i=1}^{n} \left[\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j) \right] \times p(\theta),$$

$$p(\theta, Z|x) \propto \prod_{i=1}^{n} \prod_{j=1}^{2} [\pi_j f_j(x_i|\lambda_j)]^{Z_{ij}} \times p(\theta). \tag{2.13}$$

A formulação bayesiana requer uma distribuição a *priori* para os parâmetros desconhecidos. Foram usandas as prioris conjugadas (como sugerido em Dellaportas et al., 1997), dadas por

$$\lambda_j \sim Gama(\alpha, \beta) \quad j = 1, 2$$

$$\pi \sim Beta(a, b)$$

A partir da distribuição *posteriori* conjunta para (θ, Z) na Equação (2.13), foram encontradas as distribuições condicionais completas para os parâmetros, que são utilizadas para gerar as amostras de cada distribuição de interesse, a partir do algoritmo de amostragem de Gibbs.

O algoritmo de amostragem de Gibbs é um método de simulação via cadeias de Markov que apresenta uma forma de obter uma amostra da distribuição *posteriori* conjunta de interesse, baseada em sucessivas seleções das distribuições condicionais. Esse algoritmo é utilizado no caso em que as distribuições condicionais completas marginais são conhecidas. Em modelos mais complexos, geralmente não temos formas conhecidas nestas distribuições.

As condicionais completas, ou seja, as distribuições posteriores de cada parâmetro condicionado nos demais, são facilmente encontradas e são dadas para $i = 1, ..., n$ e $j = 1, 2$, por

i) $Z_{ij}|\lambda_1, \lambda_2, \pi, x \sim Bernoulli(p_{ij})$, com $p_{ij} = \dfrac{\pi_j f_j(x_i|\lambda_j)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)}$;

ii) $\pi|\lambda_1, \lambda_2, z, x \sim Beta\left(a + \sum_{i=1}^{n} z_{i1}, b + \sum_{i=1}^{n} z_{i2}\right)$;

iii) $\lambda_j|\pi, z, x \sim Gama\left(\alpha + \sum_{i=1}^{n} z_{ij}x_i, \beta + \sum_{i=1}^{n} z_{ij}\right) I(\lambda_{j-1}, \lambda_{j+1})$.

Aqui, I(a, b) representa a função indicadora, que assume o valor 1 se λ_j pertence ao intervalo (a, b) e zero, caso contrário, admitindo-se que $\lambda_0 = 0$ e $\lambda_3 = +\infty$. Segundo Dellaportas *et al.*(1997), a forma truncada da distribuição gama para λ_j é para evitar a multimodalidade da distribuição *posteriori* resultante. Os hiperparâmetros a, b, α e β, caso não se tenham nenhuma informação prévia sobre eles, podem ser escolhidos de modo que as densidade sejam não informativas. No entanto, as *prioris* devem ser adequadas e garantir a identificabilidade parâmetros.

2.5 Abordagem Bayesiana para Modelos de Misturas Finita de Normais

No caso de uma mistura de distribuições Normais com k componentes, a função densidade de probabilidade do modelo, assumindo variâncias iguais para as componentes e médias distintas, é dada por

$$f(x|\mu, \sigma^2, \mathbf{p}) = \sum_{j=1}^{k} p_j (2\pi\sigma^2)^{-1/2} exp\left[-\frac{(x-\mu_j)^2}{2\sigma_2}\right] \tag{2.14}$$

com vetor de parâmetros $\theta = (\mu, \sigma^2, \mathbf{p})$, $\mu = (\mu_1, ..., \mu_k)$ e $\mathbf{p}$=$(p_1, ..., p_k)$. Assim, a distribuição condicional de X, dado μ_j e σ^2 é $N(\mu_j, \sigma^2)$.

Considerando, novamente, o caso $k = 2$, prioris conjugadas para os componentes do vetor de parâmetros foram adotadas, de forma que para as médias foram adotadas prioris normais, para a variância comum adotou-se uma gama-inversa e para a proporção na mistura adotou-se uma distribuição Beta, ou seja, $\mu_j \sim N(\mu_0, \sigma^2/c_0)$, j=1,2, $\sigma^2 \sim GI(n_0/2, n_0\sigma_0^2/2)$ e $p \sim Beta(a, b)$.

Com a introdução de um vetor de variáveis latentes classificadoras $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$, com $\mathbf{Z}_i = (Z_{i1}, Z_{i2})$ como descrito na seção anterior, e assumindo novamente distribuição de Bernoulli com parâmetro dado na Equação (2.10), tem-se que a f.d.p. posterior para o vetor $\theta = (\mu_1, \mu_2, \sigma^2, p)$ é dada por

$$p(\theta|\mathbf{x}, \mathbf{z}) \propto f(\mathbf{x}, \mathbf{z}|\theta) \times p(\theta)$$

$$\propto \prod_{i=1}^{n}\prod_{i=1}^{2}\left[(2\pi\sigma^2)^{-1/2} exp\left[-\frac{(x_i-\mu_j)^2}{2\sigma_2}\right]\right]^{Z_{ij}} \times \left[\left(\frac{\sigma^2}{c_0}\right)^{-1/2} exp\left(-\frac{(\mu_j-\mu_0)^2}{2(\sigma^2/c_0)}\right)\right]^2$$

$$\times(\sigma^2)^{-\frac{n_0}{2}+1} exp\left[-\frac{n_0\sigma_0^2}{2\sigma^2}\right] \times p^{a-1}(1-p)^{b-1}. \tag{2.15}$$

Logo, as condicionais completas marginais dos parâmetros e da variável latente Z, para $i = 1, 2, ..., n$, $p_1 = p$, $p_2 = 1 - p$ e j=1,2, são dadas por

iv) $Z_{ij}|\mu_1, \mu_2, \sigma^2, p, \mathbf{x} \sim Bernoulli(\phi_{ij})$, com $\phi_{ij} = \dfrac{\pi_j f_j(x_i|\lambda_j)}{\sum_{j=1}^{2} \pi_j f_j(x_i|\lambda_j)}$;

v) $p|\mu_1, \mu_2, \sigma^2, \mathbf{z}, \mathbf{x} \sim Beta\left(a + \sum_{i=1}^{n} z_{i1}, b + \sum_{i=1}^{n} z_{i2}\right)$;

vi) $\mu_j|\sigma^2, p, \mathbf{z}, \mathbf{x} \sim N\left(\dfrac{c_0\mu_0 + \sum_{i=1}^{n} z_{ij}x_i}{c_0 + \sum_{i=1}^{n} z_{ij}}, \dfrac{\frac{\sigma^2}{n}}{c_0 + \sum_{i=1}^{n} z_{ij}}\right) I(\mu_{j-1}, \mu_{j+1})$.

vii) $\sigma^2|\mu_1\mu_2, p, \mathbf{z}, \mathbf{x} \sim GI\left(\dfrac{2n+2+n_0}{2}, \dfrac{n_0\sigma_0^2 + \sum_{i=1}^{n}\sum_{i=1}^{2} z_{ij}(x_i-\mu_j)^2}{2}\right)$.

Os hiperparâmetros a, b, c_0, μ_0, n_0 e σ_0^2, caso não se tenha informação prévia sobre eles, podem ser escolhidos de forma que as prioris sejam não informativas, mas garantindo a identificabilidade dos parâmetros.

2.6 Aspectos Computacionais

A obtenção das distribuições condicionais completas *a posteriori* ocorre de forma analitica para o caso de mistura finita de modelos Poisson com o número de componentes conhecido. Para se obter as amostras destas distribuições é necessário a implementação do algoritmo de Gibbs. Vários trabalhos podem ser encontrados com o passo a passo do algoritmo. Saito e Rodrigues (2005) implementaram no software Winbugs as distribuições *condicionais* e obtiveram as estimativas para os parâmetros desconhecidos. O mesmo software também foi utilizado por Pissini (2006).

O Winbugs é um programa desenvolvido para a implementação da metodologia bayesiana, a partir das especificações do modelo, as distribuições a *priori*, os valores amostrais e os valores iniciais, tudo isso para o processo numérico de estimação.

Em todos os casos também se faz necessário estudar o comportamento da cadeia gerada no processo MCMC. Assim, neste trabalho será considerado dois termos fundamentais o "*burn in*"e a correlação entre as observações geradas. O "*burn in*"é o termo que descreve a prática de aquecimento da cadeia de Markov, ignorando algumas iterações no começo de um processo MCMC. Ou seja, deve-se começar a cadeia em algum lugar, por exemplo, em x, então roda-se a cadeia de Markov m vezes (até atingir sua estacionariedade), e descarta-se esta amostra inicial. Este período é chamado de *burn-in*. Já o controle da correlação é um modo de eliminar uma possível influência entre as iterações realizadas. Assim, após geradas as m amostras, considera-se "*saltos*"na escolha da amostra, onde este salto depende do problema estudado (PAULINO *et al.*, 2003).

Neste trabalho, a estimação dos parâmetros desconhecidos foi implementada computacionalmente com o uso de uma rotina desenvolvida no MATLAB. O **MATLAB** (MATrix LABoratory) é um software interativo de alta performance voltado para o cálculo numérico, permitindo assim que o cálculo de muitos problemas numéricos sejam resolvidos em fração de segundos. Outra van-

tagem do MATLAB é que, além disso, as soluções dos problemas são expressas quase exatamente como elas são escritas matematicamente.

Capítulo 3

Full Bayesian Siginificance Test

3.1 Introdução

O problema estatístico de teste de hipóteses consiste em rejeitar ou não uma hipótese sobre o valor de um parâmetro desconhecido com base na informação trazida pela amostra. Assim, na abordagem Clássica é bastante usada uma medida de evidência denominada *nível descritivo* ou *p-valor*, p. Esse baseia-se na distribuição amostral da estatística do teste e tem por finalidade medir a evidência trazida pelos dados em favor da hipótese nula H_0. Calcula-se uma estatística de teste, obtida a partir da observação dos dados amostrais, denominada $T = T(\mathbf{X})$, e com base no seu valor observado mede-se a evidência contra a hipótese H_0. O nível descritivo observado (p) é medido pela probabilidade de se obter um valor mais extremo do que o valor observado da estatística, ou seja, obtém-se a probabilidade, sob H_0, para os pontos do espaço amostral que são tão ou mais desfavoráveis para a hipótese nula do que o valor observado, e valores pequenos para p indicam que os dados observados não favorecem a hipótese nula. levando à decisão de rejeitar H_0.

Observa-se que todo esse processo inferencial é realizado sem levar em consideração a hipótese alternativa, pois a distribuição de probabilidades da estatística T só precisa ser conhecida sob a hipótese nula. Outra questão levantada é que o cálculo do nível p é feito levando-se em consideração a informação dos dados que poderiam ter sido observados, mas que ainda não foram, violando o Princípio da Verossomilhança, segundo qual todo processo de decisão deve ser feito nos dados efetivamente observados.

Com a abordagem Bayesiana, o cálculo da medida de evidência baseia-se na função de verossimilhança dos dados e na distribuição a *priori* para a quantidade desconhecida. A vantagem dos testes bayesianos é que eles levam em consideração a hipótese alternativa e, dentre eles, destacam-se o Fator de Bayes e a Probabilidade *Posteriori* de H_0, $Pr\{H_0|x\}$. O Fator de Bayes envolve no

seu cálculo a razão entre as probabilidades posteriores de cada hipótese e suas probabilidades a priori, e é descrito por

$$FB_{01}(x) = \left\{\frac{Pr\{H_0|x\}}{Pr\{H_1|x\}}\right\} / \left\{\frac{Pr\{H_0\}}{Pr\{H_1\}}\right\}.$$

Assim, com o Fator de Bayes tem-se uma medida que indica se os dados aumentaram ou diminuíram as chances de H_0 relativamente a H_1, onde um valor grande para $FB_{01}(x)$ significa uma maior evidência nos dados favorecendo a hipótese nula.

Alguns autores, Berger e Selke (1987), e Berger e Delampady (1987), entre outros, apresentaram e discutiram os conflitos entre a medida de evidência Clássica e as medidas de evidência Bayesiana, alertando para o fato de que em algumas situações o p-valor pode não ser uma boa medida de evidência para uma hipótese estatística precisa (hipótese com dimensão menor que a dimensão do espaço paramétrico).). Na abordagem Bayesiana, o tratamento de uma hipótese precisa leva à necessidade de introduzir uma massa de probabilidade positiva no valor definido sob H_0, quando o espaço paramétrico é contínuo, causando o desconforto de se trabalhar com um modelo de probabilidade misto. Com isso, houve o surgimento de novas medidas de evidência para teste de hipóteses precisas. Neste sentido será apresentada a medida de evidência genuinamente Bayesiana.

3.2 Definição

Proposto por Pereira e Stern (1999), o procedimento é denominado *Full Bayesian Signicance Test* (FBST), cujo objetivo é testar hipóteses precisas baseadas no cálculo da probabilidade *posteriori* da região HPD (*Highest Posterior Density*), que é tangente ao conjunto que define a hipótese nula. Segundo Pereira e Stern (1999) o teste é intuitivo, com fácil caracterização geométrica e pode ser implementado a partir de técnicas de otimização e integração numérica. Eles definiram a seguinte medida de evidência em favor de uma hipótese precisa:

Definição 1. *Considere um modelo estatístico paramétrico, isto é, uma quíntupla* $(\chi, A, F, \Theta, \pi)$*, onde* χ *é o espaço amostral,* A *é uma sigma - álgebra conveniente de subconjuntos de* χ*,* F *é uma classe de distribuições de probabilidade em* A *indexadas no espaço paramétrico* Θ *e* π *é uma densidade a priori em* Θ*. Suponha que um subconjunto* Θ_0 *de* Θ *tendo medida de Lebesgue nula é de interesse. Considere também que* $p(\theta|\boldsymbol{x})$ *é a densidade posterior de* θ*, dada a observação amostral* $\boldsymbol{x}$*, e seja o conjunto* $T(\boldsymbol{x}) = \{\theta \in \Theta : p(\theta|\boldsymbol{x}) > sup_{\Theta_0} p(\theta|x)\}$*. A medida de evidência de*

Pereira-Stern em favor de Θ_0 *é definida como* $EV(\Theta_0, \boldsymbol{x}) = 1 - Pr[\theta \in T(\boldsymbol{x})|\boldsymbol{x}]$ *e um teste (ou procedimento), denominado Full Bayesian Significance Test, é aceitar* Θ_0 *sempre que é "grande".*

Percebe-se então que a medida de evidência de Pereira - Stern leva em consideração todos os pontos do espaço paramétrico que são menos prováveis do que algum ponto em Θ_0. Se a probabilidade posterior do conjunto $T(\mathbf{x})$ é "grande ", isso significa que o conjunto de valores da hipótese nula está em uma região de baixa probabilidade, e a evidência trazida pelos dados é contra a hipótese nula. Por outro lado, se a probabilidade de $T(\mathbf{x})$ é "pequena", então o conjunto de valores da hipótese nula está em uma região de alta probabilidade posterior, o que leva a uma evidência em favor da hipótese nula.

Uma das vantagens desta metodologia é que o problema de se trabalhar com uma hipótese precisa é contornado, uma vez que não há necessidade de se introduzir uma probabilidade positiva a priori como no teste de Jeffreys. Madruga (2002) prova que:

i) O FBST não viola o Princípio da Verossimilhança;

ii) A medida de evidência, sob as condições do item anterior, é consistente;

iii) A medida de evidência corrigida é invariante sob transformações um-a-um do parâmetro de interesse.

Madruga *et al.* (2001) apresentam funções de perda cuja minimização da sua esperança posterior, sob a hipótese H0, levam ao procedimento proposto por Pereira e Stern (1999). Isto torna o FBST um teste de hipóteses genuinamente "Bayesiano", com caracterização dentro da abordagem de Teoria da Decisão.

3.3 Regra de Decisão e Validação no FBST

Após o cálculo da medida de evidência $EV(\Theta_0, x)$, é necessário construir uma regra de decisão, ou seja determinar um valor e_c de modo que:

i) Rejeita-se H_0 se $EV(\Theta_0, x) \leq e_c$

ii) Aceita-se H_0 se $EV(\Theta_0, x) > e_c$.

O valor c depende da função de perda e pode assumir valores diferentes, isso porque existem variações da função de perda com interpretações diferentes. Madruga *et al.* (2001) define o

problema da seguinte forma: Considere $D = \{d_0, d_1\}$ o espaço de decisões usual em um problema estatístico de teste de hipótese, com d_0 representado a decisão de aceitar H_0 e d_1 a de rejeitar H_0, e seja a função de perda definida por $L : D \times \Theta_0 \longrightarrow \mathbb{R}^+, L(RejeitarH_0, \theta) = \alpha[1 - \mathbf{1}(\theta \in T(x))]$ e $L(AceitarH_0, \theta) = b + c1(\theta \in T(\mathbf{x}))$, com α, b e $c > 0$. Para esta função de perda, Madruga et al. (2001) mostram que o valor de corte é $e_c = (b + c)/(\alpha + c)$.

Na prática, a escolha dos valores de α, b e c, necessários para a tomada de decisão não é simples e envolve a opinião do pesquisador sobre o erro mais (ou menos) danoso na sua decisão. Nos casos em que a evidência obtida é muito próxima de zero (ou de 1), a decisão natural é rejeitar (ou aceitar) a hipótese H0. Nas demais situações, pode-se estabelecer o nível de significância do teste, como é feito nos testes clássicos, e buscar a validação do resultado obtido.

Uma forma de validar o resultado do FBST em estudos de simulação é através de medidas empíricas, tais como o *nível de significância* e o *poder empírico* do teste. Madruga *et al.* (2005) apresentaram as seguintes medidas empírica para comparação entre resultados clássicos e o FBST:

***i*)** O nível de significância empírico de um teste de hipótese, com base em um grande número de repetições, é dado pela proporção de vezes em que a hipótese nula é rejeitada quando ela é verdadeira;

***ii*)** O poder empírico de um teste de hipótese, com base em um grande número de repetições, é dado pela proporção de vezes em que a hipótese nula é rejeitada quando ela é falsa.

Neste trabalho, o nível de significância empírico do FBST será fixado em 5%, obtendo-se um valor de corte e_c, que será utilizado na regra de decisão. Com isso, o poder empírico poderá ser obtido para validar o uso do FBST no teste de homogeneidade em modelos com mistura finita de distribuições e número de componentes conhecida.

Capítulo 4

Aplicações e Resultados

4.1 Teste de homogeneidade na mistura de duas distribuições Poisson

Seja X uma variável aleatória composta da mistura de duas distribuições Poisson com parâmetros λ_1 e λ_2 $(0 < \lambda_1 < \lambda_2)$, e parâmetro de proporção da mistura π, com função de probabilidade para x=0,1,2,..., dada por

$$f(x|\lambda_1, \lambda_2, \pi) = \pi \frac{e^{-\lambda_1}\lambda_1^x}{x!} + (1-\pi)\frac{e^{-\lambda_2}\lambda_2^x}{x!}. \tag{4.1}$$

O interesse aqui é testar $H_0 : \lambda_1 = \lambda_2$ (ou, equivalentemente, $\pi = 0$ ou $\pi = 1$) versus $H_0 : \lambda_1 \neq \lambda_2$ (ou $\pi \neq 0$). Seja $\theta = (\lambda_1, \lambda_2, \pi)$ o parâmetro de interesse e o espaço paramétrico difinido por $\Theta = \theta : \lambda_1, \lambda_2 > 0 \; e \; 0 < \pi < 1$. A hipótese nula para o caso de mistura determina o seguinte subconjunto no espaço paramétrico: $\Theta_0 = \{(\lambda_1, \lambda_2, \pi) \in \Theta : \lambda_1 = \lambda_2, \pi = 0 \; ou \; 1\}$.

Sob a hipótese H_1 , a variável aleatória X tem distribuição de Poisson conforme a função de densidade descrita em na Equação (4.1). Sob a abordagem bayesiana, é necessário atribuir uma distribuição *a priori* $p(\theta)$ para o vetor de parâmetros desconhecidos $\theta = (\lambda_1, \lambda_2, \pi)$. Como já discutido na seção 4.2, prioris conjugadas foram escolhidas de modo a garantir que as distribuições posteriores pertençam a mesma família paramétrica. Com isso adotou-se a distribuição $Gamma(\alpha, \beta)$ para cada $\lambda_i, i = 1, 2$, e uma distribuição para π. Logo, a função de probabilidade posterior p(x,z) é dada por

$$p(\theta|\mathbf{x}, \mathbf{z}) \propto f(\mathbf{x}, \mathbf{z}|\theta) \times p(\theta)$$

$$\propto \prod_{i=1}^{n}\prod_{i=1}^{2}[f_i(x_i|\lambda_j)]^{Z_{ij}} \times p(\theta)$$

$$\propto \prod_{i=1}^{n}\prod_{i=1}^{2}[f_i(x_i|\lambda_j)]^{Z_{ij}} \times \pi^{a-1}(1-\pi)^{b-1} \times \lambda_1^{\alpha-1}e^{-\frac{\lambda_1}{\beta}} \times \lambda_2^{\alpha-1}e^{-\frac{\lambda_2}{\beta}}$$

$$\propto \prod_{i=1}^{n}\prod_{i=1}^{2}(e^{-\lambda_1}\lambda_j^{x_i})^{Z_{ij}} \times \pi^{a-1}(1-\pi)^{b-1} \times \lambda_1^{\alpha-1}e^{-\frac{\lambda_1}{\beta}} \times \lambda_2^{\alpha-1}e^{-\frac{\lambda_2}{\beta}}$$

$$\propto \prod_{i=1}^{n}\left[(e^{-\lambda_1}\lambda_1^{x_i})^{Z_{i1}} \times (e^{-\lambda_2}\lambda_2^{x_i})^{Z_{i2}}\right] \times \pi^{a-1}(1-\pi)^{b-1} \times \lambda_1^{\alpha-1}e^{-\frac{\lambda_1}{\beta}} \times \lambda_2^{\alpha-1}e^{-\frac{\lambda_2}{\beta}}$$

Portanto,

$$p(\theta|\mathbf{x},\mathbf{z}) \propto e^{-\lambda_1(\sum_{i=1}^{n} Z_{i1}+\beta^{-1})}\lambda_1^{\alpha+\sum_{i=1}^{n} x_i Z_{i1}} \times e^{-\lambda_2(\sum_{i=1}^{n} Z_{i2}+\beta^{-1})}\lambda_2^{\alpha+\sum_{i=1}^{n} x_i Z_{i2}} \times \pi^{a-1}(1-\pi)^{b-1}.$$

Foram obtidas as condicionais completas para cada parâmetro desconhecido, assim como para a variável latente Z, conforme descrito no capítulo 2:

(a) $Z_{ij}|\lambda_1,\lambda_2,\pi,x \sim Bernoulli(p_{ij})$ i=1, ..., n, j=1,2 com

$$p_{ij} = \frac{\pi f_1(x_i|\lambda_1)}{\pi f_1(x_i|\lambda_1) + (1-\pi) f_2(x_i|\lambda_2)};$$

(b) $\pi|\lambda_1,\lambda_2,z,x \sim Beta\left(a+\sum_{i=1}^{n} z_{i1}, b+\sum_{i=1}^{n} z_{i2}\right)$;

(c) $\lambda_j|\pi,z,x \sim Gama\left(\alpha+\sum_{i=1}^{n} z_{ij}x_i, \beta^{-1}+\sum_{i=1}^{n} z_{ij}\right) I(\lambda_{j-1},\lambda_{j+1})$, j=1 e 2.

Conhecidas as distribuições condicionais completas apresentadas em (a) - (c), foram geradas amostras aleatórias da distribuição posterior $p(\theta|\mathbf{x},\mathbf{z})$, $((\lambda_1,\lambda_2,\pi)_1,(\lambda_1,\lambda_2,\pi)_2,...,(\lambda_1,\lambda_2,\pi)_m)$, que foram usadas para obter a medida de evidência $EV(\Theta_0,\mathbf{x})$. Essas amostras foram obtidas a partir do método de amostragem de Gibbs, com a implementação da rotina no MATLAB.

Sob a hipótese nula, a variável aleatória X tem distribuição Poisson (λ), com apenas um parâmetro desconhecido, λ. Com base em uma distribuição *a priori* $Gamma(\alpha,\beta)$ para λ, tem-se que a função de probabilidade posterior sob a hipótese nula é dada por

$$p_{H_0}(\lambda|\mathbf{x}) \propto f(\mathbf{x}|\lambda) \times p(\lambda)$$

$$\propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \times \lambda^{\alpha-1} e^{-\lambda/\beta} = \lambda^{\alpha+\sum_{i=1}^{n} x_i - 1} e^{-\lambda(n+\beta^{-1})}$$

Portanto, sob H_0 conclui-se que a distribuição de λ é Gama(A, B), com $A = \alpha + \sum_{i=1}^{n} x_i$ e $B = (n + \beta^{-1})^{-1}$.

De acordo com o procedimento FBST é necessário agora obter o valor supremo de $p_{H_0}(\lambda|\mathbf{x})$. Assim, calculou-se a moda da distribuição Gama(A, B), ou seja, um estimador Bayesiano para λ sob H_0, a partir da derivada do logaritmo neperiano de $p_{H_0}(\lambda|\mathbf{x})$,

$$p_{H_0}(\lambda|\mathbf{x}) = \frac{1}{B^A \Gamma(A)} \lambda^{A-1} e^{-\lambda/B}$$

$$lnp_{H_0}(\lambda|\mathbf{x}) = -ln(B^A \Gamma(A)) + (A-1)ln(\lambda) - \lambda/B$$

$$\frac{d}{d\lambda} ln(p_{H_0}(\lambda|\mathbf{x})) = \frac{A-1}{\lambda} - \frac{1}{B}$$

$$\frac{d}{d\lambda} ln(p_{H_0}(\lambda|\mathbf{x})) = 0 \Leftrightarrow \hat{\lambda}_{H_0} = (A-1)B = \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{n + \beta^{-1}}.$$

Assim, para o cálculo da $EV(\Theta_0, \mathbf{x})$, é necessário obter a região T do espaço paramétrico que contém os pontos θ com função de probabilidade posterior maior que $p_{H_0}(\hat{\lambda}_{H_0}|x)$.

A Seção 4.1.1. apresenta os resultados obtidos para dados simulados de uma mistura de duas distribuições Poisson, apresentando as estimativas dos parâmteros e o poder empírico associado ao FBST. A Seção 4.1.2. apresenta as estimativas dos parâmetros e a medida de evidência para um conjunto de dados reais analisado em Dellaportas et al. (1997).

4.1.1 Dados Simulados

Os dados utilizados foram simulados a partir de uma mistura de duas distribuições Poisson com parâmetros $\lambda_1 = 10$, $\lambda_2 = 18$ e $\pi = 0,3$. Os valores dos hiperparâmetros das distribuições

a priori foram escolhidos de modo que as mesmas fossem pouco informativas, de modo que a evidência maior fosse obtida a partir da amostra observada. As prioris adotadas foram $\lambda_j \sim Gama(0,01; 1000)$, j=1,2, e $\pi \sim Beta(1,1)$.

Com isso, calculou-se a evidência em favor de H_0 para diferentes tamanhos amostrais n e diferentes números de amostras (iterações) m, geradas via amostrador de Gibbs a partir das distribuições condicionais completas apresentadas em (a) - (c). No processo de amostragem adotou-se *burn in* de tamanho 100, ou seja, as 100 primeiras observações foram eliminadas para valores de m iguais a 1100 e 5100, e *burn in* de tamanho 1000 para m igual a 11000 e 51000. Considerou-se, também, saltos de tamanho 10 entre as amostras geradas para eliminar a correlação.

A Tabela 4.1 mostra o valor da evidência $EV(\Theta_0, \mathbf{x})$ em favor da hipótese H_0 de homogeneidade, segundo os valores de **n** e **m** adotados. Observa-se que a evidência em favor de H_0 é pequena, como esperado, uma vez que os dados foram gerados de uma mistura com duas componentes. Porém, seu valor decresce ("melhora") com o aumento do tamanho amostral **n**. Isto justifica-se pelo fato de que amostras pequenas dificultam a alocação de observações na componente de menor proporção na mistura, e prejudica o processo de estimação (DIEBOLT e ROBERT, 1994).

Tabela 4.1 *Valores da evidência e dos parâmetros estimados segundo o tamanho amostral* (n) *e o número de amostras geradas* (m).

				Número de Amostras				
Tamanho		**1100**				**5100**		
amostral	π_1	λ_1	λ_2	Evidência	π_1	λ_1	λ_2	Evidência
30	0,15	3,15	19,17	0,26	0,18	3,62	34,44	0,29
50	0,18	6,95	16,99	0,001	0,19	6,94	16,98	0,0002
100	0,27	8,95	17,36	$< 0,0001$	0,27	8,97	17,4	$< 0,0001$
200	0,29	9,73	18,49	$< 0,0001$	0,29	9,65	18,52	$< 0,0001$
Tamanho		**11000**				**51000**		
amostral	π_1	λ_1	λ_2	Evidência	π_1	λ_1	λ_2	Evidência
30	0,14	2,80	30,64	0,24	0,11	2,27	26,39	0,21
50	0,19	7,02	17,03	0,0001	0,21	7,21	17,07	0,0008
100	0,27	9,00	17,45	$< 0,0001$	0,27	9,01	17,45	$< 0,0001$
200	0,29	9,68	18,52	$< 0,0001$	0,3	9,66	18,51	$< 0,0001$

A partir destes resultados, calculou-se o poder empírico como definido na seção 3.3 do Capítulo 3. O valor de corte e_c para a regra de decisão foi obtido a partir do nível de significância empírico fixado em 5%. Assim, 1.000 amostras foram geradas, sob a hipótese de homogeneidade, de uma mistura de duas distribuições Poisson tomando-se $\lambda_1 = \lambda_2 = 10$. Em seguida, os 1.000 valores da $EV(\Theta_0, \mathbf{x})$, um para cada amostra gerada, foram obtidos e ordenados. O percentil de ordem 5 foi escolhido como ponto de corte e_c . Assim, a regra de decisão obtida foi:

i. Rejeita-se H_0 se $EV(\Theta_0, \mathbf{x}) \leq 0,158$;

ii. Aceita-se H_0 se $EV(\Theta_0, \mathbf{x}) > 0,158$.

Para a obtenção do poder empírico foram fixados diferentes valores para a diferença $\lambda_2 - \lambda_1$ e, para cada diferença, gerou-se 1.000 amostras e obteve-se a $EV(\Theta_0, \mathbf{x})$ associada. Com base na regra de decisão expressa acima, obteve-se a proporção de vezes em que a hipótese de homogeneidade foi rejeitada, ou seja o poder empírico. Na Tabela 4.2 tem-se o poder empírico, segundo a diferença $\lambda_2 - \lambda_1$. Observa-se que à medida que a diferença $\lambda_2 - \lambda_1$ aumenta, a proporção de vezes que se rejeita H_0, quando de fato ela é falsa, aumenta e tende para 1, ou seja, o FBST consegue evidenciar que a amostra provém de uma mistura de distribuições Poisson. A Figura 4.1 mostra o valor médio das evidências em favor de H_0 para cada diferença, com base nos 1.000 valores obtidos em cada uma delas. As Figuras 4.2 a 4.7 mostram a convergência da medida de evidência para diferentes valores da diferença $\lambda_2 - \lambda_1$.

Tabela 4.2 *Poder Empírico do Teste FBST, Segundo a Diferença* $\lambda_2 - \lambda_1$

λ_1	λ_2	$\lambda_2 - \lambda_1$	**Poder Empírico**
10	12	2	0,066
10	14	4	0,328
10	15	5	0,578
10	16	6	0,814
10	18	8	0,996
10	20	10	1,000

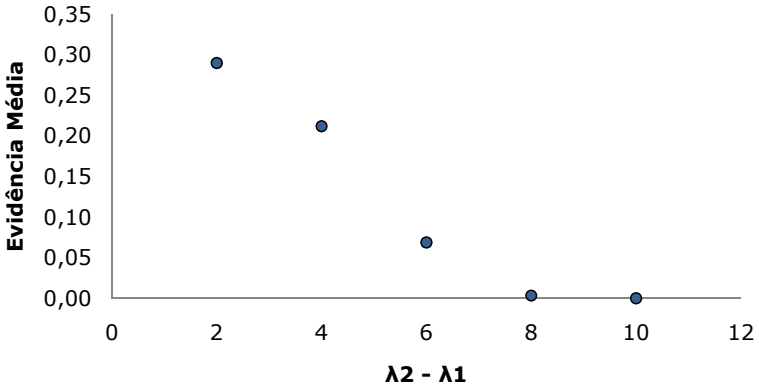

Figura 4.1 *Evidência Média, Segundo a Diferença* $\lambda_2 - \lambda_1$.

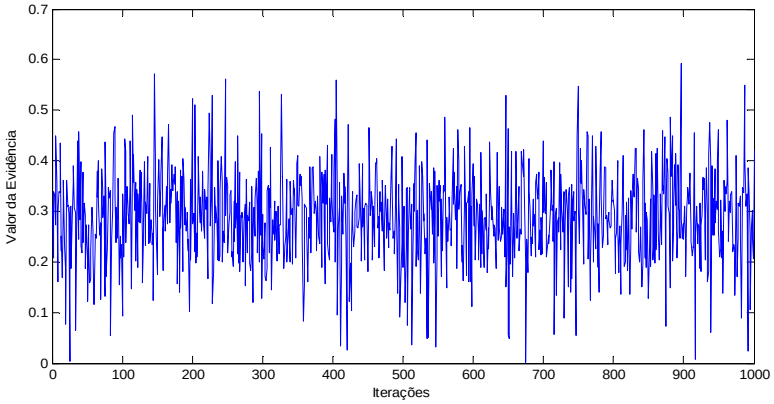

Figura 4.2 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 2$.

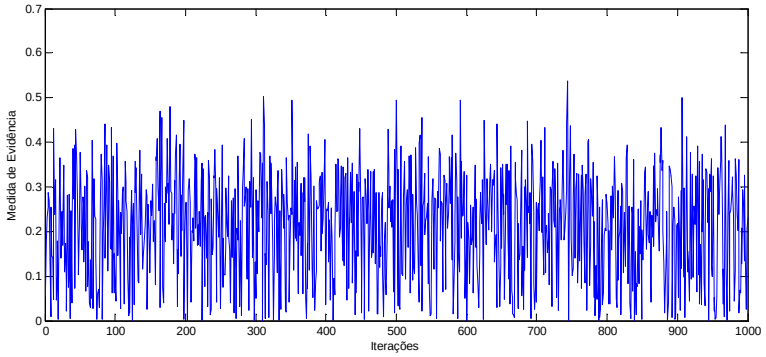

Figura 4.3 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 4$.

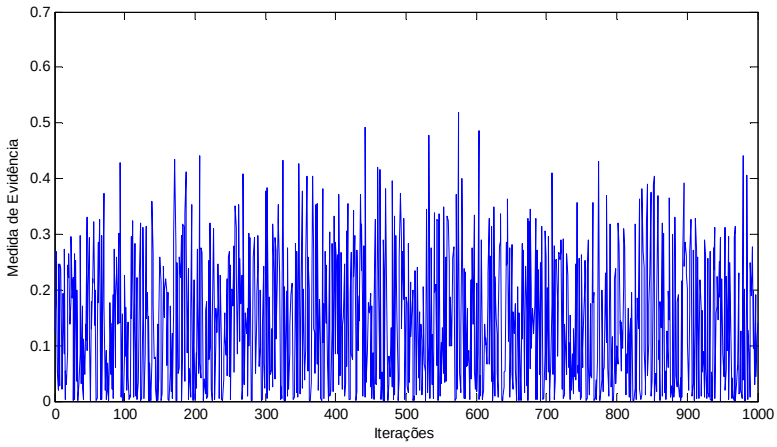

Figura 4.4 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 5$.

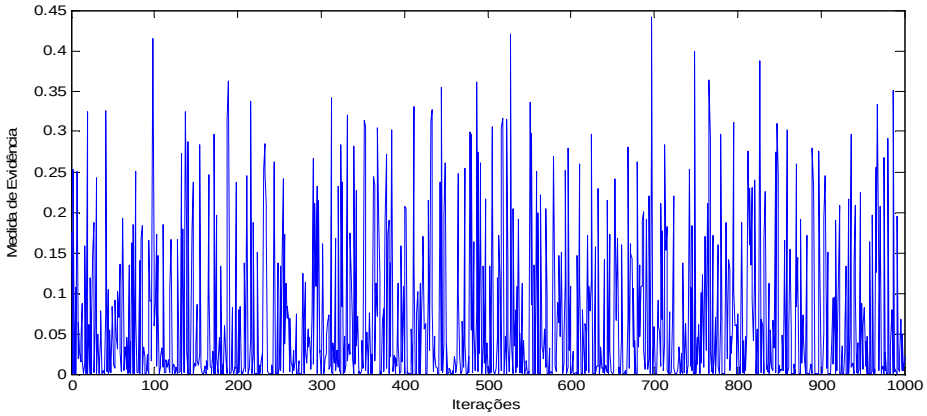

Figura 4.5 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 6$.

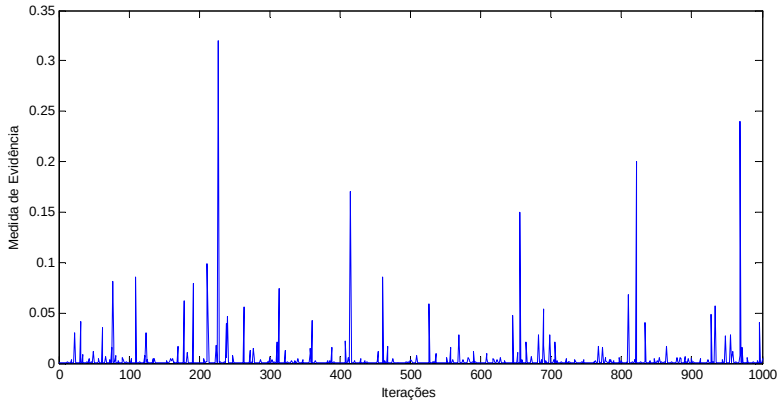

Figura 4.6 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 8$.

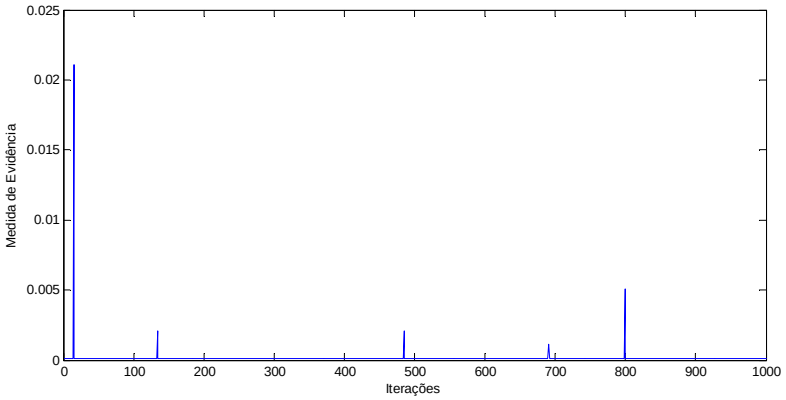

Figura 4.7 *Convergência da Evidência para* $\lambda_2 - \lambda_1 = 10$.

4.1.2 Aplicação a Dados Reais

A fim de analisar o comportamento do FBST em dados reais, foram utilizados os dados analisados em Dellaportas *et al* (1997), referentes ao número de parcelas em atraso de clientes que tomaram empréstimo em uma instituição financeira da Espanha em 1990. Em estudos de risco de crédito, é de interesse das instituições financeiras classificar um potencial cliente como "bom"ou "mau"pagador, para subsidiar a decisão sobre um pedido de empréstimo. Os dados são apresentados na Tabela 4.3.

Tabela 4.3 *Quantidade de Clientes por Número de Parcelas em Atraso em uma Instituição Financeira da Espanha em 1990.*

N° de Parcelas	Quantidade de Clientes	N° de Parcelas	Quantidade de Clientes	N° de Parcelas	Quantidade de Clientes	N° de Parcelas	Quantidade de Clientes
0	3002	7	80	14	13	21	0
1	502	8	59	15	11	22	1
2	187	9	53	16	4	23	0
3	138	10	41	17	5	24	1
4	233	11	28	18	8		
5	160	12	34	19	6	$> 24*$	4
6	107	13	10	20	3		

*28,29,30 e 34.

Os dados contêm uma grande dispersão, com média igual a 1,58 e variância igual a 9,93, o que implica a inadequação de uma única distribuição de Poisson para modelar a distribuição dos dados observados. Dellaportas *et al.* (1997) ajustaram um modelo de mistura para descrever a heterogeneidade dentro da população. Eles modelaram considerando o número de componentes

que melhor se ajustava aos dados, verificando assim que um modelo de mistura de poissons com duas componentes foi o mais adequado.

Assim, o FBST foi usado para testar a hipótese nula de homogeneidade contra a hipótese alternativa de mistura com duas componentes. Os resultados mostram que a evidência em favor de H_0 é abaixo de 0,0001, reforçando a conclusão obtida por Dellaportas *et al.* (1997) de que os dados provém de uma mistura de distribuição Poisson. A Tabela 4.4 apresenta as estimativas dos parâmetros obtidas anteriormente em Dellaportas *et al.* (1997) e aquelas atuais, obtidas neste trabalho, bem como a medida de evidência.

Vale ressaltar que no caso do conjunto de dados considerado, a própria instituição estaria interessada em determinar o número de diferentes subpopulações que compõem a população inteira do cliente. No entanto, era necessário primeiramente comprovar a heterogeneidade da população, e isto pôde ser feito através do FBST, utilizando o número de componentes sugerido em Dellaportas *et al.* (1997) para os dados analisados.

Tabela 4.4 *Estimativas dos Parâmetros Obtidas no Trabalho Anterior e no Atual, e a Medida de Evidência em Favor da Hipótese de Homogeneidade.*

	Estimativas	
Parâmetros	**Dellaportas et al. (1997)**	**Atual**
π_1	0,78	0,77
π_2	0,22	0,23
λ_1	0,24	0,20
λ_2	6,41	6,14
	Medida de Evidência (FBST)	<0,0001

4.2 Teste de Homogeneidade na Mistura de Duas Distribuições Normais

Seja X uma variável aleatória composta da mistura de duas distribuições Normais, com variâncias iguais a σ^2 e médias μ_1 e μ_2, e parâmetro de proporção da mistura p, cuja função de probabilidade para $-\infty < x < +\infty$, é dada por

$$f(x|\mu_1, \mu_2, \sigma^2, p) = p(2\pi\sigma^2)^{-1/2} e^{\frac{(x-\mu_1)^2}{2\sigma^2}} + (1-p)(2\pi\sigma^2)^{-1/2} e^{\frac{(x-\mu_2)^2}{2\sigma^2}}. \quad (4.2)$$

Novamente, o interesse aqui é testar $H_0 : \mu_1 = \mu_2$ (ou equivalentemente, p=0 ou p=1) versus $H_1 : \mu_1 \neq \mu_2$ (ou $p \neq 0$). Seja $\theta = (\mu_1, \mu_2, \sigma^2, p)$ o parâmetro de interesse, cujo espaço paramétrico é definido por $\Theta = \{\theta : -\infty < \mu_1, \mu_2 < +\infty, \sigma > 0, 0 \leq p \leq 1\}$. A hipótese nula para o caso de mistura determina o seguinte subconjunto no espaço paramétrico $\Theta_0 = \{\theta \in \Theta : \mu_1 = \mu_2, p = 0 \; ou \; 1\}$. De (2.14) tem-se que a f.d.p. posterior conjunta dos parâmetros é dada por

$$p(\mu_1, \mu_2, \sigma^2, p|\mathbf{x}) \propto \prod_{i=1}^{n} \sigma^{-(Z_{i1}+Z_{i2})} exp\{-\frac{1}{2\sigma^2}[Z_{i1}(x_i - \mu_1)^2 + Z_{i2}(x_i - \mu_1)^2]\} \times$$

$$\sigma^{-2} exp\left(-\frac{c_0}{2\sigma^2}[(\mu_1 - \mu_0)^2 + (\mu_2 - \mu_0)^2]\right) \times (\sigma^2)^{-\frac{n_0}{2}+1} exp\left[-\frac{n_0\sigma_0^2}{2\sigma^2}\right] \times p^{a-1}(1-p)^{b-1}.$$

e suas condicionais completas foram apresentadas em (IV) -(VII) na Seção 2.5. Assim, com a implementação do algoritmo de Gibbs, amostras são geradas da distribuição posterior acima, a fim de se obter a medida de evidência proposta no FBST para o teste de homogeneidade.

Sob a hipótese H_0, a variável aleatória X tem distribuição normal com parâmetros μ e σ^2. A fim de obter o máximo da distribuição *posterior* sob H0, foram atribuídas uma distribuição *a priori* $Normal(\mu_0, \sigma^2/c_0)$ para μ e uma *Gamma-Inversa* $(n_0/2, n_0\sigma_0^2/2)$ para σ^2. A função densidade de probabilidade posterior sob a hipótese nula é dada por:

$$p_{H_0}(\mu, \sigma^2|x) \propto L^{H_0} \times p(\mu, \sigma^2)$$

$$p_{H_0}(\mu, \sigma^2|x) \propto L^{H_0} \times p(\mu|\sigma^2) \times p(\sigma^2)$$

$$p_{H_0}(\mu, \sigma^2|x) \propto (\sigma^2)^{-\frac{n}{2}} \; exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \times \left(\frac{\sigma^2}{c_0}\right)^{-1/2} exp\left[-\frac{c_0}{2\sigma^2}(\mu - \mu_0)^2\right] \times$$

$$(\sigma^2)^{-(\frac{n_0}{2}+1)} exp\{-n_0\sigma_0^2/2\sigma^2\}.$$

Os estimadores de máxima a posteriori sob H_0, que tornam a f.d.p. posterior, sob H_0, máxima são dados por

$$\hat{\mu} = \frac{n\bar{x} + c_0\mu_0}{c_0 + n} \quad e \quad \hat{\sigma}^2 = \frac{(n-1)S^2 + n_0\sigma_0^2 + \frac{n_0c_0\bar{x}^2 + nc_0\mu_0^2 - n\bar{x}c_0\mu_0}{n+c_0}}{n + n_0 + 3}.$$

Assim, para o cálculo da $EV(\Theta_0, \mathbf{x})$, é necessário obter a região T do espaço paramétrico que contém os pontos θ com função de probabilidade posterior maior que $p_{H_0}(\hat{\mu}, \hat{\sigma}_0^2|\mathbf{x})$.

A Seção 4.2.1. apresenta os resultados obtidos para dados simulados de uma mistura de duas distribuições Normais, apresentando as estimativas dos parâmteros e o poder empírico associado ao FBST. A Seção 4.2.2. apresenta as estimativas dos parâmetros e a medida de evidência para um conjunto de dados reais analisado em Marin et. al. (2005).

$$ln\ f(\theta|x,z) \propto -0,5ln\ (\sigma^2)\sum_{i=1}^{n} z_{i1} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} z_{i1}(x_i-\mu_1)^2 - 0,5ln\ (\sigma^2)\sum_{i=1}^{n} z_{i2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} z_{i2}(x_i-\mu_2)^2$$
$$-0,5\ ln\ (\sigma^2) - \frac{c_0}{2\sigma^2}(\mu_1-\mu_0)^2 - 0,5\ ln\ (\sigma^2) - \frac{c_0}{2\sigma^2}(\mu_2-\mu_0)^2 \times (-0,5\ n_0+1)\ ln\ (\sigma^2) - \frac{n_0\sigma_0^2}{2\sigma^2}.$$

4.2.1 Dados Simulados

Os dados utilizados foram simulados a partir de uma mistura de duas distribuições Normais com parâmetros $\mu_1 = 30$, $\mu_2 = 35$, $\sigma^2 = 9$ e p=0,4. Os valores dos hiperparâmetros das distribuições *a priori* foram escolhidos de modo que as mesmas fossem pouco informativas, de modo que a evidência maior fosse obtida a partir da amostra observada. As prioris adotadas foram $\mu_j \sim N(0, 10\sigma^2)$, j=1,2, $\sigma^2 \sim IG(3, 20)$ e $p \sim Beta(1, 1)$.

Com isso, calculou-se a evidência em favor de H_0 para diferentes tamanhos amostrais n e diferentes números de amostras (iterações) m, geradas via amostrador de Gibbs a partir das distribuições condicionais completas apresentadas em (IV) - (VII). No processo de amostragem adotou-se *burn in* de tamanho 100, ou seja, as 100 primeiras observações foram eliminadas para valores de m iguais a 1100 e 5100, e *burn in* de tamanho 1000 para m igual a 11000 e 51000. Considerou-se, também, saltos de tamanho 10 entre as amostras geradas para eliminar a correlação.

A Tabela 4.3 mostra o valor da evidência $EV(\Theta_0, \mathbf{x})$ em favor da hipótese H_0 de homogeneidade, segundo os valores de **n** e **m** adotados. Observa-se que a evidência em favor de H_0 é pequena, para $n \geq 100$ e quando **m** cresce.

Tabela 4.5 *Valores da Evidência e dos Parâmetros Estimados Segundo o Tamanho Amostral* (n) *e o Número de Amostras Geradas* (m).

	Número de Amostras									
Tamanho amostral	**1100**					**5100**				
	p_1	μ_1	μ_2	σ^2	Evidência	p_1	μ_1	μ_2	σ^2	Evidência
30	0,04	0,95	32,47	56,98	1,00	0,05	2,41	32,57	57,36	1,00
50	0,08	4,99	32,48	32,79	1,00	0,02	1,02	32,58	35,36	1,00
100	0,44	22,01	34,38	20,91	0,28	0,56	28,23	35,11	15,92	0,07
200	0,57	29,64	35,32	11,05	<0,001	0,58	29,60	35,27	10,77	<0,001
Tamanho amostral	**11000**					**51000**				
	p_1	μ_1	μ_2	σ^2	Evidência	p_1	μ_1	μ_2	σ^2	Evidência
30	0,04	0,57	32,49	57,26	1,00	0,04	0,80	32,48	58,04	1,00
50	0,05	2,66	32,37	34,94	0,99	0,04	1,81	32,35	35,1	0,99
100	0,59	29,79	35,28	14,89	0,03	0,59	29,80	35,27	14,90	0,03
200	0,59	29,61	35,30	10,82	<0,001	0,59	29,62	35,31	10,89	<0,001

A partir destes resultados, calculou-se o poder empírico como definido na seção 3.3 do Capítulo 3. O valor de corte e_c para a regra de decisão foi obtido a partir do nível de significância empírico fixado em 5%. Assim, 1.000 amostras foram geradas, sob a hipótese de homogeneidade, de uma mistura de duas distribuições Normais tomando-se $\mu_1 = \mu_2 = 30$. Em seguida, os 1.000 valores da $EV(\Theta_0, \mathbf{x})$, um para cada amostra gerada, foram obtidos e ordenados. O percentil de ordem 5 foi escolhido como ponto de corte e_c. Assim, a regra de decisão obtida foi:

i. Rejeita-se H_0 se $EV(\Theta_0, \mathbf{x}) \leq 0,0051$;

ii. Aceita-se H_0 se $EV(\Theta_0, \mathbf{x}) > 0,0051$.

Para a obtenção do poder empírico foram fixados diferentes valores para a diferença $\mu_2 - \mu_1$ e, para cada diferença, gerou-se 1.000 amostras e obteve-se a $EV(\Theta_0, \mathbf{x})$ associada. Com base na regra de decisão expressa acima, obteve-se a proporção de vezes em que a hipótese de homogeneidade foi rejeitada, ou seja o poder empírico. Na Tabela 4.4 tem-se o poder empírico, segundo a diferença $\mu_2 - \mu_1$. Observa-se que à medida que a diferença aumenta, a proporção de vezes que se rejeita H_0, quando de fato ela é falsa, aumenta e tende para 1, ou seja, o FBST consegue evidenciar que a amostra provém de uma mistura de distribuições Normal. A Figura 4.8 mostra o valor médio das evidências em favor de H_0 para cada diferença, com base nos 1.000

valores obtidos em cada uma delas. As Figuras 4.9 a 4.12 mostram a convergência da medida de evidência para diferentes valores da diferença $\mu_2 - \mu_1$.

Tabela 4.6 *Poder Empírico do Teste FBST, Segundo a Diferença* $\mu_2 - \mu_1$.

μ_1	μ_2	$\mu_2 - \mu_1$	**Poder Empírico**
30	32	2	0,062
30	34	4	0,335
30	36	6	0,879
30	38	8	0,997
30	40	10	1,000

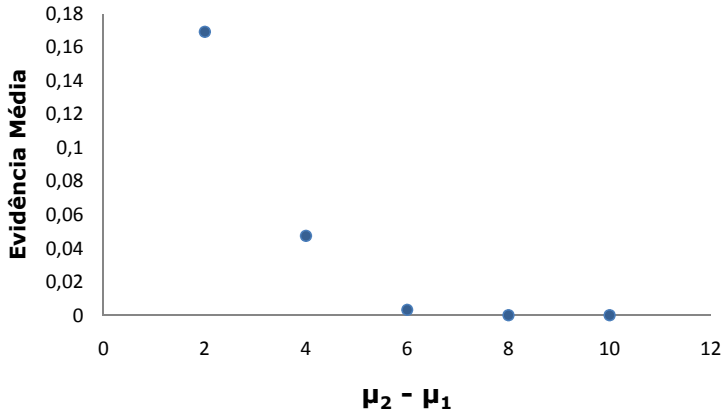

Figura 4.8 *Evidência Média, Segundo a Diferença* $\mu_2 - \mu_1$.

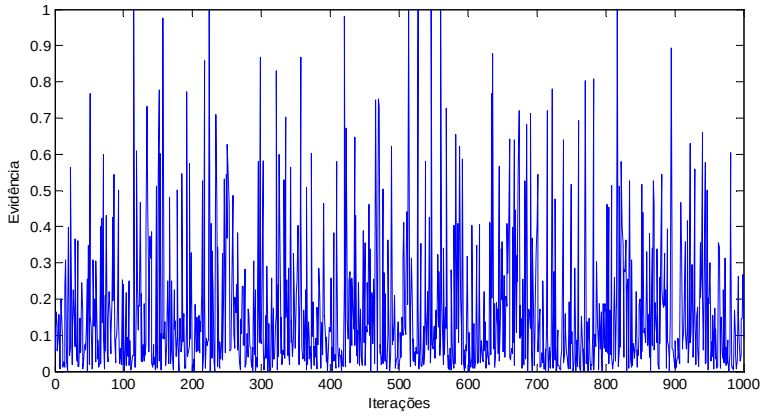

Figura 4.9 *Convergência da Evidência para* $\mu_2 - \mu_1 = 2$.

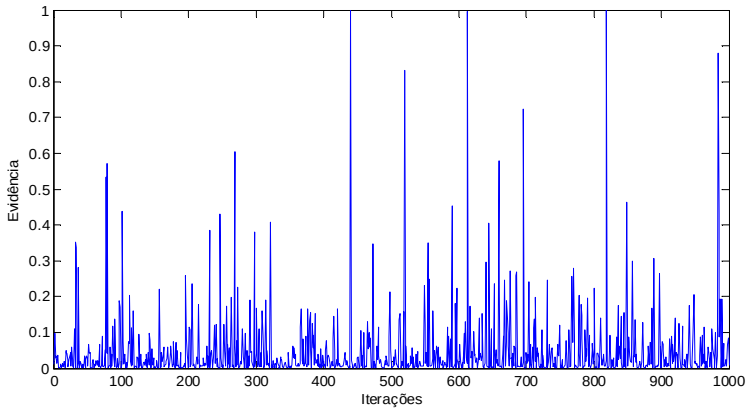

Figura 4.10 *Convergência da Evidência para* $\mu_2 - \mu_1 = 4$.

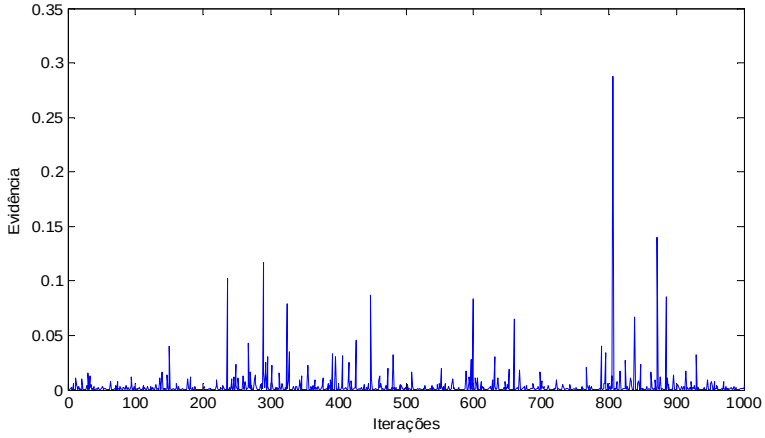

Figura 4.11 *Convergência da Evidência para* $\mu_2 - \mu_1 = 6$.

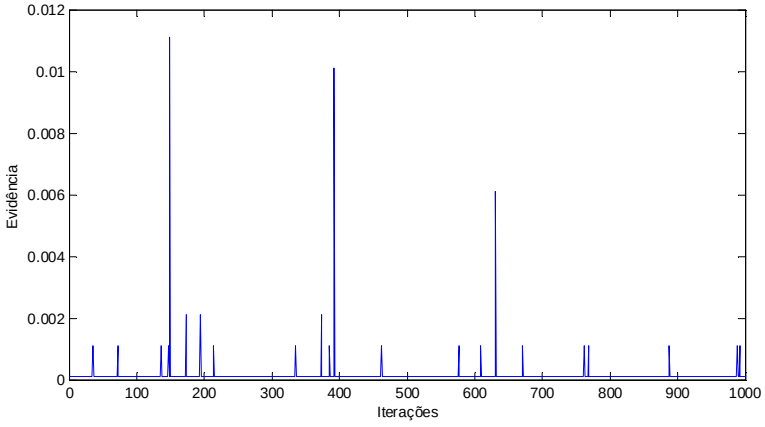

Figura 4.12 *Convergência da Evidência para* $\mu_2 - \mu_1 = 8$.

4.2.2 Aplicação a Dados Reais

A fim de analisar o comportamento do FBST em dados reais, foram utilizados os dados analisados em Marin *et al.* (2005) referentes ao estudo da galáxia feitos inicialmente por Roeder (1990) . Trata-se de 82 observações das velocidades das galáxias onde os autores consideram que as velocidades delas são realizações de variáveis aleatórias independentes e identicamente distribuídas com uma mistura de distribuições norrmais. No entanto, era necessário primeiramente comprovar a heterogeneidade entre as velocidades das galáxias. Os dados são encontrados em Roeder (1990).

Marin *et al.* (2005) analisou os dados considerando uma mistura de três distribuições normais, onde eles ajustaram um modelo de mistura para descrever o problema e estimaram os parâmetros μ_j, σ_j^2, p_j para cada distribuição. Neste trabalho, o uso do FBST é feito considerando a hipótese nula de homogeneidade contra a hipótese alternativa de mistura com duas componentes. Os resultados também mostram que a evidência em favor da hipótese de homogeneidade H_0 é abaixo de 0,0001, reforçando os estudos feitos e concluídos pelos autores que as galaxias são compostas por misturas de distribuições, assim como obtido em Marin *et al.* (2005). A Tabela 4.9 apresenta as estimativas dos parâmetros obtidas anteriormente em Marin *et al.* (2005) para o caso de uma mistura de três distribuições e aquelas atuais, obtidas neste trabalho, bem como a medida de evidência, verificando-se que nas duas propostas a proporção estimada para a segunda componente é acima de 0,80.

Destaca-se a importância deste estudo principalmente na área da astronomia, uma vez que devido à expansão do universo, as galáxias se movimentem em velocidades maiores, despertando o interesse em muitos pesquisadores que tem se preocupado em analisar o comportamento da alteração dessas velocidades. Mais detalhes podem ser encontrados em Roeder (1990).

Tabela 4.7 *Estimativas dos Parâmetros Obtidas no Trabalho Anterior e no Atual, e a Medida de Evidência em Favor da Hipótese de Homogeneidade.*

	Estimativas	
Parâmetros	**Marin *et al.* (2005)**	**Atual**
μ_1	9,50	9,68
μ_2	21,40	21,84
μ_3	26,80	-
σ_1^2	1,90	24,85*
σ_2^2	6,10	-
σ_3^2	34,10	-
p_1	0,09	0,10
p_2	0,85	0,90
p_3	0,06	-
Medida de Evidência (FBST)		<0,001

*Variâncias iguais.

Capítulo 5

Considerações Finais

5.1 Considerações Finais

Neste trabalho, o procedimento *Full Bayesian Significance Test* (FBST) proposto por Pereira e Stern (1999), foi usado para calcular a evidência em favor da hipótese nula de homogeneidade em problemas de mistura finita de distribuições, considerando misturas de modelos Poisson e normal com o número de componentes na mistura fixado. Nos dois casos estudados o FBST apresentou um bom desempenho, tanto para dados simulados como em dados reais disponíveis na literatura. Na implementação do FBST foram geradas amostras da distribuição posterior dos parâmetros do modelo via amostrador de Gibbs. Os resultados para dados simulados foram validados através do poder empírico do teste, que mostrou comportamento satisfatório, uma vez que o poder empírico aumentou à medida que os dados gerados se distanciavam mais da hipótese nula.

Em problemas envolvendo mistura de modelos, o maior interesse é determinar K, o número de componentes na mistura, que geralmente é desconhecido. Alguns autores abordam este problema incluindo K no vetor de parâmetros do modelo, e usando métodos MCMC com saltos reversíveis para sua estimação (Richardson e Green, 1997). Outra abordagem é via métodos de seleção de modelos, em que se comparam modelos com diferentes números de componentes na mistura, usando algum critério de seleção. Lauretto (2007) propõe, neste contexto, usar o FBST para a escolha do número de componentes K, e apresenta resultados satisfatórios no seu trabalho, que envolve modelos mais complexos.

Assim, o FBST mostra-se uma ferramenta eficaz no estudo de mistura finita de distribuições, com número de componentes fixado, nos casos estudados. Sua utilização em modelos multivariados de mistura foi estudado em Lauretto (2007) e mostrou bons resultados.

5.2 Trabalhos Futuros

Recomenda-se como tópicos para pesquisas futuras:

i. Fazer um estudo de sensibilidade do FBST para os modelos estudados considerando diferentes distribuições a priori;

ii. Estudar o desempenho do FBST para modelos de mistura finita envolvendo outras distribuições de probabilidade;

iii. Usar o FBST para modelos de mistura com número desconhecido de componentes;

iv. Usar o FBST em estudos de misturas, composta por distribuições de probabilidade diferentes.

Referências Bibliográficas

Berger, J. O.; Selke, T. (1987). Tesnting a Point Null Hypothesis: The Irreconcilability of P values and Evidence. Journal of the American Statistical Association, 82, 112 - 139.

Berger, J. O.; Delampady, M. (1987). Testing Precise Hypothesis: Statistical Science, 2, 317 - 352.

Bolfarine, H.; Sandoval, M. C. (2002). Introdução à Inferência Estatística. SBM.

Chen, J. (1998). Penalized Likelihood - ratio Test for Finite Mixture with Multinomial Observations. Canad. Journal of Stastistal, 26, 583 - 599.

Chen, H.; Chen, J. (1998). The Likelihood Ratio Test for Homogeneity in the Finite Mixture Models. Techinical Report STAT 98-09. Departament of Statistics and Actuarial Science, University of Waterloo, Waterloo.

Chen, J.; Kalbeisch, J. D. (2004). Modified Likelihood Ratio Test in Finite Mixture Models with a Structural Parameter. Journal of Statistical Planning and Inference 129 (2005) 93 - 107.

Chen, H., Chen, J., Kalbeisch, J.D., (2001). A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models. Journal of the Royal Statistical Society. B, 63, 19-29.

Chen, H.; Chen, J.; Kalbfeisch, J.D. (2002). Testing for a Finite Mixture Model with Two Components. Statistics and Actuarial Science Technical Report #2001-02, University of Waterloo.

Dellaportas, P.; Karlis, D.; Xekalaki, E. (1997). Bayesian Analysis of Finite Poisson Mixtures. Department of Statistics. Athens University of Economics and Business, 76 Patission Str. Athens, Greece.

Diebolt, J.; Robert, C. (1994). Estimation of Finite Mixture Distributions Through Bayesian Sampling. Journal of the Royal Statistical Society B 56, 363-375.

Dempster, A.P.; Laird, N.M.; Rubin D.B. 1977. Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39: 1-38.

Ehlers, R. S. (2003). Introdução a Inferência Bayesiana, Versão Revistada em junho de 2003.http://www.leg.ufpr.br/ paulojus/CE227/ce227.pdf.

Hasselblad, V. (1969). Estimation of Finite Mixtures from the Exponential Family. Journal of the American Statistical Association, 64, 1459-1471.

Laureto, M. S. (2007). Seleção de Modelos Através de um Teste Genuinamente Bayesiano: Misturas de Normais Multivariadas e Hipóteses Separadas. Tese de Doutorado. São Paulo. Universidade de São Paulo (USP).

Li, P. (2007). Hypothesis Testing in Finite Mixture Models. Tese de Doutorado. University of Waterloo, Ontario, Canada.

Lindsay B. (1995). Mixture Models: Theory, Geometry and Applications. Regional Conference Series in Probability and Statistics, Vol 5, Institute of Mathematical Statistics and American Statistical Association.

Madruga, M. R.; Esteves, L. G.; Wechsler, S. (2001). On the Bayesianity of Pereira-Stern tests. Test, 10, 291-299.

Madruga, M. R. (2002). Teste de Significância: Uma Proposta Genuinamente Bayesiana. Tese de Doutorado. São Paulo. Instituto de Matemática e Estatística (IME-USP).

Madruga, M. R.; Pereira, C. A. B. (2005). Power of FBST: Standard Examples. Instituto Interamericano de Estadística. Estadística (2005), 57, 168 y 169, pp. 1-9.

Magalhaes, M. N.; Lima, A. C. P. (2008). Noções de Probabilidade e Estatística. 6/3. ed. São Paulo: Edusp, 2008. v. 1.

Marin, J. M.; Mengersen, K.; Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. Handbook of Statistics, 25(16), 459-507.

McLachlan, G.; Basford, K. (1988). Mixture Models: Inference and Application to Clustering.

Min, D. (1998). The Null Distribution of the Likelihood Ratio Test for a Mixture of Two Gammas. Journal of the Korean. Data & Information Science Society. Volume 9, Nº. 2 pp. 289 - 298.

Paulino, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. (2003) Estatística Bayesiana, Lisboa:Fundação Calouste, Gulbenkian, Portugal.

Pereira, C. A. B; Stern, J. M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. Entropy 1: 99-110.

Pissini, C. F. (2006). Aplicações em Meta - Análise sob um Enfoque Bayesiano Usando Dados Médicos. Dissertação de Mestrado. Santa Catarina. Universidade de Santa Catarina (UFScar).

Qin, Y.; Smith, B.; Lei, Q. (2009). Test for Homogeneity in Normal Mixtures with Unknown Means and Variances. Journal of Statistical Planning and Inference, 139, 4165 - 4178.

Richardson S.; Green P. (1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). Journal of the Royal Statistical Society, Series B, 59, 731-792.

Roeder, K. (1990). Density Estimation with Confidence Sets Exemplified by Superclusters and Vaids in the Galaxies. Journal of the American Statistical Association, Vol. 85, No. 411, 617-624. Aplication and Case Studies.

Saito, M. Y.; Rodrigues J. (2005). Análise Bayesiana de Dados de Contagem com Excesso de Zeros e Uns. Revista Matemática e Estatística. São Paulo, v. 23, n.1, p. 47-57.

Titterington, D.; Smith, A.; Makov, U. (1985). Statistical Analysis of Finite Mixture Distributions. John Wiley, New York.