Seleção de Modelos Através de um Teste de Hipótese Genuinamente Bayesiano: Misturas de Normais Multivariadas e Hipóteses Separadas

Marcelo de Souza Lauretto

TESE APRESENTADA
AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE DOUTOR

Orientador: **Prof. Dr. Júlio Michael Stern**

Durante a elaboração deste trabalho o autor recebeu apoio financeiro da CAPES

São Paulo, outubro de 2007

Resumo

Nesta tese propomos o Full Bayesian Significance Test (FBST), apresentado por Pereira e Stern em 1999, para análise de modelos de misturas de normais multivariadas. Estendemos o conceito de modelos de misturas para explorar outro problema clássico em Estatística, o problema de modelos separados.

Nas duas propostas, realizamos experimentos numéricos inspirados em problemas biológicos importantes: o problema de classificação não supervisionada de genes baseada em seus níveis de expressão, e o problema da discriminação entre os modelos Weibull e Gompertz – distribuições clássicas em análise de sobrevivência.

Abstract

In this thesis we propose the Full Bayesian Significance Test (FBST) as a tool for multivariate normal mixture models. We extend the fundamental mixture concepts to another important problem in Statistics, the problem os separate models.

In both methods, we perform numerical experiments based on important biological problems: the unsupervised classification of genes based on their expression profiles, and the problem of deciding between the Weibull and Gompertz models – two classical distributions widely used in survival analysis.

Agradecimentos

Gostaria de manifestar meu profundo agradecimento e admiração por algumas pessoas que me ajudaram e me influenciaram enormemente ao longo desses anos.

Aos Professores Júlio Michael Stern e Carlos Alberto de Bragança Pereira, pela imprescindível orientação, apoio e paciência; pela sua profunda competência e, principalmente, pelo privilégio de sua amizade.

Aos Professores Luis Fernandez Lopez e Eduardo Massad, da Disciplina Informática Médica da Faculdade de Medicina da USP, por terem me recebido e me apoiado no Programa de Pós-Graduação da FM/USP – no qual eu havia ingressado originalmente.

Aos Professores Sérgio Wechsler e Wagner de Souza Borges, que muito admiro pela sua grande competência; pelos excelentes cursos ministrados e pelas valiosas oportunidades de aprendizado que me propiciaram nos problemas interessantes em que pudemos trabalhar juntos.

À Professora Elaine Carrer, pela forma madura e competente como tem coordenado o Programa de Pós-Graduação em Bioinformática.

A todos os professores do Programa de Pós-Graduação em Bioinformática, cuja determinação tem garantido o êxito do Programa. Foi para mim uma honra conhecer alguns desses professores.

Ao Professor Basilio de Bragança Pereira, por ter gentilmente cedido os códigos-fonte das rotinas de testes de hipóteses separadas, e pelas importantes sugestões.

Ao colega Fabio Nakano, pelo espírito de coleguismo e altruísmo demonstrado inúmeras vezes; pelas oportunidades que tivemos de trabalhar juntos ao longo do doutorado; pelas valiosas sugestões em computação e pelas valiosas trocas de experiências realizadas nesses anos.

Ao colega Silvio Rodrigues de Faria Junior, pela valiosa parceria na implementação do FBST para problemas de hipóteses separadas, e pelas várias sugestões.

À Capes, pela imprescindível bolsa de estudos concedida.

Ao Sebastião Pinho Sousa e à Patricia Martorelli, da Comissão de Pós-Graduação, pela dedicação, atenção e presteza com que sempre nos atenderam e nos orientaram.

À minha esposa Renata, que amo muito e a quem agradeço pelo companheirismo, compreensão e apoio; e também por ter me ajudado a compreender os conceitos de Biologia que aprendi neste programa.

Aos meus pais que, mesmo à distância, têm me apoiado imensamente.

Aos queridos amigos com quem tive a alegria de conviver ao longo desses anos, e em especial: Claus, Eliany, Fábio Henrique, Fátima, Francisco, Jair, Karin, Luiz Carlos, Marco Aurélio, Rafael, Said e Valguima.

A todos, meu *muito obrigado*.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Considerações Iniciais

Modelos de misturas típicos são aqueles em que a função de densidade de probabilidade de uma observação x é descrita como uma combinação linear de várias densidades, ou seja,

$$f(x) = w_1 f(x \mid \psi_1) + w_2 f(x \mid \psi_2) + \cdots + w_m f(x \mid \psi_m) \ ,$$

onde $f(. \mid \psi)$ é uma dada família paramétrica de densidades indexadas por um (vetor) parâmetro ψ. O objetivo da análise é a inferência sobre o número m de componentes, bem como seus respectivos pesos e parâmetros, w_k e ψ_k. Essa classe de modelos surge em dois contextos distintos.

No primeiro, postulamos uma *população heterogênea* constituída por grupos $i = 1, 2, ..., m$ de tamanhos proporcionais a w_k. Supõe-se que cada observação pertence a um dos m grupos, porém com classificação desconhecida. O objetivo nesses casos é inferir o número m de classes (grupos coesos entre si e separados dos demais), suas proporções w_k na população, os parâmetros ψ_k de suas densidades de probabilidade e a classe de cada observação (ou as probabilidades estimadas da observação pertencer a cada uma das classes). Este é o contexto de *clustering* ou, sob a ótica da inteligência artificial, de *aprendizado não-supervisionado* – termo usado em contraposição ao contexto de *aprendizado supervisionado*, no qual o conjunto de dados fornecido ao sistema é previamente classificado por um especialista da área ou por algum critério pré-estabelecido [49].

Em modelos de misturas, os estimadores de máxima verossimilhança ou máxima posteriori tendem a privilegiar modelos desnecessariamente complexos, isto é, com muitas componentes. A solução usualmente adotada para contornar esse problema é a adoção de critérios de regularização, que buscam um balanço entre a complexidade do modelo e seu ajuste aos dados [13].

No segundo contexto, o modelo de misturas é visto como uma representação conveniente da densidade de probabilidade, quando não se encontram distribuições *standard* de probabilidades capazes de se ajustar suficientemente bem aos dados (por exemplo, quando a distribuição dos dados possui vários máximos locais). Nesses casos, usualmente o problema de inferência é estimar os parâmetros da mistura que melhor se ajusta aos dados.

Essa segunda visão de modelos de misturas sugere um modo de resolver um problema clássico em estatística, o problema de *modelos separados* (ou *hipóteses separadas*, que consiste em escolher qual dentre duas (ou mais) distribuições de probabilidade candidatas melhor se ajusta aos dados disponíveis. Neste tipo de domínio, várias questões importantes se colocam:

- Existe evidência de que os ajustes fornecidos pelas distribuições candidatas sejam significativamente diferentes?
- Qual a distribuição que melhor se ajusta aos dados?
- Escolhendo-se uma das distribuições, existe alguma evidência de não aderência dos dados àquela distribuição, mas sim a uma outra?
- Supondo que uma das distribuições seja a verdadeira, qual é a evidência fornecida pelos dados?
- Se nenhuma das distribuições for a verdadeira, qual é a evidência fornecida pelos dados?

Testes de hipóteses separadas (ou modelos separados), como são comumente chamados problemas dessa natureza, foram originalmente desenvolvidos por Cox [18, 19], e desde então diversas abordagens têm sido utilizadas. Em especial, diversas derivações do Fator de Bayes têm sido definidas [2, 12, 66]. Na maioria desses trabalhos busca-se confrontar a hipótese nula, de que os dados provêm de uma certa distribuição F, contra a hipótese alternativa, de que os dados provêm de uma distribuição G. Em todos esses métodos, assume-se que uma das distribuições candidatas é a verdadeira, mas tais métodos não são capazes de responder à última das questões levantadas no parágrafo anterior. Atkinson [7] analisou este problema usando uma formulação de mistura exponencial, enquanto Quandt [74] foi o primeiro a propor a formulação de mistura linear – que exploraremos em nosso trabalho.

1.2 Objetivos Iniciais

Nosso trabalho possui dois objetivos principais:

1. Propor um método para determinação do número de componentes (*clusters*) em um modelo de misturas, baseado em testes de hipóteses. Em linhas gerais, o teste consiste em um procedimento seqüencial, onde se comparam os ajustes de um modelo de $m-1$ componentes contra o modelo de m componentes, escolhendo-se este último somente se houver uma evidência alta de que seu ajuste é significativamente melhor do que o modelo mais simples.

2. Propor um método para seleção de modelos separados através da formulação de misturas. Essencialmente, consideramos uma mistura de m componentes, onde cada componente k possui uma função de densidade $f_k(.|\psi_k)$ de família distinta das demais componentes. A densidade de probabilidade de uma observação x é:

$$f(x \mid w_1 ... w_k, \psi_1 ... \psi_k) = w_1 f_1(. \mid \psi_1) + w_2 f_2(. \mid \psi_2) + ... + w_k f_m(. \mid \psi_k).$$

   A hipótese de que os dados de uma amostra provêm da distribuição f_k equivale à hipótese

$$H : w_k = 1 \wedge w_j = 0, j \neq k \ .$$

Em ambos os casos, o procedimento de teste de hipótese proposto é o FBST, Full Bayesian Significance Test [70], um teste intuitivo, com caracterização geométrica, e implementável diretamente através de técnicas de otimização e integração numérica.

1.3 Contribuições

Consideramos que nossos objetivos iniciais foram alcançados, onde as contribuições desenvolvidas foram:

1. Um método baseado em testes de hipóteses para a determinação do número de componentes em uma mistura de normais multivariadas – uma família clássica e com ampla aplicação. A implementação foi feita em C++, compilável no Gnu-GCC e portável tanto para Windows como para Linux.

   Nosso método foi comparado com alguns métodos previamente existentes, em três experimentos numéricos: dois experimentos baseados em dados simulados, inspirados em *datasets* públicos, e um experimento baseado em um banco de dados de microarranjos de cDNA, onde o objetivo era classificar um conjunto de genes com base em seus perfis de expressão gênica. Nos três experimentos, nossa proposta apresentou um desempenho superior, mostrando um grande potencial na área de aprendizado não-supervisionado – especialmente em problemas de classificação de genes.

2. Um método para seleção de modelos separados, também baseado em testes de hipóteses. Apresentamos a formulação geral desse método, e implementamos uma instância particular: a seleção entre os modelos Gompertz e Weibull – dois modelos de grande importância em análise de sobrevivência. A implementação foi feita em C++, compilável no Gnu-GCC.

   Em comparação o teste de Kolmogorov-Smirnov, nossa proposta apresentou um desempenho muito superior.

3. Formulação e implementação do Full Bayesian Significance Test para o teste de equilíbrio de Hardy-Weinberg envolvendo múltiplos alelos simultaneamente, um problema clássico em genética das populações e com grande importância prática.

   Através da análise dos gradientes da densidade a posteriori (Dirichlet), conseguimos demonstrar que a otimização da posteriori sob a hipótese – uma das etapas do cálculo da evidência no FBST – se reduz a um sistema de restrições lineares. O teste implementado a partir desse resultado supera as implementações anteriores (testes sequenciais ou uso de otimizadores numéricos) tanto em desempenho computacional como em estabilidade. A implementação foi feita em linguagem C e incorporada a um sistema de cálculo de probabilidade de paternidade, desenvolvido dentro do Programa de Inovação Tecnológico (PIPE) da Fapesp.

1.4 Organização da Tese

No Capítulo 2 apresentaremos os conceitos básicos sobre o FBST e dois estudos de casos em que temos trabalhado.

No Capítulo 3 apresentaremos a formulação geral de modelos de misturas, e nossa proposta para determinação do número de componentes através do FBST. Especial atenção é dada aos experimentos numéricos, cujos resultados obtidos se mostram coerentes e robustos.

No Capítulo 4 apresentaremos o problema de hipóteses separadas sob o contexto do FBST, com especial ênfase em um problema importante em teoria do envelhecimento: a discriminação entre os modelos Weibull e Gompertz. Especialmente no caso desses dois modelos, discutiremos de que forma o processo de envelhecimento está relacionado com as características estruturais e de construção de um sistema biológico ou tecnológico.

No Apêndice apresentamos as cópias de três artigos publicados com os resultados desta tese.

Capítulo 2

O Teste de Significância FBST

2.1 Definições

O FBST - Full Bayesian Significance Test - foi apresentado por Pereira e Stern [70], como um teste intuitivo, com caracterização geométrica, e que pode ser implementado através de técnicas de otimização e integração numérica. O método é considerado *totalmente Bayesiano* por requerer apenas o conhecimento do espaço paramétrico representado pela sua distribuição *a posteriori*. Uma vantagem do FBST é a de não necessitar de premissas adicionais, como probabilidade positiva para hipóteses precisas.

Nas discussões que seguem, assume-se que o espaço paramétrico, Θ, é um subconjunto de R^n, e a hipótese é formulada através de um subconjunto de Θ definido por restrições de igualdade e desigualdade:

$$H : \theta \in \Theta_H$$

onde

$$\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}.$$

O domínio de interesse do FBST são as hipóteses precisas, nas quais se tem ao menos uma restrição de igualdade, i.e., $dim(\Theta_H) < dim(\Theta)$. $f(\theta)$ é a função de densidade de probabilidade (pdf) posterior. Por simplicidade, no restante deste trabalho Θ_H será denotado simplesmente por H.

O cálculo da medida de evidência do FBST é realizado em dois passos:

1. Otimização da função de densidade a posteriori sob H:

$$f^* = \max_{\theta \in H} f(\theta \,|\, x) = f(\theta^* \,|\, x)$$

2. Integração da função de densidade a posteriori no conjunto tangente:

$$\overline{T} = \{\theta \in \Theta : f(\theta \,|\, x) > f^*\}$$

$$\overline{\text{ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f(\theta) d\theta$$

A medida $\overline{\text{ev}}(H)$ é a evidência contra H, e $\text{ev}(H) = 1 - \overline{\text{ev}}(H)$ é a evidência suportando (ou a favor de) H. Note que $\overline{T}$ é uma região de máxima densidade de probabilidade (HDP), cujo nível de credibilidade é $\overline{\text{ev}}(H)$. Faria Jr [28] mostra que $\overline{\text{ev}}(H)$ é uma medida de distância no intervalo $[0, 1]$ entre o ponto de máxima posteriori (irrestrita), $\widehat{\theta}$, e o ponto de máxima posteriori sob a hipótese, θ^*. Intuitivamente, um grande volume de $\overline{T}$ é indício de que a região sobre a hipótese (incluindo θ^*) possui baixa densidade de probabilidade, indicando "forte" evidência contra H.

Ao definir o conjunto tangente $\overline{T}$ levando em conta exclusivamente o ponto de máxima posteriori sob H, os autores compatibilizam o FBST com o princípio jurídico do "ônus da prova", segundo o qual, em um julgamento, deve ser dada a máxima credibilidade possível à defesa do acusado, cabendo ao acusador o ônus de provar que os argumentos do acusado são falsos [88].

Uma formulação mais geral da medida de evidência no FBST é a que segue. Seja $r(\theta)$ uma função de densidade definida em Θ.

$$\overline{\text{ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f(\theta \,|\, x) d\theta \ ,$$

$$\overline{T}(x) = \{\theta \in \Theta : s(\theta \,|\, x) > s^*\},$$

$$s^* = \max_{\theta \in H} s(\theta \,|\, x) = s(\theta^* \,|\, x),$$

$$s(\theta \,|\, x) = \left(\frac{s(\theta \,|\, x)}{r(\theta)} \right) \ .$$

A função $s(\theta)$ é conhecida como a surpresa a posteriori relativa à densidade de referência $r(\theta)$. Seu papel é tornar $\text{ev}(H)$ explicitamente invariante sob certas transformações sobre o sistema de coordenadas do espaço paramétrico [56]. A primeira formulação apresentada é um caso particular desta segunda, em que se adota a função de referência uniforme, $r(\theta) \propto 1$. Ao longo deste trabalho, trabalhamos com a formulação restrita do teste.

2.2 Determinação do nível de rejeição da hipótese nula

A decisão quanto a aceitar ou não a hipótese nula H depende de um nível crítico τ: rejeita-se H se a evidência contra H estiver acima desse nível crítico, $\overline{\text{ev}}(H) > \tau$, e aceita-se H caso contrário. Discutimos a seguir alguns critérios possíveis.

2.2.1 Análise de poder empírico

Este critério busca minimizar uma combinação entre os dois tipos de erros em testes de hipóteses:

Erro do Tipo 1 : é a taxa de rejeição da hipótese H quando esta é verdadeira;

Erro do Tipo 2 : é a taxa de aceitação da hipótese quando esta é falsa.

Denotamos por θ^* e $\widehat{\theta}$ as estimativas de máxima posteriori em H e em Θ, respectivamente. Pelo critério do poder empírico, são simulados dois conjuntos de amostras (da ordem de 500 ou mais), cada uma com aproximadamente o mesmo tamanho da amostra original sobre a qual se deseja testar a hipótese. Para um certo valor $\tau \in (0, 1)$, denotemos por $\alpha(\tau)$ a proporção de amostras do primeiro grupo tais que $\overline{\text{ev}}(H) \geq \tau$, e denotemos por $\beta(\tau)$ a proporção de amostras do segundo grupo tais que $\overline{\text{ev}}(H) < \tau$. (Note que $\alpha(\tau)$ e $\beta(\tau)$ são estimativas dos erros dos Tipos 1 e 2, respectivamente.)

O nível crítico τ é calibrado de forma a minimizar o erro total $c\ \alpha(\tau) + (1 - c)\ \beta(\tau)$, onde $c \in [0, 1]$ é uma constante pré-definida. Uma alternativa poderia ser definir τ em função apenas do erro do Tipo I, fixando τ tal que $\alpha(\tau) = 1\%, 5\%$, etc.

Este método possui a grande vantagem de definir τ dinamicamente, em função da amostra X original (e seu respectivo tamanho). Isso é especialmente útil quando a amostra é pequena em relação à dimensão do espaço paramétrico. Por outro lado, seu custo computacional pode ser elevado quando a dimensão do espaço paramétrico é alta.

A análise de poder empírico foi adotada em três de nossos trabalhos [50, 51, 91].

2.2.2 Análise assintótica

A demonstração da convergência assintótica da medida de evidência do FBST está sendo feita por Stern [87]. Denotando por θ^0 o valor verdadeiro do parâmetro e por $\overline{V}(c) = \Pr(\overline{\text{ev}} \leq c)$, dado θ^0, o resultado postula que, sob condições de regularidade e $n \to \infty$, pode-se afirmar que:

- Se H é falsa, $\theta^0 \notin H$, então $\overline{\text{ev}}$ converge (em probabilidade) para 1, i.e., $\overline{V}(c) \to 1$.
- Se H é verdadeira, $\theta^0 \in H$, então $\overline{V}(c)$, o nível de confiança é aproximado pela função
$$\overline{W}(t, h, c) = \text{Q}\left(t - h, \text{Q}^{-1}(t, c)\right)$$
onde $t = \dim(\Theta)$, $h = \dim(H)$ e $\text{Q}(k, x)$ é a distribuição chi-quadrada acumulada com k graus de liberdade.

Esta propriedade sugere um critério assintótico para rejeição de H com um nível de confiança $1-\alpha$, adotando-se

$$\tau = \overline{W}^{-1}(t, h, 1-\alpha),$$

ou seja, escolhendo-se τ tal que $\overline{W}(t, h, \tau) = 1-\alpha$.

A propriedade assintótica da evidência foi utilizada em dois de nossos trabalhos [52,53]. Embora tenha um custo computacional baixo, da mesma forma que em outros resultados assintóticos [94], este método não é recomendável quando o tamanho da amostra é pequeno em relação à dimensão do espaço paramétrico.

2.2.3 Funções de perda

Madruga, Esteves e Wechsler [55] mostraram a existência de funções de perda que tornam o FBST um procedimento de testes de hipóteses compatível com a Teoria da Decisão.

Denotando por $D = \{d_0 = \text{Aceitar } H \ , \ d_1 = \text{Rejeitar } H\}$ o espaço das decisões, os autores definem a seguinte função de perda L em $D \times \Theta$ (denominada LP_1):

$$L(d_1, \theta) = \left\{ \begin{array}{c} a \text{ se } \theta \in \overline{T}(x) \\ 0 \text{ c.c.} \end{array} \right. = a[1 - \mathbf{1}(\theta \in \overline{T}(x))] \ , \ a > 0$$

e

$$L(d_0, \theta) = \left\{ \begin{array}{c} b + c \text{ se } \theta \in \overline{T}(x) \\ b \text{ C.C.} \end{array} \right. = b + c \ \mathbf{1}(\theta \in \overline{T}(x))] \ , \ b \geq 0, \ c > 0.$$

Demonstra-se que o risco de aceitação e o risco de rejeição a posteriori são, respectivamente,

$$\begin{aligned} E[L(d_0, \theta)] &= b + c(1 - \text{ev}(H)) \\ E[L(d_1, \theta)] &= a \ \text{ev}(H) \ . \end{aligned}$$

Minimizar a função de perda consiste em aceitar H se, e somente se, $E[L(d_0, \theta)] < E[L(d_1, \theta)]$, ou seja, se

$$\text{ev}(H) > \frac{b + c}{a + c} \ .$$

Em um dos testes numéricos descritos adiante, adotamos o critério de aceitar H se $\text{ev}(H) > 0.5$, o que equivale a adotar a função de perda LP_1 com $a = c = 1, b = 0$.

2.3 Exemplo 1: Teste de equilíbrio de Hardy-Weinberg

Antes de discutirmos o exemplo desta seção, apresentaremos de forma bem sucinta alguns conceitos bastante importantes em genética das populações. Um *loco* (ou *locus*) é a posição

que um gene ocupa em um cromossomo de uma espécie. Um *alelo* é cada uma das formas alternativas de um gene ou marcador genético na espécie. O *genótipo* é a informação genética de cada organismo vivo. Em espécies com reprodução sexuada, o genótipo de um indivíduo em cada *locus* é determinado pela herança de um alelo da mãe e outro do pai naquele *locus*.

Em genética das populações, uma série de resultados teóricos e aplicações práticas advêm da possibilidade de se estabelecer relações entre as freqüências gênicas e as freqüências genotípicas em uma população. Isso é extremamente importante, já que o número de genótipos possíveis excede em muito o número de genes [20].

Um problema bastante atual em que esta relação precisa ser considerada é o cálculo de probabilidade de paternidade baseado em exames de DNA, cujo estudo detalhado é encontrado em [64]. Em linhas gerais, este problema consiste no seguinte: quando uma pessoa ou seu responsável requer em juízo o reconhecimento da paternidade por parte de um terceiro (pai putativo, ou demandado), recolhem-se amostras de DNA do filho, da mãe e do pai (na falta desse, recolhem-se amostras de DNA de seus parentes próximos). Observando-se os genótipos dos envolvidos, o estatístico deve responder se há uma evidência forte de paternidade, ou se eventuais semelhanças genética são apenas frutos do acaso. Para responder a essa questão, é necessário levar-se em consideração as freqüências populacionais dos alelos observados e calcular, a partir dessas freqüências, a probabilidade dos genótipos observados sob a hipótese de que o demandado não seja o pai. Se essa probabilidade for alta, não se pode afirmar a relação de paternidade. Por outro lado, se tal probabilidade for muito baixa, então tem-se um indício forte da paternidade. Veremos nesta seção em que condições se pode relacionar as freqüências genotípicas e alélicas, bem como a formulação do teste sob a perspectiva do FBST. Finalmente, apresentaremos a contribuição realizada com a extensão do teste para múltiplos alelos simultaneamente.

Consideremos o caso em que um *locus* possui dois alelos A e a, com freqüências populacionais p e q, respectivamente, com $p + q = 1$. Se os cruzamentos ocorrerem de forma aleatória entre os indivíduos, e se a população não estiver sofrendo ação de alguma das forças evolucionárias sobre esse *locus*, as freqüências genotípicas dos pares de alelos AA, Aa e aa em uma nova geração de indivíduos podem ser representadas pela Tabela 2.1. Essa tabela indica que esperamos observar na próxima geração as freqüências:

$$f(AA) = p^2, \qquad f(aa) = q^2, \qquad f(Aa) = 2pq,$$

tais que $p^2 + 2pq + q^2 = 1$.

Esse princípio foi originalmente discutido independentemente por Yule, Pearson e Castle (entre 1902 e 1904), para alguns valores particulares de p e q, e foi demonstrado no caso geral em 1908, independentemente por Hardy e por Weinberg, razão pela qual passou a ser denominado *princípio* (ou *equilíbrio*) *de Hardy-Weinberg* (vide referências em [20]).

Esse princípio ocorre sob as seguintes condições:

Tabela 2.1: Freqüências genotípicas esperadas sob condições de cruzamento aleatório dos indivíduos de uma população

	A (p)	a (q)
A (p)	AA (p^2)	Aa (pq)
a (q)	Aa (pq)	aa (q^2)

- Cruzamentos aleatórios entre os indivíduos da população, ou que não sejam influenciados pelo produto da expressão dos alelos;
- Número de indivíduos suficientemente grande para evitar grandes flutuações aleatórias entre as freqüências de duas gerações;
- Ausência de influência das forças evolucionárias, como seleção natural, mutações e migrações.

A generalização do equilíbrio de Hardy-Weinberg quando o *locus* possui mais de 2 alelos é obtida diretamente da expansão multinomial de grau 2 das freqüências alélicas populacionais. Tomando um *locus* com k alelos, os fatores da espansão de $(p_1 + \cdots + p_k)^2$ são:

$$(\underbrace{p_1}_{f(A_1)} + \cdots + \underbrace{p_k}_{f(A_k)})^2 = \underbrace{p_1^2}_{f(A_1A_1)} + \cdots + \underbrace{p_k^2}_{f(A_kA_k)} + \sum_{i \neq j} \underbrace{2p_ip_j}_{f(A_iA_j)}$$

onde $p_i = f(A_i)$ representa a freqüência alélica populacional para o gene A_i, e $f(A_iA_j)$ representa a freqüência do genótipo A_iA_j.

Uma vez em equilíbrio, as relações entre as freqüências genotípicas e as freqüências alélicas devem satisfazer:

$$f(A_iA_j) = f(A_i)f(A_j) \qquad \forall\ i, j \in \{1, 2, \ldots, k\}\,.$$

A formulação do teste do Equilíbrio de Hardy-Weinberg pelo FBST é feita da seguinte forma: Denotemos por $n_{ij}, i = 1...k, j = 1...i$ o número de indivíduos com o genótipo A_iA_j em uma amostra. Note que os casos $j = i$ são aqueles dos indivíduos monozigotos, ou seja, que herdaram o mesmo alelo da mãe e do pai. Aqui assumimos que as freqüências dos alelos são independentes de sexo, de forma que A_iA_j equivale a A_jA_i.

Se assumirmos que a probabilidade de um indivíduo possuir o genótipo A_iA_j é independente dos demais indivíduos da mesma geração, podemos considerar que as freqüências conjuntas $n_{ij}, i = 1...k, j = 1...i$ seguem uma distribuição multinomial com parâmetros n (tamanho total da amostra) e $\pi_{ij}, i = 1...k, j = 1...i$ (probabilidade de um indivíduo possuir o genótipo A_iA_j). Admitindo a distribuição a priori Dirichlet com parâmetros (1,1,...,1) (uniforme), a distribuição a posteriori para as freqüências genotípicas

$\pi_{ij} = f(A_i A_j)$ é dada por:

$$f(\pi) \propto \prod_{i=1}^{k} \prod_{j=1}^{i} \pi_{ij}^{n_{ij}}$$

onde π é o vetor das freqüências genotípicas π_{ij}. Sob a condição de equilíbrio de Hardy-Weinberg, as freqüências genotípicas π_{ij} são produtos das freqüências alélicas π_i, $i = 1, \ldots, k$,:

$$\pi_{ij} = \pi_i \pi_j \qquad \forall i,\, j \in \{1, 2, \ldots, k\},$$

onde $\sum_{i=1}^{k} \pi_i = 1$. Reescrevendo a distribuição a posteriori sob a hipótese, temos:

$$f(\pi \mid H) \propto \prod_{i=1}^{k} \prod_{j=1}^{i} (\pi_i \pi_j)^{n_{ij}} \tag{1}$$

Para calcular a evidência da hipótese de equilíbrio de Hardy-Weinberg, é necessário encontrar o ponto ótimo θ^* dentro da hipótese. O caso com apenas 2 alelos é bastante simples, e o máximo sob a hipótese é obtido por:

$$p = (2n_{11} + n_{21})/(2n), \quad q = 1 - p; \pi_{11} = p^2, \quad \pi_{21} = 2pq, \quad \pi_{22} = q^2.$$

Até recentemente, nos casos com mais de 2 alelos, o FBST vinha sendo utilizado de forma seqüencial, testando-se o primeiro alelo contra os demais, em seguida o segundo contra os demais, etc [69]. Todavia, nesta abordagem, após aceitarmos as hipóteses parciais, não sabemos qual é a evidência geral da hipótese. Considerando também que podem ocorrer dezenas de alelos distintos, o passo sequencial impõe um critério rigoroso para definição do nível crítico sob a hipótese [22].

Através da dedução analítica da otimização da posteriori sob a hipótese, desenvolvemos um método direto para testar simultaneamente todos os alelos, evitando assim os inconvenientes citados. Esta dedução é descrita abaixo.

Primeiramente deve-se notar que a Equação (1) pode ser reescrita separando-se os genótipos homozigotos (alelos iguais) dos heterozigotos (alelos diferentes):

$$f(\pi \mid H) \propto \prod_{j=1}^{k} \pi_j^{2n_{jj}} \prod_{i=2}^{k} \prod_{j=1}^{i-1} (2\pi_i \pi_j)^{n_{ij}}.$$

Maximizar $f(\pi \mid H)$ equivale a maximizar seu logaritmo:

$$\begin{aligned}
l(\pi) &= \log(f(\pi \mid H)) \\
&\propto \sum_{j=1}^{k} 2n_{jj} \log(\pi_j) + \sum_{i=2}^{k} \sum_{j=1}^{i-1} n_{ij} \log(2\pi_i \pi_j) \\
&= \sum_{j=1}^{k} 2n_{jj} \log(\pi_j) + \sum_{i=2}^{k} \sum_{j=1}^{i-1} n_{ij} (\log(2) + \log(\pi_i) + \log(\pi_j)) \\
&= h \log(2) + \sum_{j=1}^{k} n_j \log(\pi_j),
\end{aligned}$$

onde

$$h = \sum_{i=2}^{k} \sum_{j=1}^{i-1} n_{ij}, \quad n_j = \sum_{i=1}^{j} n_{ji} + \sum_{i=j}^{k} n_{ij}.$$

Pela restrição $\sum_{i=1}^{k} \pi_i = 1$, temos:

$$l(\pi) \quad \propto \quad h \log(2) + \sum_{j=1}^{k-1} n_j \log(\pi_j) + n_k \log\left(1 - \sum_{j=1}^{k-1} \pi_j\right).$$

As componentes do gradiente de $l(\pi)$ são dadas por:

$$\frac{\partial l}{\partial \pi_i}(\pi) \propto \frac{n_i}{\pi_i} - \frac{n_k}{\left(1 - \sum_{j=1..k-1} \pi_j\right)}$$

E portanto o ponto ótimo sob a hipótese H é o vetor π que satisfaz:

$$A.\pi = b; \qquad A = \begin{pmatrix} n_1 + n_k & n_1 & \dots & n_1 \\ n_2 & n_2 + n_k & & n_2 \\ \vdots & & \ddots & \vdots \\ n_{k-1} & n_{k-1} & \dots & n_{k-1} + n_k \end{pmatrix}; \qquad b = \begin{pmatrix} n_1 \\ \vdots \\ n_{k-1} \end{pmatrix}.$$

Assim, a otimização da posteriori sob a hipótese se reduz a um sistema de restrições lineares, para o qual encontra-se grande disponibilidade de bibliotecas [73].

Este novo teste foi implementado em linguagem C e foi incorporado a uma ferramenta computacional para cálculo de probabilidade de paternidade, desenvolvida dentro do Programa de Inovação Tecnológica (PIPE) da Fapesp (http://watson.fapesp.br/PIPEM /Pipe13/genet1.htm).

2.4 Exemplo 2: Avaliação de diferenças de expressão gênica em SAGE

Esta aplicação consiste em analisar as listas de freqüências de Tags de 3 bibliotecas de SAGE, extraídas de amostras de tecidos de pescoço e cabeça, sendo uma de tecido normal, uma de tecido de câncer do tipo N_+ e uma de câncer de laringe do tipo N_0 [21]. O nosso interesse é comparar as freqüências de cada Tag sob as três condições distintas, identificando as Tags diferencialmente expressas considerando as três condições simultaneamente. A seguir apresentamos a formulação e os resultados do teste, considerando o caso geral de k bibliotecas extraídas em condições distintas.

Usaremos a seguinte notação:

- m é o número total de Tags, e k o número de condições (amostras);

- $X = [x_{ij}], i = 1..m, j = 1..k$ é uma matriz tal que x_{ij} é a freqüência observada do Tag i na amostra j;
- $N_1, N_2, ..., N_k$ são as freqüências totais de Tags em cada amostra e N é a freqüência total de Tags em todas as amostras, ou seja, $N_j = \sum_{i=1}^{m} x_{ij}$ e $N = \sum_{j=1}^{k} N_j$;
- Y_i é a freqüência total da Tag i sob todas as condições, $Y_i = \sum_{j=1}^{k} x_{ij}$.

O modelo estatístico considerado é o de que o vetor de freqüências das Tags na biblioteca $j = 1..k$ (correspondente à coluna j da matriz X) é estatisticamente independente dos demais, e segue uma distribuição multinomial de m categorias e parâmetros N_j e $p_{1j}, p_{2j}, ...p_{mj}$, onde p_{ij} é nosso parâmetro de interesse – a probabilidade de ocorrência da Tag i na biblioteca j. Duas conseqüências naturais da suposição de multinomialidade são:

1. a freqüência individual de cada Tag em uma biblioteca segue distribuição binomial: $X_{ij} \,|\, p_{ij} \sim bin(N_j, p_{ij})$, $i = 1..m,\ j = 1..k$;
2. As variáveis aleatórias $X_{i1}, X_{i2}, ..., X_{ik}$ são mutuamente condicionalmente independentes dados os parâmetros $p_{i1}, p_{i2}, ..., p_{ik}$.

Na presença dos chamados "eventos raros", em que a probabilidade individual de cada possível resultado de um processo multinomial tem uma probabilidade baixa, é usual adotar-se a Distribuição de Poisson. Dadas as elevadas quantidades de Tags possíveis, parece razoável assumir que as distribuições amostrais das Tags, $X_{i1}, X_{i2}, ..., X_{ik}, Y_i$, seguem distribuições de Poisson com médias $N_j p_{ij}$ e $\sum_j N_j p_{ij}$.

Como o teste é realizado para cada Tag independentemente, deste ponto em diante consideraremos uma única Tag de interesse e eliminaremos o índice i da notação (que estará implícito).

Utiliza-se a seguinte parametrização:

$$\theta = \sum_j N_j p_j; \quad \pi_j = \frac{N_j p_j}{\theta}.$$

Se a hipótese $H : p_1 = p_2 = ... = p_k = p$ for verdadeira, esses parâmetros se transformam, sob H, em constantes conhecidas, exceto por p, ou seja,

$$\theta = Np; \quad \pi_j = \frac{N_j}{N}.$$

Por outro lado, conhecer uma observação de $X_1, ..., X_k$ equivale a conhecer a observação de $X_1, ..., X_{k-1}, Y$. Logo, se $x_1, ..., x_j$ é uma observação de $X_1, ..., X_k$ e $y = \sum_j x_j$, então:

$$Pr\{X_1 = x_1, ..., X_k = x_k \,|\, p_1, p_2, ..., p_k\} = Pr\{X_1 = x_1, ..., X_{k-1} = x_{k-1}, Y = y \,|\, p_1, p_2, ..., p_k\}$$

$$= Pr\{X_1 = x_1, ..., X_{k-1} = x_{k-1} \,|\, Y = y, p_1, p_2, ..., p_k\}\ Pr\{Y = y \,|\, p_1, p_2, ..., p_k\},$$

isto é, tem-se o produto de uma multinomial com parâmetro $(y, \pi_1, ..., \pi_k)$, calculada no ponto $x_1, ...x_k$, por uma Poisson com parâmetro θ, calculada no ponto y.

Com a equação acima, testar a igualdade da expressão gênica na Tag de interesse corresponde a testar a hipótese H descrita acima. Ao usarmos apenas o fator multinomial, testar H é testar $H' : (\pi_1, ..., \pi_k) = (N_1, ..., N_k)/N$. Note que H' especifica completamente os valores dos parâmetros.

Para o cálculo da evidência, deve-se notar que a verossimilhança no modelo condicional, multinomial, é proporcional a

$$L(\pi_1, ..., \pi_k \,|\, x_1, ..., x_k) \propto \prod_{j=1}^{k} \pi_j^{x_j},$$

e o conjunto tangente, se considerarmos como priori a distribuição uniforme no hipercubo $0 < \pi_j < 1, j = 1..k$, seria definido pela seguinte expressão:

$$\overline{T}(x) = \{\pi_1, ..., \pi_k : L(\pi_1, ..., \pi_k \,|\, x_1, ..., x_k) > f^*\},$$

onde $f^* = \prod_{j=1}^{k} N_j^{x_j}/N^y$. A evidência contra a hipótese, $\overline{\text{ev}}(H)$, será a integral de L no conjunto tangente, em relação à sua integral no espaço paramétrico completo, e a evidência a favor da hipótese será seu complemento: $\text{ev}(H) = 1 - \overline{\text{ev}}(H)$.

Reparametrizando em termos de p, a verossimilhança pode ser reescrita como

$$L(p_1, ..., p_k \,|\, x_1, ..., x_k) \propto \prod_{j=1}^{k} \left(\frac{N_j p_j}{S}\right)^{x_j},$$

onde $S = \sum_{j=1}^{k} N_j p_j$.

Nesta formulação, o teste é implementado facilmente. O ponto ótimo sob H é dado por $p_1 = ... = p_k = y/N$. A integração no conjunto tangente é feito via importance sampling, onde são sorteados pontos uniformemente no hipercubo $0 < p_j < 1, j = 1..k$, e os volumes da verossimilhança são computados no conjunto tangente e no espaço paramétrico completo. Apenas alguns cuidados devem ser tomados na implementação, para evitar instabilidade numérica – já que os expoentes na expressão da verossimilhança são muito elevados: deve-se usar sempre que possível os logaritmos das expressões, e também deve-se subtrair da log-verossimilhança em cada ponto o valor da log-verossimilhança máxima.

Adotamos este teste sobre um total de 53.899 tags, calculando a evidência para cada Tag. Para a aceitação/rejeição das hipóteses, adotamos o nível crítico τ de acordo com o critério assintótico apresentado na seção 2.2, com um nível de significância de 5%. Considerando que o modelo de misturas completo e o modelo restrito possuem respectivamente 3 e 1 graus de liberdade, temos $\tau = \overline{W}^{-1}(3, 1, 0.95) = 0.888$. Portanto, rejeitamos H se $\overline{\text{ev}}(H) > 0.888$, ou equivalentemente, se $\text{ev}(H) < 0.112$.

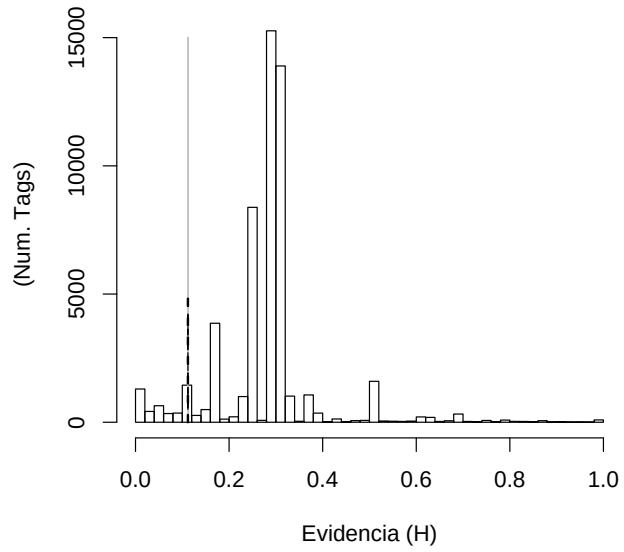


Figura 2.1: Histograma das evidências da hipótese de igualdade das proporções de Tags. A linha tracejada indica o nível de rejeição: se $\text{ev}(H) < 0.112$ então considera-se a Tag diferencialmente expressa. Um total de $8,3\%$ das Tags encontradas recaem neste grupo.

A Figura 2.1 mostra um histograma das evidências a favor hipótese de similaridade de expressão das Tags sob as três condições. O nível crítico para aceitação da hipótese está representado no histograma por uma linha vertical tracejada. 4.468 Tags foram apontadas pelo FBST como diferencialmente expressas, e esses resultados passarão agora por uma validação biológica.

Capítulo 3

FBST para Modelos de Misturas

3.1 Definições

Considere um modelo finito de dimensão d com m componentes (ou classes), e uma amostra $x^1, x^2 \dots x^n$ de tamanho n. Uma observação qualquer x^j pertence à classe k com probabilidade w_k. Assim, os pesos w_k fornecem a probabilidade de uma nova observação pertencer à classe k. Uma observação j de classe $k = c(j)$ é distribuída com densidade $f(x^j \mid \psi_k)$.

Neste parágrafo definiremos algumas notações gerais com matrizes. $1 : n$ indica o conjunto $\{1, \dots n\}$. Sejam h, i índices no conjunto $1 : d$, $k \in 1 : m$, e $j \in 1 : n$. Um array de matrizes terá um índice sobrescrito, como $S^1 \dots S^m$. Assim $S^k_{h,i}$ é o elemento da h-ésima linha e i-ésima coluna da matriz S^k. Escreveremos uma matriz retangular X, com o índice de linha subscrito e de coluna sobrescrito. Assim, x_i, x^j e x^j_i são a linha i, coluna j, e elemento (i, j) da matriz X. $\mathbf{0}$ e $\mathbf{1}$ são matrizes de zeros e uns, e $V > 0$ denota uma matriz positiva definida.

As classificações z^j_k são variáveis booleanas indicando se x^j pertence à classe k, i.e. $z^j_k = 1$ sse $c(j) = k$. Z não é observada, sendo por isso denominada *variável latente* ou *missing data*. Condicinado à variável latente, temos:

$$
\begin{aligned}
f(x^j \mid \theta) &= \sum_{k=1}^{m} f(x^j \mid \theta, z^j_k) f(z^j_k \mid \theta) &= \sum_{k=1}^{m} w_k f(x^j \mid \psi_k) \\
f(X \mid \theta) &= \prod_{j=1}^{n} f(x^j \mid \theta) &= \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f(x^j \mid \psi_k)
\end{aligned}
$$

Dados os parâmetros da mistura, $\theta = (w_1 ... w_k, \psi_1 ... \psi_k)$, e os dados observados, X, as

probabilidades condicionais de classificação, $P = f(Z \mid X, \theta)$, são:

$$p_k^j \;=\; f(z_k^j \mid x^j, \theta) \;=\; \frac{f(z_k^j, x^j \mid \theta)}{f(x^j \mid \theta)} \;=\; \frac{w_k f(x^j \mid \psi_k)}{\sum_{k=1}^m w_k f(x^j \mid \psi_k)}$$

Denotamos por y_k o número de exemplos de classe k, i.e., $y_k = \sum_j z_k^j$, or $y = Z\mathbf{1}$. A verossimilhança dos dados "completos", X, Z, é:

$$f(X, Z \mid \theta) \;=\; \prod_{j=1}^n f(x^j \mid \psi_{c(j)}) f(z_k^j \mid \theta) \;=\; \prod_{k=1}^m \left[{w_k}^{y_k} \prod_{j \mid c(j)=k} f(x^j \mid \psi_k) \right]$$

Nosso enfoque é sobre os modelos de misturas onde $f(x^j \mid \psi_k) = N(x^j \mid b^k, R^k)$, i.e., Normal com média b^k e matriz de precisão R^k (a matriz de precisão é a inversa da variância, $R^k = (V^k)^{-1}$).

3.2 Priori Dirichlet-Normal-Wishart

Considere a matriz aleatória X_i^j, $i \in 1:d$, $j \in 1:n$, $n > d$, onde cada coluna contém um elemento amostral de uma normal multivariada de dimensão d com parâmetros b (vetor de médias) e V (matriz de covariâncias), ou $R = V^{-1}$ (precisão). Denote-se por u e S as estatísticas:

$$\begin{aligned} u &= \frac{1}{n} \sum_{j=1}^n x^j = \frac{1}{n} X\mathbf{1} \\ S &= \sum_{j=1}^n (x^j - b) \otimes (x^j - b)' = (X - b)(X - b)' \end{aligned}$$

O vetor aleatório u segue uma distribuição normal com média b e precisão nR. A matriz aleatória S segue uma distribuição Wishart com n graus de liberdade e matriz de precisão R. As fdps Normal, Wishart e Normal-Wishart têm expressões:

$$\begin{aligned} N(u \mid n, b, R) &= (\frac{n}{2\pi})^{d/2} |R|^{1/2} \exp(-\frac{n}{2}(u-b)'R(u-b)) \\ W(S \mid e, R) &= c^{-1} |S|^{(e-d-1)/2} \exp(-\frac{1}{2}\mathrm{tr}(S\,R)) \text{ onde} \\ c &= |R|^{-e/2}\, 2^{ed/2}\, \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma(\frac{e-i+1}{2}) \end{aligned}$$

Agora considere a matriz X como acima, com média b e precisão R desconhecidas, e a estatística

$$S = \sum_{j=1}^{n} (x^j - u) \otimes (x^j - u)' = (X - u)(X - u)'$$

A família conjugada de prioris para distribuições normais multivariadas é a Normal-Wishart [22]. Adotamos como distribuição a priori para a matriz de precisão R a distribuição Wishart com $\dot{e} > d - 1$ graus de liberdade e matriz de precisão $\dot{S}$ e, dada R, adotamos como priori para b uma normal multivariada com média $\dot{u}$ e precisão $\dot{n}R$. Ou seja, usamos como priori a Normal-Wishart

$$NW(b, R \,|\, \dot{n}, \dot{e}, \dot{u}, \dot{S}).$$

Então, a distribuição a posteriori para R é uma distribuição Wishart com $\ddot{e}$ graus de liberdade e precisão $\ddot{S}$, e a distribuição posterior para b, dada R, é k-Normal com média $\ddot{u}$ e precisão $\ddot{n}R$. Assim, temos a posteriori Normal-Wishart:

$$\begin{aligned} NW(b, R \,|\, \ddot{n}, \ddot{e}, \ddot{u}, \ddot{S}) &= W(R \,|\, \ddot{e}, \ddot{S})\; N(b \,|\, \ddot{n}, \ddot{u}, R) \\ \ddot{n} &= \dot{n} + n \quad , \quad \ddot{e} = \dot{e} + n \quad , \quad \ddot{u} = (nu + \dot{n}\dot{u})/\ddot{n} \\ \ddot{S} &= S + \dot{S} + \frac{n\dot{n}}{n + \dot{n}} (u - \dot{u}) \otimes (u - \dot{u})' \end{aligned}$$

As matrizes de covariância e precisão devem ser positivas definidas, e prioris próprias satisfazem $\dot{e} \geq d$, e $\dot{n} \geq 1$. Prioris impróprias não informativas são dadas por $\dot{n} = 0$, $\dot{u} = 0$, $\dot{e} = 0$, $\dot{S} = 0$, i.e. com uma Wishart com 0 graus de liberdade como priori para R, e uma priori constante para b [22]. Então, a posteriori para R é uma Wishart com n graus de liberdade e precisão S, e a posteriori para b, dado R, é d-Normal com média u e precisão nR.

A priori conjugada para uma distribuição multinomial é uma distribuição Dirichlet:

$$\begin{aligned} M(y \,|\, n, w) &= \frac{n!}{y_1! \ldots y_m!} (w_1)^{y_1} \ldots (w_m)^{y_m} \\ D(w \,|\, y) &= \frac{\Gamma(y_1 + \ldots + y_k)}{\Gamma(y_1) \ldots \Gamma(y_k)} \prod_{k=1}^{m} {w_k}^{y_k - 1} \quad , \quad w > \mathbf{0} \;,\; w\mathbf{1} = 1 \end{aligned}$$

Se tivermos informação sobre o processo multinomial dado pelo parâmetro a priori $\dot{y}$, e então observamos y, nossa informação posterior é dada por $\ddot{y} = \dot{y} + y$. A priori não-informativa é dada por $\dot{y} = \mathbf{1}$.

Finalmente, para o modelo de misturas de normais com priori Dirichlet-Normal-Wishart, podemos escrever a posteriori,

$$\begin{aligned} f(\theta \,|\, X, \dot{\theta}) &\propto f(X \,|\, \theta) f(\theta \,|\, \dot{\theta}) \\ f(X \,|\, \theta) &= \prod_{j=1}^{n} \sum_{k=1}^{m} w_k N(x^j \,|\, b^k, R^k) \\ f(\theta \,|\, \dot{\theta}) &= D(w \,|\, \dot{y}) \prod_{k=1}^{m} NW(b^k, R^k \,|\, \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k) \\ p_k^j &= \frac{w_k N(x^j \,|\, b^k, R^k)}{\sum_{k=1}^{m} w_k N(x^j \,|\, b^k, R^k)} \end{aligned}$$

e posteriori completa:

$$\begin{aligned} f(\theta \,|\, X, Z, \dot{\theta}) &\propto f(\theta \,|\, X, Z) f(\theta \,|\, \dot{\theta}) = \\ &= D(w \,|\, \ddot{y}) \prod_{k=1}^{m} NW(b^k, R^k \,|\, \ddot{n}_k, \ddot{e}_k, \ddot{u}^k, \ddot{S}^k) \\ y &= Z\mathbf{1} \;,\;\; \ddot{y} = \dot{y} + y \;,\;\; \ddot{n} = \dot{n} + y \;,\;\; \ddot{e} = \dot{e} + y \\ u^k &= \frac{1}{y_k} \sum_{j=1}^{n} z_k^j x^j \;,\;\; S^k = \sum_{j=1}^{n} z_k^j (x^j - u^k) \otimes (x^j - u^k)' \\ \ddot{u}^k &= \frac{\dot{n}_k \dot{u}^k + y_k u^k}{\ddot{n}_k} \;,\;\; \ddot{S}^k = S^k + \dot{S}^k + \frac{\dot{n}_k y_k}{\ddot{n}_k} (u^k - \dot{u}^k) \otimes (u^k - \dot{u}^k)' \end{aligned}$$

Na implementação do FBST para modelos de misturas, o uso de prioris não informativas pode fazer com que o MCMC caia em estados onde uma ou mais componentes do modelo tem um número muito pequeno de pontos altamente colineares (i.e. com matriz de covariância singular) resultando em uma densidade posterior singular. Uma forma de evitar este tipo de situação é utilizar prioris minimamente informativas ao invés de prioris não informativas [80]. Em nosso trabalho, usamos os seguintes parâmetros a priori:

$$\dot{y} = \mathbf{1},\; \dot{n} = 1,\; \dot{u} = u,\; \dot{e} = d,\; \dot{S} = (1/n)S \;,$$

onde d é a dimensão de X.

3.3 Integração e Otimização

Para a integração da função a posteriori no espaço paramétrico, utilizamos o método de amostragem de Gibbs [38]. Dado θ, calculamos P. Dado P, $f(z^j \,|\, p^j)$ é uma distribuição

multinomial. Dada a variável latente, Z, tem-se as expressões da densidade a posteriori condicional para os parâmetros da mistura:

$$\begin{aligned} f(w \mid Z, \dot{y}) &= D(w \mid \ddot{y}) \\ f(R^k \mid X, Z, \dot{e}_k, \dot{S}^k) &= W(R \mid \ddot{e}_k, \ddot{S}^k) \\ f(b^k \mid X, Z, R^k, \dot{n}_k, \dot{u}^k) &= N(b \mid \ddot{n}_k, \ddot{u}^k, R^k) \end{aligned}$$

Um dos cuidados a se tomar durante a integração no modelo de misturas é a simetria do espaço paramétrico: dado um modelo de misturas, é possível obter um modelo equivalente simplesmente renumerando as componentes $1:m$ por uma permutação $\sigma([1:m])$. Stephens [86] sugere a adoção de critérios para a ordenação e permutação das componentes, a fim de garantir modelos identificáveis.

A etapa de otimização é realizada através do algoritmo EM, que maximiza a função log-posterior $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$ [81]. O algoritmo EM é derivado da log-verossimilhança condicional e da desigualdade de Jensen: se $w, y > \mathbf{0}, w'\mathbf{1} = 1$ então $\log w'y \geq w' \log y$.

Sejam θ a estimativa atual do MLE, $p_k^j = f(z_k^j \mid x^j, \theta)$ as probabilidades condicionais de classificação atuais, e $\tilde{\theta}$ a próxima iteração. O incremento da log-posteriori na iteração corrente é:

$$\begin{aligned} \delta(\tilde{\theta}, \theta \mid X, \dot{\theta}) &= fl(\tilde{\theta} \mid X, \dot{\theta}) - fl(\theta \mid X, \dot{\theta}) = \delta(\tilde{\theta}, \theta \mid X) + \delta(\tilde{\theta}, \theta \mid \dot{\theta}) \\ \delta(\tilde{\theta}, \theta \mid \dot{\theta}) &= fl(\tilde{\theta} \mid \dot{\theta}) - fl(\theta \mid \dot{\theta}) \\ \delta(\tilde{\theta}, \theta \mid X) &= fl(X \mid \tilde{\theta}) - fl(X \mid \theta) = \sum_j \delta(\tilde{\theta}, \theta \mid x^j) \\ \delta(\tilde{\theta}, \theta \mid x^j) &= fl(x^j \mid \tilde{\theta}) - fl(x^j \mid \theta) = \log \frac{f(x^j \mid \tilde{\theta})}{f(x^j \mid \theta)} \\ &= \log \frac{\sum_k \tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{f(x^j \mid \theta)} = \log \sum_k \left(\frac{p_k^j}{p_k^j} \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{f(x^j \mid \theta)} \right) \geq \\ \Delta(\tilde{\theta}, \theta \mid x^j) &= \sum_k p_k^j \log \frac{\tilde{w}_k f(x^j \mid \tilde{\psi}_k)}{p_k^j f(x^j \mid \theta)} \end{aligned}$$

Logo, $\Delta(\tilde{\theta}, \theta \mid X, \dot{\theta}) = \Delta(\tilde{\theta}, \theta \mid X) + \delta(\tilde{\theta}, \theta \mid \dot{\theta})$, é um limite inferior para $\delta(\tilde{\theta}, \theta \mid X, \dot{\theta})$. Também $\Delta(\theta, \theta \mid X, \dot{\theta}) = \delta(\theta, \theta \mid X, \dot{\theta}) = 0$, Assim, sob condições de diferenciabilidade, ambas as superfícies são tangentes, garantindo a converência para um máximo local, se o algoritmo for iniciado suficientemente perto desse máximo. Note que maximizar $\Delta(\tilde{\theta}, \theta \mid X, \dot{\theta})$ sobre $\tilde{\theta}$ é o mesmo que maximizar

$$Q(\tilde{\theta}, \theta) = \sum_{k,j} p_k^j \log \left(\tilde{w}_k f(x^j \mid \tilde{\psi}_k) \right) + fl(\tilde{\theta} \mid \dot{\theta})$$

Cada iteração do algoritmo EM é dividida em dois passos:

- passo E: Calcula $P = E(Z \mid X, \theta)$.
- passo M: Otimiza $Q(\tilde{\theta}, \theta)$, dado P.

Para o modelo de misturas gaussiano, com uma priori Dirichlet-Normal-Wishart,

$$\begin{aligned} Q(\tilde{\theta}, \theta) &= \sum_{k=1}^{m} \sum_{j=1}^{n} p_k^j \left[\log \tilde{w}_k + \log N(x^j \mid \tilde{b}^k, \tilde{R}^k) \right] + fl(\tilde{\theta} \mid \dot{\theta}) \\ fl(\tilde{\theta} \mid \dot{\theta}) &= \log D(\tilde{w} \mid \dot{y}) + \sum_{k=1}^{m} \log NW(\tilde{b}^k, \tilde{R}^k \mid \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k) \end{aligned}$$

O passo E é realizado como descrito anteriormente:

$$p_k^j = E(z_k^j \mid x^j, \theta) = \frac{w_k f(x^j \mid \psi_k)}{\sum_{k=1}^{m} w_k f(x^j \mid \psi_k)}$$

Para misturas de distribuições normais, o passo M é obtido analiticamente:

$$\begin{aligned} y &= P\mathbf{1} \;,\; \tilde{w}_k = \frac{y_k + \dot{y}_k - 1}{n - m + \sum_{k=1}^{m} \dot{y}_k} \\ u^k &= \frac{1}{y_k} \sum_{j=1}^{n} p_k^j x^j \;,\; S^k = \sum_{j=1}^{n} p_k^j (x^j - \tilde{b}^k) \otimes (x^j - \tilde{b}^k)' \\ \tilde{b}^k &= \frac{\dot{n}_k \dot{u}^k + y_k u^k}{\dot{n}_k + y_k} \;,\; \tilde{V}^k = \frac{S^k + \dot{n}_k (\tilde{b}^k - \dot{u}^k) \otimes (\tilde{b}^k - \dot{u}^k)' + \dot{S}^k}{y_k + \dot{e}_k - d} \end{aligned}$$

O EM é um algoritmo para otimização local, mas através da geração de pontos iniciais através do MCMC, temos um otimizador global estocástico [72].

3.4 Seleção de modelos

Neste trabalho, o problema de interesse é determinar o número de componentes (ou classes) da população a partir de uma amostra X. Assume-se que cada componente k segue uma distribuição normal multivariada, cujo vetor de média b^k e matriz de covariância V^k também devem ser estimados. Na formulação do problema pelo FBST, o espaço paramétrico Θ contém m componentes, e a hipótese a ser testada é a restrição de haver apenas $m - 1$ componentes, ou seja,

$$H = \{\theta = (w_1 \ldots w_m, \psi_1 \ldots \psi_m) \in \Theta : w_m = 0\}$$

O FBST seleciona o modelo de m componentes, rejeitando H, se a evidência contra a hipótese estiver acima do nível crítico $\overline{\text{ev}}(H) > \tau$, e seleciona o modelo de $m-1$ componentes, aceitando H, caso contrário. A determinação do número final de componentes é feita iterativamente, testando-se $m-1$ contra m componentes, para $m = 2, 3, ...$, até que $H : w_m = 0$ seja aceita. Denotando por m_f o menor valor de m para o qual a hipótese H é aceita, o modelo eleito será o de $m_f - 1$ componentes.

Em dois estudos de caso a serem apresentados nas próximas seções, comparamos o desempenho do FBST com o do *software Mclust* – Model-Based Clustering [31]. A escolha do Mclust como *benchmark* se deve a diversos fatores, sendo os principais: disponibilidade do pacote no ambiente R, facilidade instalação e utilização e aplicações bem sucedidas descritas na literatura [92]. O Mclust é baseado na formulação proposta por Banfield e Raftery [9], onde as matrizes de covariâncias das componentes são expressas em termos de suas decomposições autovalores/autovetores,

$$V^k = \lambda^k Q^k D^k (Q^k)' \ .$$

Q^k é a matriz (ortogonal) de autovetores, D^k é a matriz diagonal cujos elementos são proporcionais aos autovalores de V^k, e λ^k é um escalar. A matriz Q^k determina a orientação da componente, D^k determina sua forma e λ^k determina seu volume. Sistemas distintos de restrições sobre esses parâmetros (permitindo alguns mas não todos os parâmetros variar), resultam em um conjunto de modelos distintos [16]. O Mclust implementa alguns desses modelos possíveis. Para cada modelo e número de componentes m, os pesos das componentes w_k, as médias b^k e as covariâncias (restritas) V^k são estimadas via Algoritmo EM. No Mclust, a estrutura da matriz de covariância e o número de componentes são selecionados através do Critério Bayesiano de Informação (BIC), que consiste em uma aproximação para a verossimilhança integrada [83]:

$$BIC(M) = 2\Lambda(M) - \kappa(M) \log n,$$

onde Λ é a máxima log-verossimilhança sob o modelo M, κ é sua complexidade (número de parâmetros), e n o tamanho da amostra. O BIC é um critério de regularização, que faz um balanço entre o ajuste do modelo e seu número de parâmetros. Quanto maior o valor dessa função, mais forte a evidência a favor do modelo correspondente.

3.5 Resultados numéricos: *Iris virginica*

O primeiro teste de desempenho do FBST foi realizado sobre o conjunto de dados do gênero de flores *Iris*, disponível no ambiente R. Os dados originais foram coletados por Anderson [4], e consistem em quatro medidas (comprimento e largura da pétala, comprimento e largura da sépala) de um total de 150 exemplares coletados eqüitativamente de três espécies: *setosa*, *versicolor* e *virginica*. Um problema biológico comumente explorado

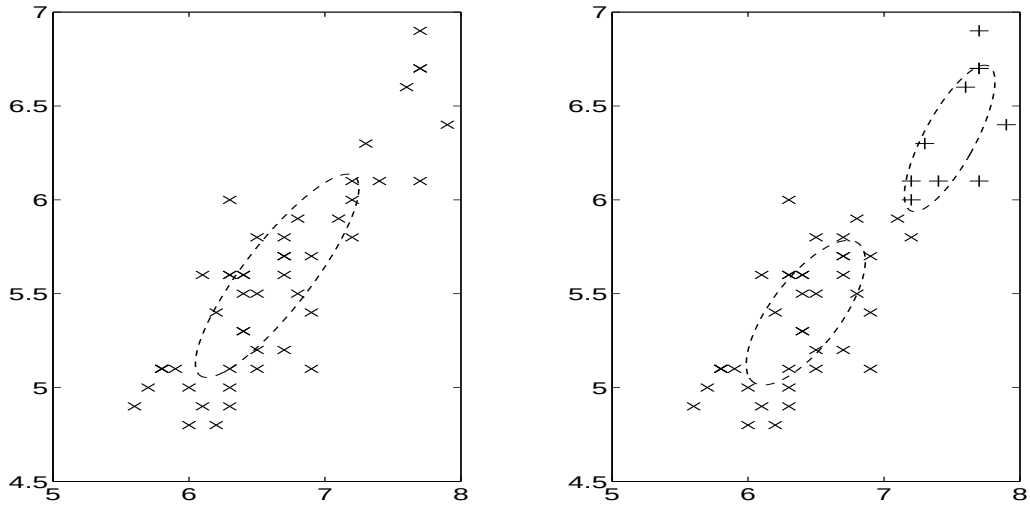


Figura 3.1: Dados públicos da espécie *Iris virginica* e curvas de nível dos modelos com uma componente (esquerda) e duas componentes (direita)

em análises de classificadores usando esse conjunto de dados é determinar se existem ou não dois subgrupos distintos da espécie *Iris virginica* [4, 60]. Dessa forma, a exemplo de Stephens [86], selecionamos 49 exemplares da espécie *Iris virginica* – um exemplar foi descartado, por ter sido considerado *outlier*. As medidas selecionadas foram o comprimento da pétala e da sépala. Nesse estudo de caso, nosso interesse é realizar uma análise de sensibilidade da precisão do FBST em função do tamanho da amostra.

Sob a formulação do FBST, o modelo paramétrico consiste em duas componentes ($m = 2$), e a hipótese a ser testada é a restrição de haver apenas uma componente ($H : w_2 = 0$). A Figura 3.1 apresenta os dados originais, juntamente com as curvas de níveis da densidade posterior dos parâmetros ótimos dos modelos com 1 componente (θ^*) e com 2 componentes($\widehat{\theta}$). O FBST seleciona o modelo de duas componentes, rejeitando H, se a evidência contra H estiver acima de um certo nível crítico τ.

Neste exemplo, o nível crítico τ foi escolhido através da análise de poder empírico, onde se procura estabelecer um balanço entre os erros do tipo 1 (rejeição de uma hipótese verdadeira) e do tipo 2 (aceitação de uma hipótese falsa). No presente caso, isso foi feito da seguinte forma: denotemos por θ^* e $\widehat{\theta}$ os parâmetros de máxima posteriori (MAP) do modelo sob a hipótese (1 componente) e do modelo irrestrito (2 componentes), ajustados sobre os dados da Iris. Foram geradas duas coleções de $t = 500$ amostras simuladas de tamanho n cada uma, sendo a primeira coleção simulada usando o parâmetro θ^* e a segunda simulada sob o parâmetro $\widehat{\theta}$. $\alpha(\tau)$ e $\beta(\tau)$, são os erros empíricos do tipo 1 e do

tipo 2 (taxa de rejeição da hipótese na primeira coleção e taxa de aceitação da hipótese na segunda coleção). O nível crítico τ foi escolhido de forma a minimizar o erro total, $(\alpha(\tau) + \beta(\tau))/2$.

Juntamente com o FBST, analisamos o desempenho de três critérios de seleção de modelos, que se baseiam no balanço entre o ajuste do modelo (verossimilhança) e sua complexidade (número de parâmetros) [13]. Nas definições abaixo, denotaremos por $\Lambda(M)$ a log-verossimilhança máxima do modelo M e por $\kappa(M)$ o número de parâmetros.

Akaike Information Criterion – AIC: Proposto por Akaike [3], este critério tem a forma:

$$AIC(M) = -2\Lambda(M) + 2\kappa(M) \ .$$

AIC3: Modificação proposta no AIC por Bozdogan [14]:

$$AIC3(M) = -2\Lambda(M) + 3\kappa(M) \ .$$

Bayesian Information Criterion – BIC: A forma tradicional de escolha de modelos na escola Bayesiana é calcular a verossimilhança integrada [46]. A verossimilhança integrada dos dados X dado o modelo M é:

$$pr(X \,|\, M) = \int pr(X \,|\, M, \theta)\pi(\theta \,|\, M)d\theta \ ,$$

onde $\pi(\theta \,|\, M)$ denota a *priori* para θ sob o modelo M. Uma forma alternativa proposta por Schwarz [83], é utilizar a aproximação

$$\log pr(X \,|\, M) = \log pr(X \,|\, M, \widehat{\theta}) - \frac{\kappa(M)}{2} \log n + O(1) \ ,$$

onde $\widehat{\theta}$ denota a estimativa de máxima verossimilhança para θ. Assim, o critério BIC é dado por

$$BIC(M) = -2\Lambda(M) + \kappa(M) \log n \ .$$

A Figura 3.2 apresenta as taxas de erro do FBST e dos critérios de regularização descritos, para amostras de tamanho $n = 50, 75, 100, 150$. O gráfico à esquerda mostra as taxas de erro do tipo 1, α; o gráfico do meio mostra as taxas de erro do tipo 2, β; o gráfico à direita apresenta as taxas médias de erro, $(\alpha+\beta)/2$. Nota-se no FBST uma convergência dos erros para zero mais rápida que a dos demais critérios. Para amostras pequenas, o BIC tem uma alta tendência a aceitar modelos de uma componente (alto erro do tipo 2). O AIC mostrou resultados melhores do que os demais critérios de regularização, alcançando a performance do FBST para amostras acima de $n = 150$.

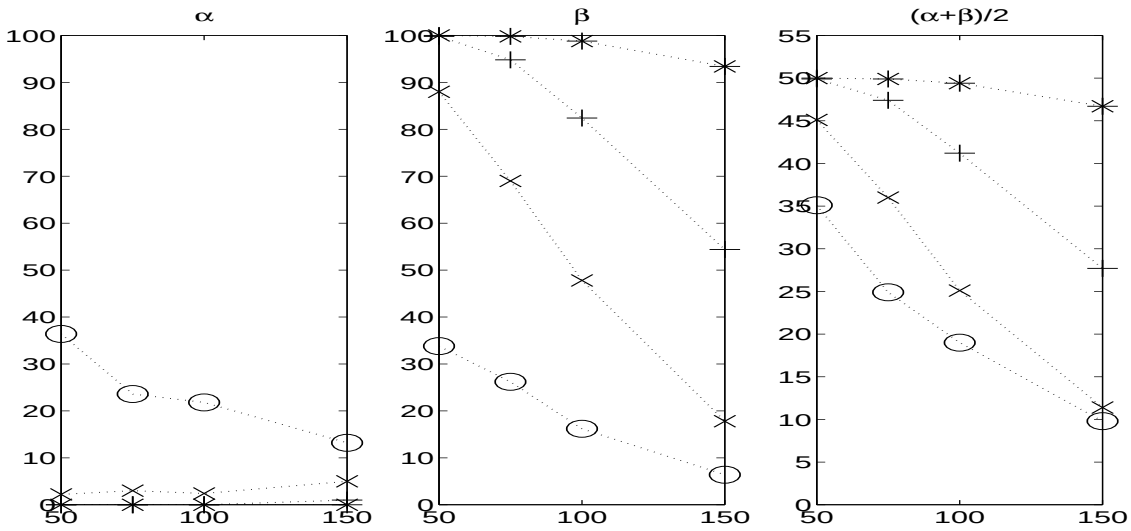


Figura 3.2: Dados *Iris virginica*: taxas de erro do tipo 1 (esquerda), tipo 2 (meio) e erro médio total (direita) em diferentes tamanhos de amostras. Critérios analisados: FBST (O), AIC ($\times$), AIC3 ($+$) e BIC ($*$).

3.6 Resultados numéricos: *Old Faithful*

Um segundo experimento numérico foi realizado a partir dos dados públicos do gêiser *Old Faithful* [39], composto por 272 observações de erupções desse gêiser localizado no Parque Nacional de Yellowstone (EUA). Cada observação consiste na duração da erupção e no tempo de espera até a próxima ocorrência. Um problema clássico em processos seqüenciais é determinar se os eventos são condicionalmente independentes e identicamente distribuídos (e, neste caso, sob a ótica de modelos de mistura, a população de eventos é formada apenas por uma componente) ou se existem sub-classes de eventos, distribuídos sob parâmetros distintos.

A Figura 3.3 apresenta dois gráficos de dispersão dos dados, cada ponto representando uma erupção. Uma vez que as observações formam claramente dois grandes agrupamentos, indicando a existência de pelo menos duas classes de erupção, são apresentados dois gráficos: o gráfico esquerdo mostra os dados classificados e as curvas de nível da densidade a posteriori sob o modelo de misturas ajustado com duas componentes; o gráfico direito apresenta as classificações e as curvas de nível da posteriori sob o modelo de misturas ajustado com três componentes. A classe prevista para cada observação é indicada pelo marcador do ponto ($+$, $\times$ ou $*$). As estimativas de máxima posteriori dos modelos de 2

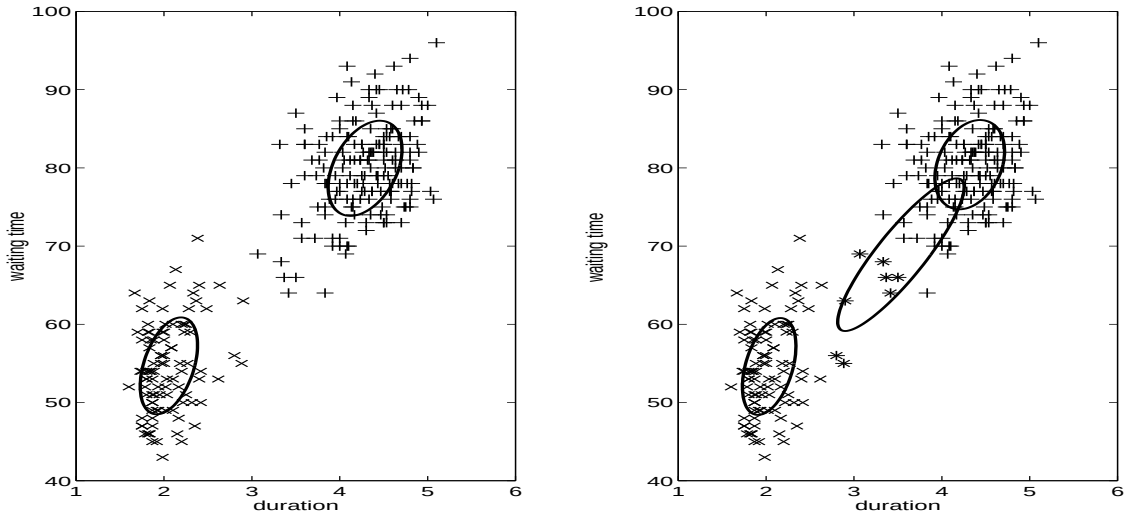


Figura 3.3: Dados de erupção do gêiser *Old Faithful* e curvas de nível dos modelos de misturas ajustados com 2 componentes (esquerda) e 3 componentes (direita).

e 3 componentes são denotadas por θ^* e $\widehat{\theta}$, respectivamente.

Para avaliação de desempenho do FBST, nosso interesse é estimar a proporção de acertos do teste em cada um dos dois possíveis cenários:

- quando os dados provêm de duas classes distintas, distribuídos sob o parâmetro θ^*; e
- quando os dados provêm de três classes distintas, distribuídos sob o parâmetro $\widehat{\theta}$.

Para aceitação ou rejeição da hipótese $H : w_m = 0$, adotamos o critério baseado em análise assintótica discutida na Seção 2.2, usando o nível de confiança $1 - \alpha$ para $\alpha = 1\%$. A tabela 3.1 apresenta as evidências críticas contra as hipóteses $w_2 = 0$, $w_3 = 0$ e $w_4 = 0$, onde t é o número de graus de liberdade do modelo de misturas completo (com m componentes), h é o número de graus de liberdade do modelo de misturas sob a hipótese (com $m - 1$ componentes) e τ é o quantil 99% da distribuição aproximada de $\overline{\text{ev}}(H)$. Definido o valor de τ para cada número de componentes, temos a decisão:

$$\text{Se } \overline{\text{ev}}(H) > \tau, \text{ rejeite } H\text{; caso contrário, aceite } H.$$

Dois experimentos numéricos foram realizados a partir dos parâmetros ajustados sobre os dados do *Old Faithful*, onde o desempenho do FBST foi comparado ao do MClust. No

Tabela 3.1: Níveis de evidência para rejeição da hipótese $H : w_m = 0$, para $m = 2, 3, 4$, definidos pelo critério assintótico.

m	t	h	$\tau = \overline{W} - 1(t, h, 0.99)$
2	11	5	0.885
3	17	11	0.532
4	23	17	0.182

Tabela 3.2: Números de amostras de 2 componentes (θ^*) e 3 componentes ($\widehat{\theta}$) segundo o número de componentes estimadas pelo FBST e pelo Mclust

Componentes	FBST		Mclust	
estimadas	θ^*	$\widehat{\theta}$	θ^*	$\widehat{\theta}$
1	0	0	0	0
2	499	96	500	287
3	1	404	0	213
4	0	0	0	0
Taxas de erro	0%	19%	0%	57%

primeiro experimento, avaliamos as taxas de erros (previsões a maior ou a menor) do número real de componentes, sobre amostras com o mesmo tamanho dos dados originais do *Old Faithful*. Para isso, foi sorteada uma coleção de 500 amostras de 272 pontos cada, sob o parâmetro θ^* (2 componentes), e uma segunda coleção de 500 amostras com 272 pontos cada, usando o parâmetro $\widehat{\theta}$ (3 componentes).

A Tabela 3.2 apresenta as freqüencias de amostras simuladas, segundo o número de componentes estimado pelo FBST e pelo Mclust. Cada coluna corresponde a uma das coleções, sob θ^* (2 componentes) e $\widehat{\theta}$ (3 componentes), e cada linha corresponde ao número estimado de componentes. A última linha contém a taxa de erro de cada método sobre cada coleção. Para amostras sorteadas com 2 componentes, o FBST acertou 499 e errou somente 1, enquanto o Mclust acertou todos os 500 casos. Por outro lado, para as amostras sorteadas com 3 componentes, o FBST teve uma taxa de erro de 19% (96 casos incorretos) contra 57% (287%) do Mclust, indicando um grande viés do Mclust em selecionar modelos mais simples, sob um alto custo em termos de precisão. É interessante observar que nem o FBST nem o Mclust elegeram modelos com 1 ou 4 componentes.

No segundo experimento numérico, procuramos analisar a convergência das taxas de erro do FBST e do Mclust para zero à medida em que o tamanho da amostra n aumenta. Para cada $n \in \{200, 300, 400, 500, 600\}$, simulamos duas coleções de 400 amostras de tamanho n, uma usando o parâmetro θ^* e outra usando o parâmetro $\widehat{\theta}$. A Tabela 3.3

Tabela 3.3: Convergência das taxas de erro com tamanhos crescentes de amostras.

Tamanho	% Erro FBST		% Erro Mclust	
amostra	θ^*	$\widehat{\theta}$	θ^*	$\widehat{\theta}$
200	1%	66%	0%	76%
300	1%	16%	0%	49%
400	1%	9%	0%	35%
500	1%	1%	0%	16%
600	1%	1%	0%	6%

apresenta as taxas de erro do FBST e do Mclust para cada uma das 10 coleções de amostras geradas, onde considera-se como taxa de erro estimada o percentual de amostras em que o método (FBST/Mclust) prevê um número de componentes diferente daquele sob o qual a amostra foi gerada. Observa-se que as taxas de erro do FBST são inferiores às taxas de erro do Mclust, bem como sua convergêngia para zero é consideravelmente mais rápida.

3.7 Classificação de genes baseada em níveis de expressão

O estudo de caso discutido nesta seção é baseado em um banco de dados de microarranjos de cDNA planejado e obtido por Ideker *et al.* [40]. Os autores analisaram a via de utilização de galactose (GAL) na levedura *Saccharomyces cerevisiae*, uma via bem conhecida (vide referências em [40]), onde as principais enzimas envolvidas são expressas somente na presença de galactose e na ausência de açúcares repressores, como glicose.

A utilização de galactose, representada na Figura 3.4, envolve uma via metabólica que converte a galactose em glicose-6-fosfato, e um mecanismo regulador que controla se a via está ligada ou desligada. Esse processo envolve pelo menos três tipos de proteínas. Um gene transportador (GAL2), um grupo de genes enzimáticos (GAL1, GAL7, GAL10 e GAL5) e genes reguladores (GAL3, GAL4, GAL6 e GAL80).

Foram preparadas 10 cepas de *Saccharomyces cerevisiae*, sendo uma selvagem (*wt*), e nove cepas geneteticamente modificadas, cada uma com a deleção completa de um dos genes citados acima (as cepas são denominadas $gal1\Delta$, $gal2\Delta$, $gal3\Delta$, $gal4\Delta$, $gal5\Delta$, $gal6\Delta$, $gal7\Delta$, $gal10\Delta$ e $gal80\Delta$). As variações na expressão de mRNA resultantes de cada perturbação foram analisadas através de microarranjos de DNA de aproximadamente 6200 genes. Em cada experimento, o cDNA marcado de uma cepa com perturbação foi hibridizada com o cDNA da cepa de referência (*wt*, crescimento em meio +gal). Para maior robustez, foram realizadas quatro hibridizações replicadas para cada perturbação.

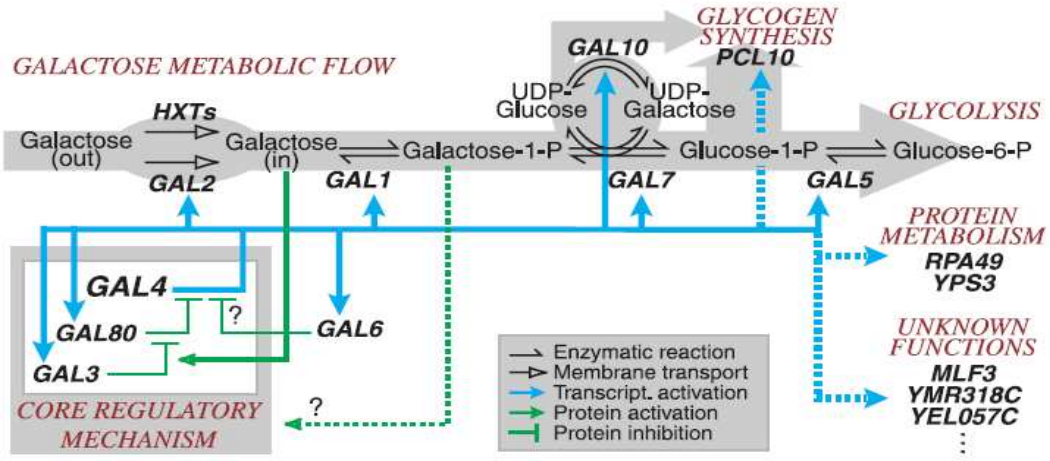


Figura 3.4: Modelo de utilização de galactose, e identificação dos principais genes envolvidos: *GAL2* (transporte), *GAL1, GAL7, GAL10, GAL5* (genes enzimáticos), *GAL4, GAL80, GAL3 E GAL6* (genes reguladores). Ideker *et al.* , 2001.

Sobre os dados originais, Yeung *et al.* [96] selecionaram um subconjunto de 205 genes cujos padrões de expressão refletem quatro categorias funcionais dentro da terminologia do consórcio *Gene Onthology* (GO) [6]. As quatro categorias funcionais e as correspondentes quantidades de genes contidos em cada categoria são apresentadas na Tabela 3.4. Sobre esse conjunto de dados, nosso interesse é agrupar os genes de forma não supervisionada (isto é, sem utilizar no método o conhecimento prévio das classes) e verificar:

- Se o FBST e o Mclust determinam corretamente o número de classes;
- O percentual de genes classificados corretamente, se o número previsto de classes estiver correto.

Os dados processados por Yeung *et al.* formam uma matriz de dimensão $205 \times 10 \times 4$ (205 genes, 10 cepas, 4 réplicas para cada cepa), onde $X_{i,j,r}$ corresponde ao nível de expressão normalizado do gene i na réplica r da cepa j. Tanto o FBST quanto o Mclust e a grade maioria dos algoritmos de *clustering* assumem que, dentro de cada agrupamento, as observações são (condicionalmente) independentes e identicamente distribuídas – o que não ocorre na presença de réplicas. Dentre as estratégias possíveis para tratar dados replicados, uma das mais simples e mais empregadas é usar a média das réplicas [96], tendo sido essa a nossa opção. Assim, obtemos uma matriz Y de dimensão 205×10, onde $Y_{i,j}$ corresponde ao nível médio de expressão do gene i nas 4 réplicas obtidas sob a condição j.

Tabela 3.4: Categorias funcionais das classes selecionadas por Yeung *et al.* [96], e número de genes na amostra.

Cod. classe	Categoria funcional	Quant. genes
1	biossíntese; metabolismo e modificação de proteínas	83
2	vias energéticas; metabolismo de carboidratos; catabolismo	15
3	metabolismo de nucleosídeos, nucleotídeos e ácidos nucleicos	93
4	transporte	14
Total		205

Em análises de dados envolvendo múltiplas variáveis, é muito comum que duas ou mais variáveis carreguem informações muito similares. Especificamente no tipo de estudo realizado por Ideker *et al.* , é possível que a inativação de dois genes possam ter efeitos similares nos níveis de mRNA da célula. De fato, observamos no presente conjunto de dados que o índice de correlação de Pearson dos níveis de expressão gênica dos 205 genes entre as cepas $gal7\Delta$ e $gal10\Delta$ é 0.92, indicando efeitos muito semelhantes da deleção desses genes no mRNA celular. Essa alta correlação explica-se pela interdependência entre esses dois genes [75], observável na Figura 3.4 (ambos estão envolvidos na conversão de Galactose-1-P em Glicose-1-P).

A presença de variáveis com alto nível de redundância frequentemente introduz efeitos indesejáveis (p. ex. *overfitting*, matrizes de covariância singulares, quebra de condições de regularidade) e usualmente precisa ser considerada [42, 47, 84]. Neste trabalho, para reduzir a dimensionalidade do espaço, utilizamos a Análise de Componentes Principais (PCA) [27], cuja idéia pode ser resumida como segue.

Dada a matriz Y $m \times d$ (m observações sobre d variáveis), e sua matriz de covariâncias V, a matriz de autovetores A de V é uma base ortogonal, e a matriz Z resultante do produto

$$Z' = A'X'$$

equivale a uma rotação das coordenadas (linhas) de X, de tal forma que $Cov(Z_{j_1,j_2}) = 0$ para $j_1 \neq j_2$. Ou seja, o produto de X pela matriz de autovetores de sua matriz de covariância é simplesmente uma mudança de coordenadas (rotação) que resulta em uma matriz cujas colunas são independentes. A operação inversa ($X' = A^{-1}Z'$) resulta nos dados originais, sem perda de informação. Isso sugere uma forma de redução da dimensão dos dados. Se duas variáveis (colunas) de X são altamente correlacionadas (negativa ou positivamente), é possível que $d-1$ dos d autovetores formem hiperplanos ortogonais tais que as distâncias (i.e. resíduos) entre as observações e os hiperplanos sejam relativamente pequenas. Então, se um dos autovetores for eliminado da matriz de autovetores A, a matriz Z resultante do produto $Z' = A'X'$ terá $d-1$ colunas; naturalmente, ocorre perda

de informação na transformação de X para Z, e essa perda de informação é proporcional às distâncias entre os pontos originais $X_{i,j}$ e os hiperplanos formados por A. Logo, se o objetivo é eliminar uma ou mais variáveis da matriz original, deve-se manter o conjunto de $d-1$ autovetores que, conjuntamente, explicam a maior parte da variância dos dados originais. Em PCA, utiliza-se usualmente um método incremental:

- Escolhe-se o autovetor cujo hiperplano minimiza o resíduo em relação aos dados originais (esse autovetor é aquele correspondente ao maior autovalor da matriz de covariância);

- A seguir, são escolhidos os próximos autovetores, em ordem decrescente de seus autovalores, incluindo-os em A até que todos os autovetores tenham sido selecionados ou até que um certo critério de parada seja satisfeito.

As colunas da matriz Z resultante do produto $Z' = A'X'$ são denominadas componentes principais. É comum que, na presença de variáveis originais altamente correlacionadas, a matriz resultante Z contenha uma grande parte da informação contida na matriz original, com um número consideravelmente menor de colunas. Vários critérios podem ser adotados para a seleção do número de componentes principais relevantes [29], sendo um deles parar a inclusão de componentes quando o percentual da variância acumulada nas componentes principais ultrapassar um certo limite $\alpha\%$ – usualmente 95% [42].

Nesta breve descrição, procuramos apenas comentar as idéias centrais da técnica de PCA, sem mencionar alguns aspectos específicos. Descrições mais detalhadas são amplamente apresentadas na literatura [27]. A análise de componentes principais é amplamente empregada nas mais diversas áreas, e um exemplo de sua aplicação em análise de dados de *microarrays* pode ser visto em [76].

Nos dados analisados, as variâncias percentuais acumuladas nas componentes principais são as seguintes: a primeira componente principal responde por 75.5% da variância dos dados; com $2, 3, 4...9$ componentes, as variâncias acumuladas são, respectivamente, 88.8%, 92.5%, 95.0%, 96.8%, 98.1%, 98.8%, 99.4% e 99.7%. Ou seja, temos 95% da informação original em apenas 4 componentes principais, e 98% da informação em 6 componentes principais.

Para que se tenha uma visão aproximada dos dados e se possa verificar informalmente o nível de segregação das quatro classes funcionais, apresentamos na Figura 3.5 as projeções dos níveis de expressão dos 205 genes nas duas primeiras componentes principais, e suas respectivas categorias.

Nossa análise comparativa do desempenho do FBST e do Mclust consistiu em aplicar os dois métodos sobre os dados originais e sobre os dados processados com 2, 3 ... 9 e 10 componentes principais. O objetivo foi verificar em que casos cada método previu o

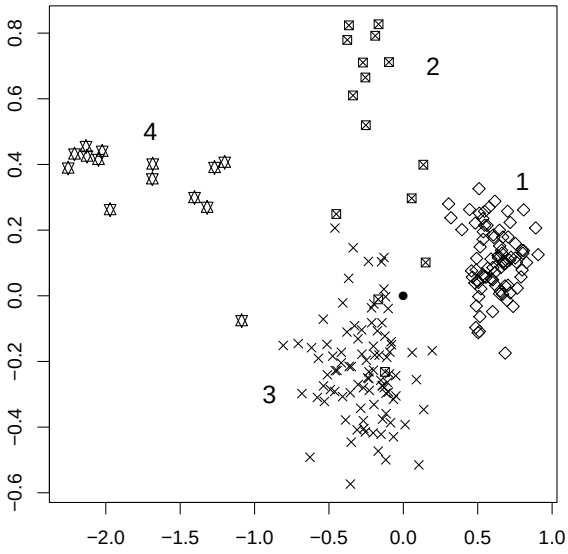


Figura 3.5: Níveis de expressão de 205 genes de *Saccharomyces cerevisiae*, projetados nas duas primeiras componentes principais, e seu agrupamento nas classes funcionais (Gene Ontology).

número correto de classes e a taxa de acerto na classificação quando o método acertava o número de classes.

Para determinar o nível de rejeição da evidência contra a hipótese $H : w_m = 0$, para este conjunto de dados não é possível usar a aproximação assintótica das evidências como fizemos na seção anterior, pois em alguns casos a dimensão do espaço paramétrico ultrapassa o tamanho da amostra (por exemplo, um modelo de misturas de distribuições normais de 10 variáveis e 4 componentes possui 263 parâmetros, número superior ao tamanho da amostra). Por essa razão, adotamos o critério de aceitar a hipótese caso $\overline{\mathrm{ev}}(H) > 0.5$. Esse critério é coerente com as funções de perda LP_1 apresentada na seção 2.2, com coeficientes $a = c = 1, b = 0$.

A Tabela 3.5 apresenta os resultados numéricos dessa análise. Cada linha contém o número de componentes principais usadas na amostra, o número de classes previstas pelo FBST e pelo Mclust sobre aquelas componentes, e a taxa de acerto na classificação quando o número de classes previstas é 4. A última linha da tabela contém as classes previstas sobre a amostra original, sem aplicação de PCA. O FBST acertou o número de classes em quatro situações (com 3,4,5 e 6 componentes principais), enquanto o Mclust acertou apenas em duas (com 2 e 3 componentes principais). (Note-se ainda que as quatro primeiras componentes principais respondem por 95% da variância acumulada dos dados, e seria

Tabela 3.5: Validação das classificações do FBST e do Mclust: quantidade prevista de classes e percentual de genes classificados corretamente (quando o número previsto de classe está correto), em função do número de componentes principais selecionadas.

Numero PCAs	FBST		Mclust	
	Classes prev.	% Acerto classif.	Classes prev.	% Acerto classif.
2	5	–	4	96%
3	4	99%	4	97%
4	4	97%	5	–
5	4	84%	7	–
6	4	83%	7	–
7	3	–	6	–
8	3	–	6	–
9	3	–	7	–
10	3	–	7	–
Dados orig.	3	–	3	–

razoável interromper a inclusão de novas componentes com esta quantidade.) O FBST mostrou uma estabilidade maior em suas respostas: Inicialmente, com uma quantidade muito pequena de informação (2 componentes), o FBST tende a encontrar um número maior de componentes. À medida que é agregada mais informação relevante ao modelo (3 e 4 componentes), o FBST consegue inferir corretamente o número de classes e também consegue uma alta taxa de acerto de classificação. Quando novas componentes principais agregando pouca informação são incluídas, o FBST começa a perder um pouco de sua precisão, e a partir de 7 componentes principais, o FBST passa a prever 3 classes. O Mclust apresenta uma boa taxa de acerto com duas e três componentes principais, mas a partir de 4 componentes principais o modelo escolhido oscila no número de classes previstas, entre 5, 6 e 7 classes, porém sem um comportamento monotônico. Curiosamente, quando utilizamos as 10 componentes principais (portanto toda a informação dos dados originais, porém em um sistema de coordenadas diferentes), o Mclust prevê 7 classes; mas quando aplicado sobre os dados originais, o Mclust prevê apenas 3 classes, o que indica uma instabilidade na seleção de modelo pelo Mclust conforme a configuração da amostra. Já o FBST mantém uma grande coerência, indicando consistentemente o mesmo número de classes sob as duas configurações dos dados.

Embora esses resultados nos pareçam favoráveis ao FBST, a questão da dimensionalidade dos dados tem grande importância, e torna-se necessário adotar critérios consistentes para determinação do número ideal de componentes. Esse é um tema interessante para pesquisa futura, que nos motiva a implementar o FBST também para este problema, oferecendo assim uma ferramenta integrada de redução de dimensão (*feature extraction*) e

de classificação não supervisionada.

Capítulo 4

Modelos de Hipóteses Separadas

4.1 Introdução

Um problema importante em inferência estatística consiste em decidir qual dentre m modelos alternativos, $f_k(x, \psi_k)$, melhor se ajusta a uma amostra. As formas mais fáceis e usuais são adotar métodos visuais (p.ex *QQ-plots*) ou utilizar testes do tipo *Goodness-of-fit*, como o Chi-quadrado ou o teste de Kolmogorov-Smirnov [23], que são baseados em uma medida de distância entre a distribuição empírica observada (amostra) e a distribuição teórica candidata. Note-se que, nessas duas abordagens, a análise de cada distribuição candidata é feita independentemente das demais. Ambas são alternativas simples, eficientes em muitos casos e que contam com vasta disponibilidade de ferramentas (a linguagem R possui pacotes para as duas alternativas). Todavia, em problemas em que a escolha da distribuição não é apenas uma questão de conveniência, mas sim um problema de inferir a dinâmica e estrutura do processo estocástico que gerou as observações, a análise descritiva se torna imprecisa. Ao mesmo tempo, os testes do tipo *goodness-of-fit* apresentam uma taxa de convergência baixa, ou seja, usualmente tendem a aceitar a distribuição candidata e demandam amostras grandes para dar a resposta correta (vide seção 4.5).

Uma forma mais robusta é adotar métodos para confrontar dois possíveis modelos. Se os modelos candidatos f_k possuem formas funcionais distintas e não relacionadas (no sentido de que um modelo não seja caso particular do segundo), é usual denominá-los *Modelos separados*, ou *Hipóteses separadas*. Vários métodos de discriminação – clássicos e bayesianos – têm sido desenvolvidos, nos quais o modelo $f_1(x, \psi_1)$ é testado contra um modelo alternativo $f_2(x, \psi_2)$, fornecendo uma medida de evidência favorecendo o modelo 1 em relação ao modelo 2 [5, 19, 46, 68]. Todavia, esses métodos não são capazes de fornecer uma resposta direta quando nenhum dos dois modelos candidatos se ajusta bem aos dados.

Neste capítulo, apresentamos a formulação de modelos de hipóteses separadas como um problema de testes de hipótese em modelos de misturas, onde cada componente cor-

responde à função de densidade de um modelo candidato. Determinar se os dados provêm de uma distribuição específica é testar se os pesos das demais distribuições são iguais a 0. Sob esta formulação, se nenhum dos modelos descrevem adequadamente os dados, o teste é capaz de fornecer uma resposta direta – uma evidência alta contra todos os modelos candidatos.

A origem desta abordagem remonta ao trabalho de Cox [19], que sugeriu que, na presença de dois modelos alternativos, a função de densidade de cada observação poderia ser uma combinação exponencial dos modelos candidatos,

$$f(x \mid w, \psi_1, \psi_2) \propto f_1(x \mid \psi_1)^w f_2(x \mid \psi_2)^{1-w},$$

onde $0 \leq w \leq 1$. Atkinson [7] desenvolveu esta idéia para algumas distribuições da família exponencial, escrevendo a densidade do modelo como

$$f(x \mid w_1, w_2, \psi_1, \psi_2) = \frac{f_1(x \mid \psi_1)^{w_1} f_2(x \mid \psi_2)^{w_2}}{\int f_1(y \mid \psi_1)^{w_1} f_2(y \mid \psi_2)^{w_2} dy}.$$

Quandt [74] foi o primeiro a explorar a formulação de modelos separados via misturas, onde a função de densidade é uma combinação linear dos modelos candidatos,

$$f(x \mid w, \psi_1, \psi_2) \propto w f_1(x \mid \psi_1) + (1-w) f_2(x \mid \psi_2).$$

Um problema clássico em teoria da confiabilidade diz respeito à comparação entre as funções Weibull e Gompertz de envelhecimento, e será explorada neste capítulo. Na próxima seção, discutimos a importância dessas distribuições e em que medida elas são capazes de explicar as características estruturais e o processo de construção de um sistema. Nas seções seguintes, discutiremos a formulação e a implementação do FBST para modelos de hipóteses separadas, apresentando como estudo de caso o teste Weibull *vs.* Gompertz. Boa parte dos conceitos apresentados neste capítulo são extensões daqueles apresentados no Capítulo 3, e inevitavelmente haverá alguma redundância. Todavia, optamos por reapresentá-los, a fim de ressaltar as diferenças conceituais e práticas desta formulação em relação a modelos de misturas tradicionais. Ao final deste capítulo, apresentaremos os resultados numéricos do desempenho do FBST, tanto em situações em que os dados provêm de uma das duas distribuições candidatas como em situações em que nenhuma das duas distribuições é a verdadeira.

Parte dos conceitos e resultados deste capítulo foram apresentados no Congresso Maxent'07 (*http://www.maxent2007.org*) e aceitos para publicação [53].

4.2 Relação entre o envelhecimento e o desenvolvimento ontogênico

Tanto os organismos vivos como os sistemas manufaturados devem necessariamente ser construídos e mantidos a partir de componentes básicos. Todavia, existem diferenças profundas em seu processo de desenvolvimento que resultarão em diferentes processos de envelhecimento. Nesta seção, examinamos as similaridades e diferenças estruturais entre essas duas classes de sistemas, e como essas estruturas podem explicar algumas das propriedades do ciclo de vida sistêmico. Esta seção dá especial atenção ao desenvolvimento sistêmico na fase adulta ou pós-construção, conhecido como envelhecimento. Em nossa análise, acompanhamos os conceitos apresentados por Gavrilov [32, 33].

"The first fundamental feature of biosystems is that, in contrast to technical (artificial) devices which are constructed out of previously manufactured and tested components, organisms form themselves in ontogenesis through a process of self-assembly out of de novo *forming and externally untested elements (cells). The second property of organisms is the extraordinary degree of miniaturization of their components (the microscopic dimensions of cells, as well as the molecular dimensions of information carriers like DNA and RNA), permitting the creation of a huge redundancy in the number of elements. Thus, we expect that for living organisms, in distinction to many technical (manufactured) devices, the reliability of the system is achieved not by the high initial quality of all the elements but by their huge numbers (redundancy)."* Gavrilov (2001,p.531.)

Ao longo deste capítulo, x é o instante de falha, $f(x)$ e $F(x)$ são as funções de densidade e de distribuição acumulada do instante de falha, $S(x) = 1 - F(x)$ é a função de sobrevivência e

$$h(x) = \frac{d\ S(x)}{S(x)\ dx} = \frac{d\ \log S(x)}{dx}$$

é a taxa de falha, ou força de mortalidade, ver Barlow e Proschan [10].

Em sistemas ou componentes mais simples, a taxa de falha em um instante x não depende de quanto tempo o sistema esteve em funcionamento [10]. Por essa razão, esses componentes são chamados sistemas sem memória/sem envelhecimento. Logo, esses sistemas são caracterizados pela distribuição exponencial, com taxa de falha constante,

$$h(x) = \kappa,\ S(x) = \exp(-\kappa x) \ \text{ e } \ f(x) = \kappa \exp(-\kappa x),\ \kappa, x \geq 0.$$

Sistemas complexos são caracterizados por regimes diferentes de envelhecimento que, por sua vez, refletem as características estruturais desse sistema. Dois regimes de envelhecimento são de especial interesse para nós:

- O regime Weibull, ou lei de potência, com $h(x) = \kappa x^{\alpha}$, $\kappa, \alpha > 0$, é característico

de sistemas complexos *top-down*, montados por agentes externos (sistemas alopoiéticos).

- O regime Gompertz-Makeham, $h(x) = A + R\exp(\alpha x)$, $A, R, \alpha > 0$, é característico de sistemas complexos *bottom-up*, auto-construídos (sistemas autopoiéticos). Em modelos biológicos, o parâmetro Makeham, A, indica uma força de mortalidade externa, enquanto o regime Gompertz puro, com $A = 0$, modela a função de falha interna ou sistêmica.

Descreveremos a seguir alguns modelos estruturais que explicam esses dois regimes e os testaremos em alguns sistemas de engenharia e biológicos.

As duas estruturas básicas em teoria da confiabilidade são composições em paralelo e em série. Sistemas complexos são composições recursivas de blocos em série e em paralelo. Um bloco em paralelo falha se todos os seus componentes falharem, enquanto um bloco em série falha se um de seus componentes falhar. Dito de outra forma, um bloco em paralelo falha quando o último de seus componentes falhar, e um bloco em série falha quando o primeiro de seus componentes falhar. Portanto, as regras composicionais série-paralelo são:

- A função de distribuição acumulada de um sistema em paralelo é igual ao produto das funções acumuladas de seus componentes.
- A função de falha de um sistema em série é igual à soma das funções de falha de seus componentes.

Vamos agora considerar o "mais simples sistema complexo", modelando um organismo ou máquina com múltiplas, m, funções, onde cada função é realizada por um bloco de componentes simples redundantes. Ou seja, tal sistema é montado como uma série de m blocos, $b_j, j = 1 \ldots m$, e o bloco b_j é montado com n_j componentes simples em paralelo.

Projetos *top-down* tipicamente usam um pequeno número de unidades redundantes, a fim de otimizar custos de produção, bem como respeitar outras restrições de projeto tais como limitação de espaço ou de peso. Logo, os componentes precisam seguir padrões rigorosos, alcançados por diversas formas de controle de qualidade no processo de fabricação. Em tais sistemas, todos os componentes estão inicialmente funcionando já que, caso contrário, teriam sido rejeitados pelos controles de qualidade. Um diagrama de blocos desse tipo de sistema é apresentado na Figura 4.1A. Neste exemplo, cada bloco possui o mesmo número de componentes redundantes, $n_j = i$.

Uma vez que cada componente tem uma distribuição de falha exponencial, as regras composicionais de confiabilidade resultam nas seguintes funções de falha sistêmica para cada bloco:

$$F_j = \left(1 - e^{-\kappa x}\right)^i \quad , \quad h_j(x) = \frac{i\kappa e^{-\kappa x}\left(1 - e^{-\kappa x}\right)^{i-1}}{1 - \left(1 - e^{-\kappa x}\right)^i} \ ;$$

e resultam na seguinte função de falha para o sistema completo:

$$h(x) = \sum_{j=1}^{m} h_j(x) = \frac{mi\kappa e^{-\kappa x}\left(1 - e^{-\kappa x}\right)^{i-1}}{1 - \left(1 - e^{-\kappa x}\right)^{i}} \ .$$

Usando as aproximações assintóticas de pré-infância (early-life) e velhice avançada (late-life), $1 - \exp(-\kappa x) \approx \kappa x$, para $x << 1/\kappa$, e $1 - \exp(-\kappa x) \approx 1$, para $x >> 1/\kappa$, as funções de falha de um bloco com i elementos paralelos e do sistema completo podem ser aproximadas como

$$h_i(x) \approx \begin{cases} i\kappa^i x^{i-1} & \text{se } x << 1/\kappa \\ \kappa & \text{se } x >> 1/\kappa \end{cases} \quad , \quad h(x) \approx \begin{cases} mi\kappa^i x^{i-1} & \text{se } x << 1/\kappa \\ m\kappa & \text{if } x >> 1/\kappa \end{cases} \quad .$$

Consideremos agora os sistemas biológicos, com construção *bottom-up*, que possuem uma altíssima redundância interna, porém com uma alta freqüência de componentes defeituosas desde o nascimento. Esses sistemas estão representados na Figura 4.1B – onde os componentes defeituosos são marcados por um "X". Consideremos que o número i de elementos inicialmente em funcionamento segue uma distribuição de Poisson com parâmetro $\lambda = nq$, $P(i) = \exp(-\lambda)\lambda^i/i!$. Devemos também truncar a Poisson, observando o fato de que o organismo está inicialmente vivo, implicando a exclusão do caso $i = 0$. A constante de normalização correta para esta Poisson truncada é $c^{-1} = 1 - \exp(-\lambda) - \exp(-\lambda)\sum_{i=n+1}^{\infty} \lambda^i/i!$.

Como no modelo anterior, a função de falha sistêmica é a soma da função de falha dos seus blocos, onde cada bloco começa com i elementos operantes (distribuídos segundo a Poisson). Logo, a função de falha sistêmica esperada pode ser escrita como:

$$h(x) = \sum_{j=1}^{m} h_j(x) = m \sum_{i=1}^{n} P(i) h_i(x) \ .$$

A substituição de $h_i(x)$ resulta na taxa de falha sistêmica e aproximações:

$$h(x) = cm\kappa\lambda e^{-\lambda} e^{-\kappa x} \sum_{i=1}^{n} \frac{\lambda^{i-1}\left(1 - e^{-\kappa x}\right)^{i-1}}{(i-1)!\left(1 - \left(1 - e^{-\kappa x}\right)^{i}\right)} \ ,$$

$$h(x) \approx \begin{cases} cm\kappa\lambda e^{-\lambda}\sum_{i=1}^{n} \frac{(\kappa\lambda x)^{i-1}}{(i-1)!} = R\left(e^{\alpha x} - \epsilon(x)\right) & \text{se } x << 1/\kappa \text{ e} \\ m\kappa & \text{se } x >> 1/\kappa \ ; \end{cases} \quad .$$

Na última expressão,

$$R = cm\kappa\lambda \exp(-\lambda) \ , \quad \alpha = \kappa\lambda \ , \text{ e } \epsilon(x) = \sum_{i=n+1}^{\infty} (\kappa\lambda x)^{i-1}/(i-1)!.$$

Para κ e λ fixados e x suficientemente pequeno, $\epsilon(x)$ se aproxima de zero. Logo, na pré-infância, $h(x) \approx R\exp(\alpha x)$, como ocorre no regime Gompertz puro.

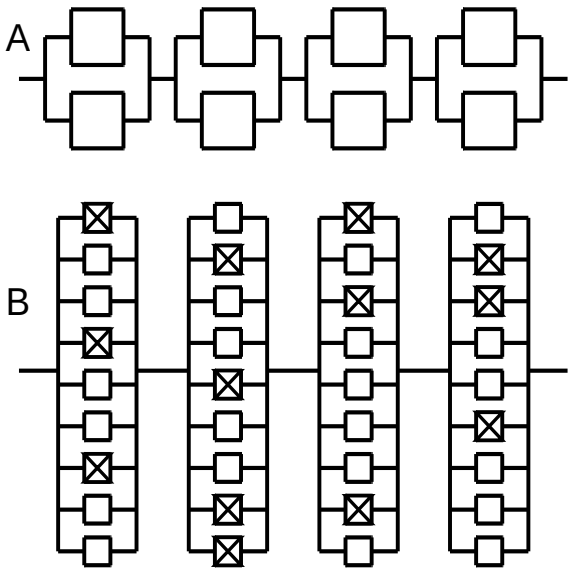


Figura 4.1: Diagramas ilustrativos das diferenças na estrutura de confiabilidade entre dispositivos manufaturados (A) e sistemas biológicos (B). Enquanto nos primeiros a confiabilidade é mantida através do controle de qualidade alta, porém uma baixa redundânciam, nos sistemas biológicos a confiabilidade é mantida por uma alta redundância de componentes, porém com várias componentes defeituosas (representadas pelas caixas com X) desde o início da vida do organismo. *Gavrilov 2001.*

4.3 Distribuições Weibull e Gompertz

Neste capítulo, adotamos por conveniência parametrizações ligeiramente diferentes das apresentadas na seção anterior.

Para o modelo Weibull, as funções de densidade, de sobrevivência e de falha, dados os parâmetros de forma de de escala, $\beta > 0, \gamma > 0$, são:

$$\begin{aligned} f_W(x \,|\, \beta, \gamma) &= (\beta t^{\beta-1}/\gamma^\beta) \exp(-(x/\gamma)^\beta) \\ r_W(x \,|\, \beta, \gamma) &= \exp(-(x/\gamma)^\beta) \\ h_W(x \,|\, \beta, \gamma) &= f_W()/r_W() = \beta x^{\beta-1}/\gamma^\beta \end{aligned}$$

No modelo Gompertz, as funções de densidade de probabilidade, de sobrevivência e

de falha, dados os parâmetros $\alpha > 1, \lambda > 0$, são:

$$\begin{aligned} f_G(x \,|\, \alpha, \lambda) &= \lambda\alpha^x \exp(-(\alpha^x - 1)\lambda/\log\alpha) \\ r_G(x \,|\, \alpha, \lambda) &= \exp(-(\alpha^x - 1)\lambda/\log\alpha) \\ h_G(x \,|\, \alpha, \lambda) &= f_G()/h_G() = \lambda\alpha^x \end{aligned}$$

Para ilustrar a forma das densidades dessas distribuições, apresentamos nas Figuras 4.2 e 4.3(esq.) os gráficos de curvas de nível das densidades Weibull e Gompertz, obtidas com uma amostra Weibull($\hat{\beta} = 4.54, \hat{\gamma} = 75$) com tamanho $n = 50$. A distribuição Gompertz apresenta uma forte dependência não-linear entre os parâmetros α and λ. Esta associação explica a chamada *lei da compensação da mortalidade*, segundo a qual valores altos do parâmetro α são compensados por valores baixos do parâmetro λ em diferentes populações de uma espécie: $\log\lambda = \log M - B\alpha$, onde B e M são invariantes espécie-específicos universais [33]. Um efeito dessa associação é que a Gompertz, em sua formulação original, não é log-côncava. Como veremos adiante, no passo de integração do FBST utilizamos amostradores adaptativos para os parâmetros, que dependem fortemente da forma da densidade – preferencialmente distribuições log-côncavas. Para diminuir essa dependência não-linear e tornar a forma da função de densidade mais apropriada para efeitos de amostragem, adotamos a reparametrização $u = 1/\log\alpha$ e $v = \log((\log\alpha)/\lambda)$, sugerida por Meeker and Escobar [62]. O gráfico à direita da Figura 4.3 apresenta as curvas de nível da densidade Gompertz reparametrizada.

As log-verossimilhanças e respectivos gradientes da distribuição Weibull e da distribuição Gompertz reparametrizada são:

$$\begin{aligned} L_W(\beta, \gamma \,|\, X) &= n\log\beta - n\beta\log\gamma + (\beta+1)\sum_j \log x - \sum_j (x_j/\gamma)^\beta \\ dL_W/d\beta &= n/\beta - n\log\gamma + \sum_j \log x_j - \sum_j (x_j/\gamma)^\beta \log(x_j/\gamma) \\ dL_W/d\gamma &= -n\beta/\gamma + \beta/\gamma \sum_j (x_j/\gamma)^\beta \end{aligned}$$

e

$$\begin{aligned} L_G(u, v \,|\, X) &= -n\log u - nv + \sum_j x_j/u + n/exp(v) - \sum_j \exp(x_j/u - v) \\ dL_G/du &= -n/u - \sum_j x_j/u^2 + \sum_j x_j/u^2 \exp(x_j/u - v) \\ dL_G/dv &= -n - n/exp(v) + \sum_j \exp(x_j/u - v). \end{aligned}$$

Os gradientes acima são necessários para o ajuste de máxima verossimilhança, que precisa ser feito numericamente.

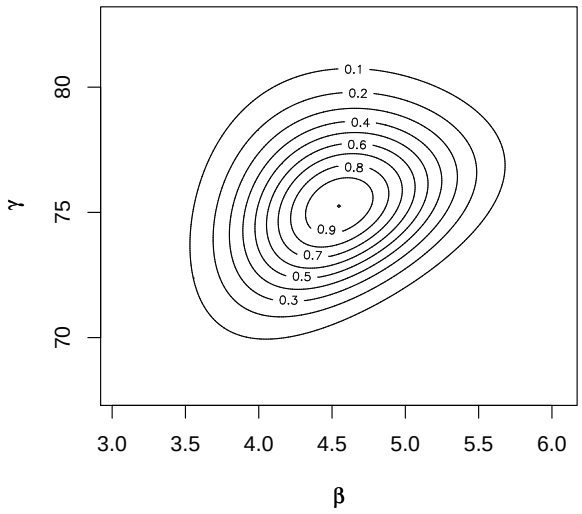


Figura 4.2: Curvas de nível das densidades Weibull em função de seus parâmetros.

4.4 Mixturas de modelos separados

Nesta seção apresentaremos a formulação de hipóteses separadas no contexto de misturas. Dada uma amostra $X = \{x_1, x_2, \ldots, x_n\}$ e m distribuições de probabilidade alternativas, $F_1(x \mid \psi_1), F_2(x \mid \psi_1) \ldots F_m(x \mid \psi_m)$, onde ψ_k é o parâmetro (possivelmente um vetor) da distribuição F_k, o problema é medir a evidência em favor de cada modelo quanto ao seu ajuste aos dados.

Em nossa proposta, consideramos que a função de densidade de probabilidade (p.d.f) de cada observação x é uma combinação linear das densidades de x sob os modelos candidatos. Denotando $\phi = (w, \phi_1 \ldots \phi_m)$,

$$f(x \mid \theta) = w_1 f_1(x \mid \psi_1) + \ldots + w_m f_m(x \mid \psi_m),\ w_k \geq 0, \sum w_k = 1.$$

A verossimilhança será

$$f(X \mid \theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f_k(x_j \mid \psi_k).$$

Neste ponto, é importante relembrar alguns conceitos-chave em modelos de misturas. Em análise de misturas para classificação não-supervisionada, assumimos que os dados provêm de uma ou mais subpopulações (classes), distribuídas sob densidades distintas. A

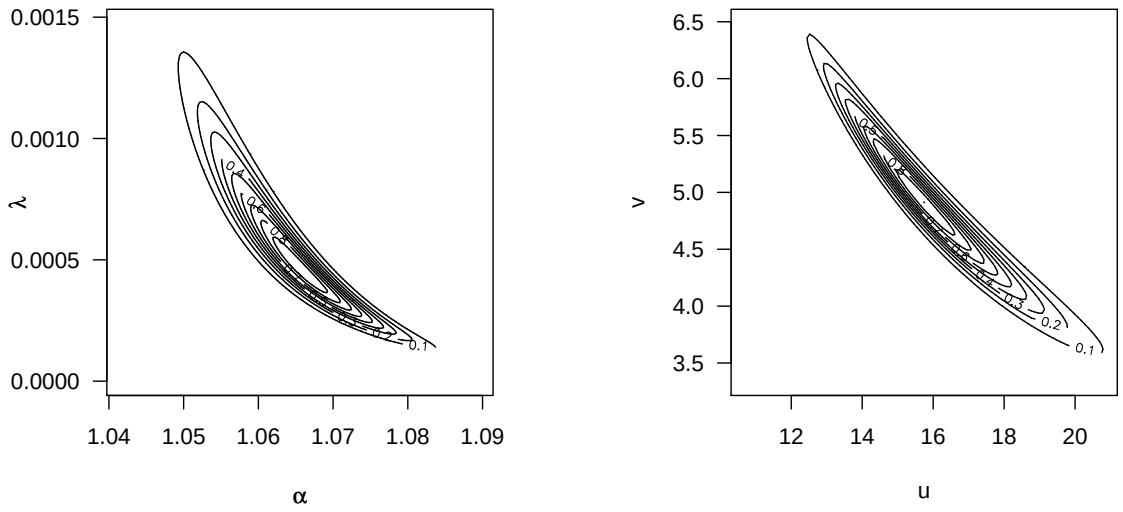


Figura 4.3: Curvas de nível das densidades da Gompertz com parâmetros α e λ (esq) e com parâmetros u e v (dir).

evidência a favor da existência de mais de uma subpopulação será maior se alguns subconjuntos dos dados forem melhor ajustados por uma componente particular da mistura, enquanto outros subconjuntos forem melhor ajustados por outras componentes. Para incorporar esta característica, o modelo de misturas precisa ser capaz de inferir as probabilidades de classificações individuais desses dados. O problema de decidir se uma única distribuição candidata se ajusta adequadamente aos dados é análogo a decidir o número de componentes em um modelo de misturas tradicional, e o comportamento do sistema será também similar: se a distribuição candidata não se ajusta bem aos dados, alguns pontos observados poderão ser melhor ajustados por outras componentes. Dessa forma, assim como foi apresentado no capítulo anterior, a mistura de modelos separados precisa considerar a "classificação" (ou componente) de cada observação.

Um exemplo j de classe $k = c(j)$ é distribuído com densidade $f_k(x_j \,|\, \psi_k)$. As classificações z_j^k são variáveis booleanas indicando se x_j pertence à classe k, ou seja, $z_j^k = 1$ sse $c(j) = k$. Denotamos por $Z = \{z_j^k\}$ a matriz de classificação. Condicionadas às variáveis latentes, teremos:

$$f(x_j \,|\, \theta) \;\; = \;\; \sum_{k=1}^{m} f_k(x_j \,|\, \theta, z_k^j) f(z_j^k \,|\, \theta) \;\; = \;\; \sum_{k=1}^{m} w_k f_k(x_j \,|\, \psi_k)$$

$$f(X \,|\, \theta) \;=\; \prod_{j=1}^{n} f(x_j \,|\, \theta) \;=\; \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f_k(x_j \,|\, \psi_k)$$

Dados os parâmetros da mistura, θ, e os dados observados, X, a matriz de probabilidades condicionais de classificação $P = f(Z \,|\, X, \theta)$, é dada por:

$$p_k^j \;=\; f(z_j^k \,|\, x_j, \theta) \;=\; \frac{f_k(z_j^k, x_j \,|\, \theta)}{f(x_j \,|\, \theta)} \;=\; \frac{w_k f_k(x_j \,|\, \psi_k)}{\sum_{k=1}^{m} w_k f(x_j \,|\, \psi_k)}$$

Denotamos por y_k o número de exemplos de classe k, ou seja, $y_k = \sum_j z_j^k$, ou $y = Z\mathbf{1}$.

A verossimilhança para os dados "completos", X, Z, é:

$$f(X, Z \,|\, \theta) \;=\; \prod_{j=1}^{n} f_{\psi_{c(j)}}(x_j \,|\, \psi_{c(j)}) f(z_j^k \,|\, \theta) \;=\; \prod_{k=1}^{m} \left[{w_k}^{y_k} \prod_{j \,|\, c(j)=k} f_k(x_j \,|\, \psi_k) \right]$$

Apresentada a formulação geral acima, discutiremos agora alguns detalhes da formulação do FBST para a mistura Weibull *vs.* Gompertz.

A distribuição a priori conjugada para uma distribuição multinomial é uma distribuição Dirichlet (da qual a Beta é um caso particular):

$$\begin{aligned} M(y \,|\, n, w) &= \left(n! / y_1! \ldots y_m!\right) {w_1}^{y_1} \ldots {w_m}^{y_m} \\ D(w \,|\, y) &= \left(\Gamma(y_1 + \ldots + y_k) / \Gamma(y_1) \ldots \Gamma(y_k)\right) \prod_{k=1}^{m} {w_k}^{y_k - 1} \end{aligned}$$

onde $w > \mathbf{0}$ e $w\mathbf{1} = 1$. A informação a priori dada por $\dot{y}$ e a observação y resultam na posteriori Dirichlet com parâmetro $\ddot{y} = \dot{y} + y$. Aqui consideramos também uma distribuição a priori (imprópria) uniforme para (β, γ, u, v), e portanto a distribuição a posteriori é

$$\begin{aligned} f(\theta \,|\, X) \;\propto\; f(X \,|\, \theta) &= \prod_{j=1}^{n} \left(w_1 f_W(x_j \,|\, \beta, \gamma) + w_2 f_G(x_j \,|\, \alpha, \lambda)\right) \\ p_j^1 &= \frac{w_1 f_W(x_j \,|\, \beta, \gamma)}{w_1 f_W(x_j \,|\, \beta, \gamma) + w_2 f_G(x_j \,|\, \alpha, \lambda)}, \quad p_j^2 = 1 - p_j^1. \end{aligned}$$

As hipóteses de interesse são:

$$\begin{aligned} H_1 &: w_1 = 1 \wedge w_2 = 0 \;\text{ e} \\ H_2 &: w_1 = 0 \wedge w_2 = 1 \end{aligned}$$

O procedimento do FBST para testar $H_k,\ k = 1, 2$ consiste em dois passos:

- Estimar o valor da máxima log-verossimilhança L_k^* sob H_k, que corresponde à máxima log-verossimilhança sob a distribuição candidata testada;
- Estimar o valor da evidência (e-valor) suportando a hipótese H_k, ou seja, a razão

$$\mathrm{ev}(H_k) = \frac{\int_{T_k} f(\theta \,|\, X) d\theta}{\int_{\Theta} f(\theta \,|\, X) d\theta} \ , \quad T_k = \{\theta \in \Theta \,|\, L(\theta) \leq L_k^*\}.$$

Note que a constante de normalização da verossimilhança é a mesma, tanto no numerador como no denominador, e portanto é cancelada; logo, essa constante pode ser ignorada no cálculo da evidência.

No passo de otimização, usamos o solver `CG_DESCENT` [36,37], baseado em métodos de gradientes conjugados. Este solver é distribuído livremente, com código-fonte em C ou Fortran (vide referências).

Para o passo de integração, adotamos o Markov Chain Monte Carlo (MCMC) por amostragem de Gibbs [35], na forma descrita a seguir. Dado o vetor de parâmetros corrente, θ^i, computamos P. Dado P, sorteanos Z onde cada coluna z_j tem distribuição multinomial $M(1, p_j)$. Dadas as variáveis latentes, Z, separamos os exemplos de classes 1 e 2. Na componente Weibull, sorteamos o valor do parâmetro $[\beta^{i+1}, \gamma^{i+1}]$ com densidade proporcional à verossimilhança parcial $\prod_{j|c(j)=1} f_W(x_j \,|\, \beta, \gamma)$. A mesma idéia é aplicada para sortear os parâmetros da Gompertz $[\alpha^{i+1}, \lambda^{i+1}]$. Dado $\ddot{y} = Z\mathbf{1} + \dot{y}$, geramos um novo vetor de pesos $[w_1^{i+1}, w_2^{i+1}]$ usando uma distribuição Dirichlet $D(w \,|\, \ddot{y}_1, \ddot{y}_2)$. No final da iteração (i), temos um novo vetor de parâmetros $\theta^{i+1} = [w_1^{i+1}, w_2^{i+1}, \beta^{i+1}, \gamma^{i+1}, \alpha^{i+1}, \lambda^{i+1}]$, podendo assim iniciar a iteração $(i+1)$.

Nós não conhecemos um método direto para gerar os parâmetros da Weibull e da Gompertz com a densidade desejada. Para isso, usamos o amostrador adaptativo `HITRO` [45,93]. O `HITRO` combina o método da Ratio-of-Uniforms com o amostrador Hit-and-Run. A transformação Ratio-of-Uniforms mapeia a região sob a p.d.f f, i.e. $G(f) = \{(x, y) : 0 < y < f(x)\}$ na região

$$A(f) = A_{r,m}(f) = \left\{ (u, v) : 0 < v < f\left(\frac{u}{v^r} + m\right)^{1/(rn+1)} \right\}$$

por meio da transformação

$$(u, v) \mapsto (x,\ y) = \left(\frac{u}{v^r} + m,\ v^{rn+1}\right) \ .$$

O vetor m precisa ser um ponto próximo à moda; em nossa implementação, m é a própria moda da distribuição. O método se baseia no teorema de que, se (u, v) é uniformemente distribuído sobre $A(f)$, então $x = u/v^r + m$ tem p.d.f $f(x)/\int f(z)dz$. O amostrador Hit-and-run é responsável por gerar pontos (u, v) uniformemente em $A(f)$.

4.5 Experimentos numéricos

Apresentamos nesta seção os resultados obtidos com os experimentos numéricos de performance do FBST para o problema de modelos separados. Os experimentos foram baseados na tábua de mortalidade masculina brasileira, do ano de 2005, disponível no endereço *http://www.ibge.gov.br/home/estatistica/populacao/tabuadevida/2005/default.shtm*. A tábua é composta pelas taxas de mortalidade (número de óbitos/1000) entre a idade i e $i+1$, de 0 a 80 anos, bem como a mortalidade acumulada acima dos 80 anos. As taxas de mortalidade abaixo dos 5 anos foram ignoradas, evitando assim o período de mortalidade infantil, em que as causas de mortalidade são preponderantemente externas [10, 59]. Nós também complementamos a tábua com as taxas anuais de mortalidade acima dos 80 anos, ajustando um spline cúbico [48] sobre as taxas dos 60 aos 80 anos e usando a função ajustada para projetar as taxas de mortalidade entre as idades i e $i+1$, para $i > 80$.

Os experimentos foram baseados em dados simulados, gerados de quatro distribuições ajustadas sobre a tábua de mortalidade. Essa foi uma forma de padronizar as amostras, fazendo não apenas com que amostras geradas a partir de distribuições diferentes pudessem estar situadas em faixas de valores compatíveis entre si, mas que também pudessem reproduzir da melhor forma possível as taxas de mortalidade humanas. As quatro distribuições ajustadas foram a Weibull, a Gompertz, a Gama e a Beta (com ajuste de escala para o intervalo de 5 a 93 anos).

A Figura 4.4 apresenta as taxas de mortalidade do IBGE, no qual a linha horizontal representa a idade de morte x e a linha vertical representa a distribuição acumulada $F(x) = P(X \leq x)$. As quatro distribuições ajustadas também são apresentadas.

Para cada uma das distribuições ajustadas aos dados, geramos conjuntos de amostras (usando como parâmetro a estimativa de máxima verossimilhança da distribuição) com tamanhos distintos: $n = 30$, 50, 75, 100, 150, 200, 300, 400 and 500. Em cada tamanho, foram geradas e testadas 500 amostras. O objetivo nesses testes é verificar a convergência das taxas de decisões corretas com o FBST, tanto em situações em que uma das distribuições é a verdadeira como em situações em que os dados não provêm de nenhuma das distribuições candidatas.

Para a aceitação/rejeição das hipóteses, adotamos o nível crítico τ de acordo com o critério assintótico apresentado na seção 2.2, com um nível de significância de 5%: considerando que o modelo de misturas completo e o modelo restrito (uma componente) possuem respectivamente 5 e 2 graus de liberdade, temos $\tau = \overline{W}^{-1}(5, 2, 0.95) = 0.83$. Portanto, rejeitamos H se $\overline{\text{ev}}(H) > 0.83$, ou equivalentemente, se $\text{ev}(H) < 0.17$.

Nos experimentos numéricos realizados, comparamos o desempenho do FBST com o do teste de Kolmogorov–Smirnov (KS). Neste teste, a qualidade do ajuste é calculada a

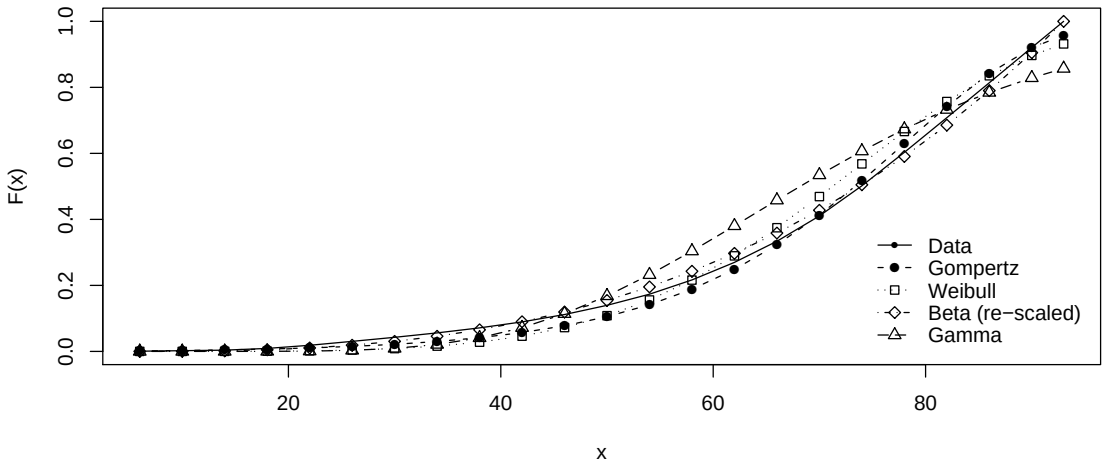


Figura 4.4: Tábua de mortalidade masculina brasileira, com taxas de mortalidade reais dos 5 aos 80 anos e estimadas dos 81 aos 93; distribuições Weibull, Gompertz, Gama e Beta ajustadas.

partir da distância de Kolmogorov,

$$D_n^* = D(F_n, F^*) = sup_x|F_n(x) - F^*(x \,|\, \theta)|,$$

onde F_n denota a distribuição da amostra e F^* denota a distribuição teórica a ser testada. Devido a dificuldades em estimar o parâmetro θ que minimiza $D(F_n, F^*)$, é usual utilizar-se o estimador de máxima verossimilhança para θ. Kolmogorov e Smirnov demonstraram que, sob a hipótese nula $F(X) = F^*(X \,|\, \theta)$, tem-se

$$lim_{n\to\infty} Pr(\sqrt{n}D_n^* \leq t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp(-2i^2t^2).$$

A distribuição no lado direito da equação permite que se calcule a significância (p-valor) de D_n^*. Em nossos testes, adotamos o nível de rejeição de 5% de significância.

As Tabelas 4.1, 4.2, 4.3 e 4.4 mostram os resultados numéricos em cada experimento simulado. Nas tabelas, a primeira coluna contém o tamanho da amostra n; as colunas 2, 3, 4 e 5 apresentam as evidências médias a favor dos modelos Weibull e Gompert, juntamente com os p-valores médios do teste KS; as colunas 6 a 11 apresentam os percentuais de casos em cada tamanho de amostra em que o FBST e o KS tomaram a decisão correta em relação à Weibull e à Gompertz. As colunas 8 e 11 representam a quantidade de casos em que

Tabela 4.1: Dados simulados com distribuição Weibull(4.540, 74.184): Evidências médias a favor dos modelos Weibull e Gompertz e percentual de decisões corretas (aceitação da Weibull e rejeição da Gompertz)

	Evidências/p-valores médios				Proporção de decisões corretas					
	FBST		KS test		FBST			KS test		
n	Ev.W	Ev.G	pv.W	pv.G	Ac.W	Rej.G	Final	Ac.W	Rej.G	Final
30	0.95	0.74	0.82	0.76	98.4%	5.0%	4.2%	100.0%	0.0%	0.0%
50	0.95	0.57	0.81	0.69	99.0%	28.0%	27.4%	100.0%	0.0%	0.0%
75	0.95	0.47	0.79	0.62	98.8%	42.8%	41.8%	99.8%	0.2%	0.0%
100	0.97	0.35	0.79	0.57	99.2%	58.6%	58.0%	100.0%	1.2%	1.2%
150	0.98	0.22	0.79	0.47	99.4%	76.2%	76.2%	100.0%	0.8%	0.8%
200	0.98	0.14	0.77	0.39	98.8%	85.0%	84.0%	100.0%	4.2%	4.2%
300	0.98	0.04	0.74	0.27	99.0%	95.6%	94.8%	100.0%	8.6%	8.6%
400	0.98	0.01	0.73	0.22	99.2%	99.4%	98.6%	100.0%	14.4%	14.4%
500	0.97	0.00	0.71	0.17	98.8%	99.4%	98.2%	99.6%	22.8%	22.8%

ambas as decisões foram tomadas corretamente (por exemplo, aceitar a Weibull e rejeitar a Gompertz, caso a amostra tenha sido gerada através da Weibull).

A tabela 4.1 mostra os resultados com dados gerados pela distribuição Weibull. A decisão correta deve ser:

- aceitar o modelo Weibull – portanto, a evidência média a favor do modelo Weibull e sua taxa de aceitação devem ser altos;
- rejeitar o modelo Gompertz – a evidência média desse modelo deve ser baixa e sua taxa de rejeição alta.

O FBST apresenta o comportamento esperado: à medida que o tamanho da amostra cresce, a taxa de rejeição do modelo falso (Gompertz) converge para 1. O acerto geral do teste atinge 95% de acerto com amostras de tamanho $n = 300$. Por outro lado o teste KS tende a aceitar o modelo falso muito mais freqüentemente, apresentando uma convergência consideravelmente mais lenta.

Na tabela 4.2 (dados gerados pela Gompertz), os resultados são similares. A taxa de rejeição do falso modelo (Weibull) converge um pouco mais lentamente do que no experimento 1, provavelmente em razão da grande flexibilidade da distribuição Weibull [25]. Na tabela 4.3 (dados gerados pela Gama), as taxas inferiores de rejeição do modelo Weibull são esperadas, pois as formas dessas distribuições são muitas vezes semelhantes [62]. Na tabela 4.4 (dados gerados pela Beta), observa-se a rápida convergência do FBST para a decisão correta (rejeição dos dois modelos).

Tabela 4.2: Dados simulados com distribuição Gompertz$(1.076, 4.255e-5)$: Evidências médias a favor dos modelos Weibull e Gompertz e percentual de decisões corretas (rejeição da Weibull e aceitação da Gompertz)

	Evidências/p-valores médios				Proporção de decisões corretas					
	FBST		KS test		FBST			KS test		
n	Ev.W	Ev.G	pv.W	pv.G	Rej.W	Ac.G	Final	Rej.W	Ac.G	Final
30	0.67	0.92	0.70	0.83	21.6%	98.2%	20.0%	0.4%	100.0%	0.4%
50	0.54	0.92	0.62	0.82	37.0%	98.4%	36.4%	1.6%	99.8%	1.6%
75	0.45	0.92	0.53	0.81	47.0%	95.2%	44.6%	2.4%	100.0%	2.4%
100	0.39	0.93	0.49	0.80	54.0%	96.8%	52.4%	4.2%	100.0%	4.2%
150	0.25	0.94	0.42	0.80	71.2%	97.0%	68.4%	5.6%	100.0%	5.6%
200	0.15	0.95	0.33	0.78	83.4%	96.6%	80.8%	8.6%	100.0%	8.6%
300	0.07	0.94	0.21	0.75	92.0%	96.6%	89.2%	22.4%	100.0%	22.4%
400	0.02	0.95	0.16	0.74	97.6%	98.0%	95.6%	33.4%	100.0%	33.4%
500	0.02	0.93	0.12	0.73	98.0%	97.6%	95.6%	45.8%	100.0%	45.8%

Tabela 4.3: Dados simulados com distribuição Gamma$(10.05, 0.15)$: Evidências médias a favor dos modelos Weibull e Gompertz e percentual de decisões corretas (rejeição da Weibull e da Gompertz)

	Evidências/p-valores médios				Proporção de decisões corretas					
	FBST		KS test		FBST			KS test		
n	Ev.W	Ev.G	pv.W	pv.G	Rej.W	Rej.G	Final	Rej.W	Rej.G	Final
30	0.94	0.40	0.74	0.53	4.6%	49.8%	4.6%	0.2%	0.8%	0.0%
50	0.89	0.19	0.66	0.36	9.2%	80.4%	9.2%	0.2%	5.2%	0.2%
75	0.90	0.06	0.63	0.25	8.8%	94.0%	8.8%	0.2%	14.4%	0.2%
100	0.85	0.01	0.54	0.14	13.2%	98.8%	13.2%	0.8%	31.0%	0.8%
150	0.82	0.00	0.46	0.06	15.4%	99.8%	15.4%	1.6%	62.0%	1.6%
200	0.73	0.00	0.40	0.02	24.2%	100.0%	24.2%	3.8%	86.0%	3.8%
300	0.56	0.00	0.26	0.00	41.4%	100.0%	41.4%	11.6%	99.2%	11.6%
400	0.37	0.00	0.18	0.00	60.0%	100.0%	60.0%	23.0%	100.0%	23.0%
500	0.29	0.00	0.16	0.00	68.2%	100.0%	68.2%	31.0%	100.0%	31.0%

Tabela 4.4: Dados simulados com distribuição Beta$(2.816, 1.017)$ (com mudança de escala): Evidências médias a favor dos modelos Weibull e Gompertz e percentual de decisões corretas (rejeição da Weibull e da Gompertz)

	Evidências/p-valores médios				Proporção de decisões corretas					
	FBST		KS test		FBST			KS test		
n	Ev.W	Ev.G	pv.W	pv.G	Rej.W	Rej.G	Final	Rej.W	Rej.G	Final
30	0.58	0.95	0.58	0.71	23.6%	2.8%	2.8%	0.6%	0.0%	0.0%
50	0.37	0.89	0.47	0.63	50.0%	8.6%	8.6%	1.4%	0.0%	0.0%
75	0.21	0.77	0.34	0.55	76.4%	20.0%	20.0%	4.6%	0.0%	0.0%
100	0.10	0.65	0.26	0.47	87.4%	31.8%	31.8%	10.6%	0.6%	0.4%
150	0.02	0.35	0.14	0.34	98.0%	62.6%	62.6%	28.4%	3.4%	3.4%
200	0.00	0.18	0.09	0.26	99.6%	80.0%	80.0%	47.8%	7.4%	6.6%
300	0.00	0.04	0.03	0.15	100.0%	95.0%	95.0%	78.2%	19.0%	18.0%
400	0.00	0.02	0.01	0.10	100.0%	97.2%	97.2%	95.0%	35.0%	34.8%
500	0.00	0.00	0.01	0.07	100.0%	99.6%	99.6%	98.2%	53.6%	53.2%

Capítulo 5

Conclusões

Neste tabalho, desenvolvemos duas propostas para resolução de problemas distintos – o problema de classificação não supervisionada e o problema de modelos separados – baseadas na aplicação do teste de significância FBST em modelos de misturas.

Os resultados obtidos com o FBST nos deixaram bastante otimistas. Nos experimentos em problemas de classificação, o FBST converge mais rapidamente para a decisão correta em função do tamanho da amostra, e também demonstrou uma estabilidade e robustez maiores do que o Mclust (algoritmo de classificação também baseado em misturas de normais multivariadas, porém com critérios de decisão do número de componentes diferente de nossa abordagem). Em modelos de hipóteses separadas, o FBST também apresentou boas taxas de convergência em comparação com o teste de Kolmogorov-Smirnov. Todos esses resultados nos motivam a continuar trabalhando no desenvolvimento do FBST para problemas correlatos aos apresentados nesta tese.

Dentre as possibilidades de desenvolvimento futuro, gostaríamos de destacar:

- Transformação do FBST em um pacote para distribuição: As principais rotinas que realizam os dois passos fundamentais para cálculo das evidências – otimização e integração – foram implementadas em Linguagem C/C++, o que torna o teste satisfatoriamente rápido se considerarmos a demanda computacional envolvida. Ao longo do meu doutorado, um grande esforço foi empreendido para tornar as implementações estáveis e robustas numericamente. Em vista do grande potencial prático que nosso teste possui – em especial, em problemas de classificação de genes e de amostras celulares baseada em expressão gênica –, a próxima etapa será transformar essas rotinas em um pacote de distribuição livre. Essa será seguramente uma forma de prestar uma contribuição efetiva à produção científica, e ao mesmo tempo um importante instrumento de difusão de nossa proposta. Para isso, uma nova etapa de amadurecimento do código e de sua documentação precisa ser iniciada.

- Outra linha interessante para desenvolvimento será a formulação e implementação do FBST para o problema de seleção e extração de características de uma amostra, com o objetivo de diminuir a dimensão do problema. No Capítulo 3 discutimos a influência de uma dimensão excessivamente grande sobre a inferência. Ali apresentamos brevemente a Análise de Componentes Principais, onde são obtidas combinações lineares das variáveis originais, de forma a se obter um conjunto muito menor de variáveis (componentes principais) com baixa perda de informação. Uma possibilidade é desenvolver um teste para determinação do número de componentes principais adequado, dentro do paradigma do FBST.

- Implementação de modelos de misturas e de hipóteses separadas para outras distribuições além das apresentadas nesta tese.

Referências Bibliográficas

[1] J. Aitchison and S. M. Shen. Logistic-normal distributions: some properties and use. *Biometrika*, 67:261–272, 1980.

[2] M. Aitkin. Posterior bayes factors. *J.R.Statist.Soc B*, 53:111–142, 1991.

[3] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[4] E. Anderson. The irises of the gaspé peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

[5] M. A. Araujo and B. B. Pereira. A comparison of bayes factors for separated models: some simulation results. *Communications in Statistics – Simulation and Computation*, 36(2):297–309, 2007.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Swight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29, 2000.

[7] A. C. Atkinson. A method for discriminating between models. *J.R.Statist.Soc B*, 32:323–354, 1970.

[8] A. C. Atkinson and D. R. Cox. Planning experiments for discriminating between models. *J.R.Statist.Soc B*, 36:321–348, 1974.

[9] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[10] R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing Probability Models*. Silver Spring, 1981.

[11] A. Ben-Dor and Z. Yakhini. Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 33–42, Lion, France, 1999.

[12] J. O. Berger and L. R. Pericchi. The intrinsic bayes factor for linear model. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smoth, editors, *Bayesian Statistics*, volume 5, pages 25–44. Oxford University Press, 1996.

[13] C. Biernacki and G. Govaert. Choosing models in model-based clustering and discriminant analysis. Technical Report 3509, INRIA, 1998.

[14] H. Bozdogan. On the information-based measure of covariance complexity and its application to the ecaluation of multiple linear models. *Communications in Statistics, Theory and Methods*, 19:221–278, 1987.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International, CA, 1984.

[16] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the em algorithm. an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55:287–314, 1996.

[17] W. G. Cochran. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23:315–345, 1952.

[18] D. R. Cox. Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, Berkeley, CA, 1961. University of California Press.

[19] D. R. Cox. Further results on tests of separated families of hypotheses. *J.R.Statist.Soc B*, 24:406–423, 1962.

[20] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, 1972.

[21] N. J. F. da Silveira, D. G. Pinheiro, R. V. Rodrigues, A. Vidotto, G. M. Polachini, M. A. Zago, W. A. Silva Jr, , and E. H. Tajara. Identification of molecular markers in head and neck squamous cell carcinomas using data derived from serial analysis of gene expression, microarrays and proteomics, 2007. Submitted.

[22] M. H. DeGroot. *Optimal Statistical Decisions*. New York, 1970.

[23] M. H. DeGroot. *Probability and Statistics*. Addison–Wesley, 1986.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J.R.Statist.Soc B*, 24:1–38, 1977.

[25] B. Dodson. *Weibull Analysis*. ASQC Quality Press, Milwaukee, 1994.

[26] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, and J. M. Trent. Inference from clustering with application to gene-expression microarrays. *Journal of Computational Biology*, 9(1):105–126, 2002.

[27] B. S. Everitt and G. Dunn. *Applied Multivariate Data Analysis*. Oxford University Press, New York, 1992.

[28] S.R. Faria Jr. Um ambiente computacional para um teste de significância bayesiano. Master's thesis, Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, 2006.

[29] L. Ferré. Selection of components in principal component analysis: A comparison of methods. *Computational Statistics & Data Analysis*, 19:669–682, 1995.

[30] R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98:39–54, 1935.

[31] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *J. Classif.*, 16:297–306, 1999.

[32] L. A. Gavrilov and N. S. Gavrilova. *The Biology of Life Span: A Quantitative Approach*. Harwood Academic Publisher, New York, 1991.

[33] L. A. Gavrilov and N. S. Gavrilova. The reliability of aging and longevity. *J. Theor. Biol.*, 213:527–545, 2001.

[34] J. E. Gentle. *Random Number Generator and Monte Carlo Methods*. Springer, New York, 1998.

[35] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Press, New York, 1996.

[36] W. W. Hager and H. Zhang. *cg$_d$escent* version 1.4 - user's guide, 2005. Available at *http://www.math.ufl.edu/ hager/papers/CG*.

[37] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficien line search. *SIAM Journal on Optimization*, 16:170–192, 2005.

[38] O. Häggström. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2002.

[39] W. Härd. *Smoothing Techniques with Implementation in S*. Springer-Verlag, New York, 1991.

[40] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. E. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, 292:511–514, 1985.

[41] T. Z. Irony, M. S. Lauretto, C. A. B. Pereira, and J. M. Stern. A weibull wearout test: full bayesian approach. In *Series in Quality, Reliability & Engineering Statistics: System and Bayesian Reliability*, pages 287–300. World Scientific, 2001. v. 5.

[42] Jr J. F. Hair, R. E. Anderson, and R. L. Tatham. *Multivariate Data Analysis*. Maxwell Macmillan International, 1987.

[43] M. E. Johnson. *Multivariate Statistical Simulation*. Wiley, New York, 1987.

[44] M. C. Jones. Generating inverse wishart matrices. *Comm. Statist. Simula. Comput*, 14:511–514, 1985.

[45] R. Karawatzki, J. Leydold, and K. Potzelberger. Automatic markov chain monte carlo procedures for sampling from multivariate distributions. Technical Report 27, Department of Statistics and Mathematics Wirtschaftsuniversitat Wien Research Report Series, December 2005. Software available at *http://statistik.wu-wien.ac.at/arvag/software.html*.

[46] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[47] S. Kim, E. R. Dougherty, J. Barrera, Y. Chen, M. L. Bittner, and J. Trent. Strong feature sets from small samples. *Journal of Computational Biology*, 9(1):127–146, 2002.

[48] P. Lancaster and K. Salkauskas. *Curve and Surface Fitting: An Introduction*. Academic Press, 1986.

[49] M. S. Lauretto. Árvores de classificação para escolha de estratégias de operação em mercados de capitais. Master's thesis, Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, 1996.

[50] M. S. Lauretto, C. A. B. Pereira, J. M. Stern, and S. Zacks. Full bayesian significance test applied to multivariate normal structure models. *Brazilian Journal of Probability and Statistics*, 17:147–168, 2003.

[51] M. S. Lauretto and J. M. Stern. Fbst for mixture model selection. *American Institute of Physics Conference Proceedings*, 803:121–128, 2005.

[52] M. S. Lauretto and J. M. Stern. Testing significance in bayesian classifiers. In K. Nakamatsu and J. M. Abe, editors, *Frontiers in Artificial Intelligence and Applications*, volume 132. IOS Press, 2005.

[53] M. S. Lauretto, J. M. Stern, S.R. Faria Jr, C. A. B. Pereira, and B. B. Pereira. The problem of separate hypotheses via mixture models. In *Submitted*, 2007.

[54] A. P. Luz, E. M. P. Ciapina, R. C. Gamba, M. S. Lauretto, E. W. C. Farias, M. C. Bicego, S. Tanigushi, R. C. Montone, and V. H. Pellizari. Potential bioremediation of hydrocarbon polluted soils in the maritime antarctic. *Antarctic Science*, 18(3):335–343, 2006.

[55] M. Madruga, L. G. Esteves, and S. Wechsler. On the bayesianity of pereira-stern tests. *Test*, 10:291–299, 2001.

[56] M. R. Madruga, C. A. B. Pereira, and J. M. Stern. Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, 117:185–198, 2003.

[57] J. P. Magalhães, J. A. S. Cabral, and D. Magalhães. The influence of genes on the aging process of mices: A statistical assessment of the genetics of aging. *Genetics*, 169:265–274, 2005.

[58] J. M. Martinez. Box-quacan and the implementation of augmented lagrangian algorithms for minimization with inequality constraints. *Computational and Applied Mathematics*, 19:31–56, 2000.

[59] G. Masuy-Stroobant. The determinants of infant mortality: how far are conceptual frameworks really modelled? Technical report, Université catholique de Louvain, Département des Sciences de la Population et du Développement, 2001. Avaliable at http://www.uclouvain.be/6913.html.

[60] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[61] M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

[62] W. Q. Meeker and L. A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics, 1998.

[63] X. L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

[64] F. Nakano. *Um novo modelo para cálculo de probabilidade de paternidade - concepção e implementação*. PhD thesis, Programa Interunidades em Bioinformática, Universidade de São Paulo, 2006.

[65] A. O'Hagan. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, London, 1994.

[66] A. O'Hagan. Fractional bayes factor for model comparison (with discussion). *J.R.Statist.Soc B*, 57:99–138, 1995.

[67] D. Ormoneit and V. Tresp. Improved gaussian mixtures density estimates using bayesian penalty terms and network averaging. *Advances in Neural Information Processing Systems*, 8:542–548, 1995.

[68] B. B. Pereira. Separate families of hypotheses. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*, volume 7, pages 4881–4886. Wiley, New York, 2005. 2nd Ed.

[69] C. A. B. Pereira, F. Nakano, J. M. Stern, and M. R. Whittle. Genuine bayesian multiallelic significance test for the hardy-weinberg equilibrium law. *Genetics and Molecular Research*, 5(4):619–631, 2006.

[70] C. A. B. Pereira and J. M. Stern. Evidence and credibility: Full bayesian significance test for precise hypotheses. *Entropy Journal*, 1:69–80, 1999.

[71] C. A. B. Pereira and J. M. Stern. Model selection: Full bayesian approach. *Environmetrics*, 12:559–568, 2001.

[72] G. C. Pflug. *Optimization of stochastic models: The interface between simulation and optimization*. Kluwer, Boston, 1996.

[73] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. 1992. 2nd Ed.

[74] R. E. Quandt. A comparison of methods for testing nonnested hypotheses. *The Review of Economics and Statistics*, 56(1):92–99, 1974.

[75] B. E. Rabinow, P. W. K. Wong, E. R. Maschgan, and S. Natelson. Screening for errors in galactose metabolism with the erythrocyte. *Clinical Chemistry*, 22(12):2010–2017, 1976.

[76] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal component analysis to summarize microarray experiments: Application to sporulation time series. *Pacific Symposium on Biocomputing*, 5:452–463, 2000.

[77] S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59:731–792, 1997.

[78] R. E. Ricklefs and A. Scheuerlein. Comparison of aging-related mortality among birds and mammals. *Experimental Gerontology*, 36:845–857, 2001.

[79] R. E. Ricklefs and A. Scheuerlein. Biological implications of the weibull and gompertz models of aging. *Journal of Gerontoloty*, 57(2):B69–B76, 2002.

[80] C.P. Robert. Mixture of distributions: Inference and estimation. In D.J.Spiegelhalter W.R.Gilks, S.Richardson, editor, *Markov Chain Monte Carlo in Practice*. CRC Press, 1996.

[81] S. Russel. Machine learning: The em algorithm. Unpublished note, 1988.

[82] J. R. Schott. Weighted chi-squared tests for partial common principal component subspaces. *Biometrika*, 90(2):411–421, 2003.

[83] G. Schwarz. Estimating a dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[84] P. J. S. Silva, R. F. Hashimoto, S. Kim, J. Barrera, L. O. Brandão, E. Suh, and E. R. Dougherty. Feature selection algorithms to find strong genes. *Pattern Recognition Letters*, 26(10):1444–1453, 2005.

[85] J. C. Spall. *Introduction to Stochastic Search and Optimization*. Wiley, Hoboken, 2003.

[86] M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, Oxford University, 1997.

[87] J. M. Stern. Fbst asymptotics. Em preparação.

[88] J. M. Stern. Significance tests, belief calculi and burden of proof in legal and scientific discourse. In *Frontiers in Artificial Intelligence and Applications*, volume 101, pages 139–147. IOS Press, Amsterdan, 2003.

[89] J. M. Stern. Cognitive constructivism, eigen-solutions and sharp statistical hypotheses. In *Foundations of Information Science*, volume 61, pages 1–23. MDPI, Basel, 2005.

[90] J. M. Stern and C. A. B. Pereira. Fbst asymptotics. Under preparation, 2005.

[91] J. M. Stern and S. Zacks. Testing the independence of poisson variates under the holgate bivariate distribution. the power of a new evidence test. *Statistical and Probability Letters*, 60:313–320, 2002.

[92] A. Thalamuhu, I. Mukhopadhyay, and X. Zheng. 2006. *Bioinformatics*, 22(19):2405–2412, 2006.

[93] I. Vaduva. Computer generation of random vectors based on transformation of uniformly distributed vectors. In *Probability Theory, Proc. 7th Conf.*, pages 589–598, 1984.

[94] A. B. Yates. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, Suppl. 1:217–235, 1934.

[95] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[96] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(R34), 2003. Article and supplementary data available at *http://genomebiology.com/2003/4/5/R34*.

Apêndice A

Artigos Resultantes da Tese

- M. S. Lauretto and J. M. Stern (2005). FBST for mixture model selection. *AIP Conference Proceedings* 803, 121–128.

- M. S. Lauretto and J. M. Stern (2005). Testing significance in Bayesian classifiers. *Frontiers in Artificial Intelligence and Applications* 132, IOS Press.

- M. S. Lauretto, S. R. de Faria Jr, C. A. B. Pereira, B. B. Pereira, J. M. Stern (2007) The problem of separate hypotheses via mixture models. *AIP Conference Proceedings* 954, 268–275.

FBST for Mixture Model Selection

Marcelo S. Lauretto and Julio M. Stern†

BIOINFO and Computer Science Dept., São Paulo University
lauretto@ime.usp.br, jstern@ime.usp.br

Abstract. The Fully Bayesian Significance Test (FBST) is a coherent Bayesian significance test for sharp hypotheses. This paper proposes the FBST as a model selection tool for general mixture models, and compares its performance with Mclust, a model-based clustering software. The FBST robust performance strongly encourages further developments and investigations.

THE FBST EVIDENCE VALUE

The Fully Bayesian Significance Test (FBST) is presented by Pereira and Stern (1999), as a coherent Bayesian significance test. The FBST is intuitive and has a geometric characterization. In this article the parameter space, Θ, is a subset of R^n, and the hypothesis is defined as a further restricted subset defined by vector valued inequality and equality constraints: $H : \theta \in \Theta_H$ where $\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}$. For simplicity, we often use H for Θ_H. We are interested in precise hypotheses, with $\dim(H) < \dim(\Theta)$. $f(\theta)$ is the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps: The optimization step consists of finding f^* and $\widehat{f}$, the constrained (over H) and unconstrained maxima of the posterior. The integration step consists of integrating the posterior density over the Tangential Set, $\overline{T}$ where the posterior is higher than anywhere in H, i.e., $\overline{T} = \{\theta \in \Theta : f(\theta) > f^*\}$, $f^* = \max_H f(\theta) = f(\theta^*)$, $\widehat{f} = \max_\Theta f(\theta) = f(\widehat{\theta})$,

$$\overline{\mathrm{Ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f(\theta) d\theta \quad .$$

$\overline{\mathrm{Ev}}(H)$ is the evidence against H, and $\mathrm{Ev}(H) = 1 - \overline{\mathrm{Ev}}(H)$ is the evidence supporting (or in favour of) H. Intuitively, if $\overline{\mathrm{Ev}}(H)$ is "large", $\overline{T}$ is "heavy", and the hypothesis set is in a region of "low" posterior density, meaning a "strong" evidence against H.

Let us consider the cumulative distribution of the evidence value against the hypothesis, $\overline{V}(c) = \Pr(\overline{\mathrm{Ev}} \leq c)$, given θ^0, the true value of the parameter. Under appropriate regularity conditions, for increasing sample size, $n \to \infty$, we can say the following:

- If H is false, $\theta^0 \notin H$, then $\overline{\mathrm{Ev}}$ converges (in probability) to one, that is, $\overline{V}(c) \to \delta(1)$.
- If H is true, $\theta^0 \in H$, then $\overline{V}(c)$, the confidence level, is approximated by the function $\overline{W}(t,h,c) = \mathrm{Chi2}\left(t-h, \mathrm{Chi2}^{-1}(t,c)\right)$, where $t = \dim(\Theta)$, $h = \dim(H)$ and $\mathrm{Chi2}(k,x)$ is the cumulative chi-square distribution with k degrees of freedom.

Several FBST applications and examples, efficient computational implementation, interpretations, and comparisons with other techniques for testing sharp hypotheses, can be found in the authors' papers in the reference list.

DIRICHLET-NORMAL-WISHART MIXTURE MODELS

In a d-dimensional multivariate finite mixture model with m components (or classes), and sample size n, any given sample x^j is of class k with probability w_k; the weights, w_k, give the probability that a new observation is of class k. A sample j of class $k = c(j)$ is distributed with density $f(x^j \mid \psi_k)$.

This paragraph defines some general matrix notation. Let $r{:}s{:}t$ indicate either the vector $[r, r+s, r+2s, \ldots t]$ or the corresponding index range from r to t with step s; $r{:}t$ is a short hand for $r{:}1{:}t$. A matrix array has a superscript index, like $S^1 \ldots S^m$. So $S^k_{h,i}$ is the h-row, i-column element of matrix S^k. We may write a rectangular matrix, X, with the row (or shorter range) index subscript, and the column (or longer range) index superscript. So x_i, x^j, and x^j_i are row i, column j, and element (i, j) of matrix X. $\mathbf{0}$ and $\mathbf{1}$ are matrices of zeros and ones which dimensions are given by the context. In this paper, let h, i be indices in the range $1{:}d$, k in $1{:}m$, and j in $1{:}n$.

The classifications z^j_k are boolean variables indicating whether or not x^j is of class k, i.e. $z^j_k = 1$ iff $c(j) = k$. Z is not observed, being therefore named latent variable or missing data, see Robert (1996). Conditioning on the missing data, we get:

$$\begin{aligned} f(x^j \mid \theta) &= \sum_{k=1}^m f(x^j \mid \theta, z^j_k) f(z^j_k \mid \theta) = \sum_{k=1}^m w_k f(x^j \mid \psi_k) \\ f(X \mid \theta) &= \prod_{j=1}^n f(x^j \mid \theta) = \prod_{j=1}^n \sum_{k=1}^m w_k f(x^j \mid \psi_k) \end{aligned}$$

Given the mixture parameters, θ, and the observed data, X, the conditional classification probabilities, $P = f(Z \mid X, \theta)$, are:

$$p^j_k = f(z^j_k \mid x^j, \theta) = \frac{f(z^j_k, x^j \mid \theta)}{f(x^j \mid \theta)} = \frac{w_k f(x^j \mid \psi_k)}{\sum_{k=1}^m w_k f(x^j \mid \psi_k)}$$

We use y_k for the number of samples of class k, i.e. $y_k = \sum_j z^j_k$, or $y = Z\mathbf{1}$. The likelihood for the "completed" data, X, Z, is:

$$f(X, Z \mid \theta) = \prod_{j=1}^n f(x^j \mid \psi_{c(j)}) f(z^j_k \mid \theta) = \prod_{k=1}^m \left(w_k^{y_k} \prod_{j \mid c(j)=k} f(x^j \mid \psi_k) \right)$$

We will see in the following sections that considering the missing data Z, and the conditional classification probabilities P, is the key for successfully solving the numerical integration and optimization steps of the FBST. In this article we will focus on Gaussian finite mixture models, where $f(x^j \mid \psi_k) = N(x^j \mid b^k, R^k)$, a Normal density with mean b^k and variance matrix V^k, or precision $R^k = (V^k)^{-1}$. Next we specialize the theory of general mixture models to the Dirichlet-Normal-Wishart case.

Consider the random matrix X^j_i, i in $1{:}d$, j in $1{:}n$, $n > d$, where each column contains a sample element from a d-multivariate Normal distribution with parameters b (mean) and V (covariance), or $R = V^{-1}$ (precision). Let u and S denote the statistics:

$$u = (1/n) \sum_{j=1}^n x^j = (1/n) X\mathbf{1} \ , \ S = \sum_{j=1}^n (x^j - b) \otimes (x^j - b)' = (X - b)(X - b)'$$

The random vector u has Normal distribution with mean b and precision nR. The random matrix S has Wishart distribution with n degrees of freedom and precision matrix

R. The Normal, Wishart and Normal-Wishart pdfs have expressions:

$$\begin{aligned} N(u\,|\,n,b,R) &= (\tfrac{n}{2\pi})^{d/2}|R|^{1/2}\exp\big(-(n/2)(u-b)'R(u-b)\big) \\ W(S\,|\,e,R) &= c^{-1}\,|S|^{(e-d-1)/2}\exp\big(-(1/2)\mathrm{tr}(SR)\big) \end{aligned}$$

with normalization constant $c = |R|^{-e/2}\,2^{ed/2}\,\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma((e-i+1)/2)$.

Now consider the matrix X as above, with unknown mean b and unknown precision matrix R, and the statistic

$$S = \sum_{j=1}^{n}(x^j-u)\otimes(x^j-u)' = (X-u)(X-u)'$$

The conjugate family of priors for multivariate Normal distributions is the Normal-Wishart. Take as prior distribution for the precision matrix R the Wishart distribution with $\dot{e} > d-1$ degrees of freedom and precision matrix $\dot{S}$ and, given R, take as prior for b a multivariate Normal with mean $\dot{u}$ and precision $\dot{n}R$, i.e. let us take the Normal-Wishart prior $NW(b,R\,|\,\dot{n},\dot{e},\dot{u},\dot{S})$. Then, the posterior distribution for R is a Wishart distribution with $\ddot{e}$ degrees of freedom and precision $\ddot{S}$, and the posterior for b, given R, is k-Normal with mean $\ddot{u}$ and precision $\ddot{n}R$, i.e., we have the Normal-Wishart posterior:

$$\begin{aligned} NW(b,R\,|\,\ddot{n},\ddot{e},\ddot{u},\ddot{S}) &= W(R\,|\,\ddot{e},\ddot{S})\;N(b\,|\,\ddot{n},\ddot{u},R) \\ \ddot{n} &= \dot{n}+n\;,\;\ddot{e}=\dot{e}+n\;,\;\ddot{u}=(nu+\dot{n}\dot{u})/\ddot{n} \\ \ddot{S} &= S+\dot{S}+(n\dot{n}/\ddot{n})(u-\dot{u})\otimes(u-\dot{u})' \end{aligned}$$

All covariance and precision matrices are supposed to be positive definite, and proper priors have $\dot{e}\geq d$, and $\dot{n}\geq 1$. Non-informative Normal-Wishart improper priors are given by $\dot{n}=0$, $\dot{u}=0$, $\dot{e}=0$, $\dot{S}=0$, i.e. we take a Wishart with 0 degrees of freedom as prior for R, and a constant prior for b, see DeGroot (1970). Then, the posterior for R is a Wishart with n degrees of freedom and precision S, and the posterior for b, given R, is d-Normal with mean u and precision nR.

The conjugate prior for a multinomial distribution is a Dirichlet distribution:

$$\begin{aligned} M(y\,|\,n,w) &= \big(n!/y_1!\ldots y_m!\big)\;{w_1}^{y_1}\ldots{w_m}^{y_m} \\ D(w\,|\,y) &= \big(\Gamma(y_1+\ldots+y_k)/\Gamma(y_1)\ldots\Gamma(y_k)\big)\prod_{k=1}^{m}{w_k}^{y_k-1} \end{aligned}$$

with $w>\mathbf{0}$ and $w\mathbf{1}=1$. Prior information given by $\dot{y}$, and observation y, result in the posterior parameter $\ddot{y}=\dot{y}+y$. A non-informative prior is given by $\dot{y}=\mathbf{1}$.

Finally, we can write the posterior and completed posterior for the model as:

$$f(\theta\,|\,X,\dot{\theta}) = f(X\,|\,\theta)\,f(\theta\,|\,\dot{\theta})$$

$$f(X\,|\,\theta) = \prod_{j=1}^{n}\sum_{k=1}^{m}p_k^j w_k N(x^j\,|\,b^k,R^k)$$

$$f(\theta\,|\,\dot{\theta}) = D(w\,|\,\dot{y})\prod_{k=1}^{m}NW(b^k,R^k\,|\,\dot{n}_k,\dot{e}_k,\dot{u}^k,\dot{S}^k)$$

$$p_k^j = w_k N(x^j\,|\,b^k,R^k)\;/\;\sum_{k=1}^{m}w_k N(x^j\,|\,b^k,R^k)$$

$$f(\theta \mid X,Z,\dot{\theta}) = f(\theta \mid X,Z)\, f(\theta \mid \dot{\theta}) = D(w \mid \ddot{y}) \prod_{k=1}^{m} NW(b^k, R^k \mid \ddot{n}_k, \ddot{e}_k, \ddot{u}^k, \ddot{S}^k)$$

$$y = Z\mathbf{1} \;\; , \;\; \ddot{y} = \dot{y} + y \;\; , \;\; \ddot{n} = \dot{n} + y \;\; , \;\; \ddot{e} = \dot{e} + y$$

$$u^k = (1/y_k) \sum_{j=1}^{n} z_k^j x^j \;\; , \;\; S^k = \sum_{j=1}^{n} z_k^j (x^j - u^k) \otimes (x^j - u^k)'$$

$$\ddot{u}^k = (1/\ddot{y}_k)(\dot{n}_k \dot{u}^k + y_k u^k) \;\; , \;\; \ddot{S}^k = S^k + \dot{S}^k + (\dot{n}_k y_k / \ddot{n}_k)(u^k - \dot{u}^k) \otimes (u^k - \dot{u}^k)'$$

GIBBS SAMPLING, INTEGRATION AND OPTIMIZATION

In order to integrate a function over the posterior measure, we use an ergodic Markov Chain. The form of the Chain below is known as Gibbs sampling, and its use for numerical integration is known as Markov Chain Monte Carlo, or MCMC.

Given θ, we can compute P. Given P, $f(z^j \mid p^j)$ is a simple multinomial distribution. Given the latent variables, Z, we have simple conditional posterior density expressions for the mixture parameters:

$$f(w \mid Z,\dot{y}) = D(w \mid \ddot{y}) \;\; , \;\; f(R^k \mid X,Z,\dot{e}_k,\dot{S}^k) = W(R \mid \ddot{e}_k,\ddot{S}^k)$$

$$f(b^k \mid X,Z,R^k,\dot{n}_k,\dot{u}^k) = N(b \mid \ddot{n}_k,\ddot{u}^k,R^k)$$

Gibbs sampling is the MCMC generated by cyclically updating variables Z, θ, and P, by drawing θ and Z from the above distributions, Häggström (2002), Johnson (1987).

Given a mixture model, we obtain an equivalent model renumbering the components $1:m$ by a permutation $\sigma([1:m])$. This symmetry must be broken in order to have an identificable model, Stephens (1997). Let us assume there is an order criterion that can be used when numbering the components. If the components are not in the correct order, Label Switching is the operation of finding permutation $\sigma([1:m])$ and renumbering the components, so that the order criterion is satisfied.

If we want to look consistently at the classifications produced during a MCMC run, we must enforce a label switching to break all non-identificability symmetries. For example, in the Dirichlet-Normal-Mixture model, we could choose to order the components (switch labels) according to the the rank given by: 1- A given linear combination of the vector means, $c' * b^k$; 2- The variance determinant $|V^k|$. The choice of a good label switching criterion should consider not only the model structure and the data, but also the semantics and interpretation of the model.

The semantics and interpretation of the model may also dictate that some states, like certain configurations of the latent variables Z, are either meaningless or invalid, and shall not be considered as possible solutions. The MCMC can be adapted to deal with forbidden states by implementing rejection rules, that prevent the chain from entering the forbidden regions of the complete and/or incomplete state space, see Bennett (1976), Meng and Wong (1996).

The EM algorithm optimizes the log-posterior function $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$, see Dempster et al. (1977), Ormoneit and Tresp (1995), Russel (1988). The EM is derived from the conditional log-likelihood, and the Jensen inequality: If $w,y > \mathbf{0}, w'\mathbf{1} = 1$ then $\log w'y \geq w' \log y$. Let θ and $\tilde{\theta}$ be our current and next estimate of the MAP (Maximum

a Posteriori), and $p_k^j = f(z_k^j \,|\, x^j, \theta)$ the conditional classification probabilities. At each iteration, the log-posterior improvement is:

$$\delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta}) \;=\; fl(\tilde{\theta} \,|\, X, \dot{\theta}) - fl(\theta \,|\, X, \dot{\theta}) \;=\; \delta(\tilde{\theta}, \theta \,|\, X) + \delta(\tilde{\theta}, \theta \,|\, \dot{\theta})$$

$$\delta(\tilde{\theta}, \theta \,|\, \dot{\theta}) \;=\; fl(\tilde{\theta} \,|\, \dot{\theta}) - fl(\theta \,|\, \dot{\theta})$$

$$\delta(\tilde{\theta}, \theta \,|\, X) \;=\; fl(X \,|\, \tilde{\theta}) - fl(X \,|\, \theta) \;=\; \textstyle\sum_j \delta(\tilde{\theta}, \theta \,|\, x^j)$$

$$\delta(\tilde{\theta}, \theta \,|\, x^j) \;=\; fl(x^j \,|\, \tilde{\theta}) - fl(x^j \,|\, \theta) \;=\; \log \textstyle\sum_k \tilde{w}_k f(x^j \,|\, \tilde{\psi}_k) \;- fl(x^j \,|\, \theta) \;=$$

$$= \log \textstyle\sum_k \dfrac{p_k^j}{p_k^j} \dfrac{\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)}{f(x^j \,|\, \theta)} \;\geq\; \Delta(\tilde{\theta}, \theta \,|\, x^j) \;=\; \textstyle\sum_k p_k^j \log \dfrac{\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)}{p_k^j f(x^j \,|\, \theta)}$$

Hence, $\Delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta}) = \Delta(\tilde{\theta}, \theta \,|\, X) + \delta(\tilde{\theta}, \theta \,|\, \dot{\theta})$, is a lower bound to $\delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta})$. Also $\Delta(\theta, \theta \,|\, X, \dot{\theta}) = \delta(\theta, \theta \,|\, X, \dot{\theta}) = 0$. So, under mild differentiability conditions, both surfaces are tangent, assuring convergence of EM to the nearest local maximum. But maximizing $\Delta(\tilde{\theta}, \theta \,|\, X, \dot{\theta})$ over $\tilde{\theta}$ is the same as maximizing

$$Q(\tilde{\theta}, \theta) \;=\; \textstyle\sum_{k,j} p_k^j \log\left(\tilde{w}_k f(x^j \,|\, \tilde{\psi}_k)\right) + fl(\tilde{\theta} \,|\, \dot{\theta})$$

and each iteration of the EM algorithm breaks down in two steps:
E-step: Compute $P = E(Z \,|\, X, \theta)$. M-step: Optimize $Q(\tilde{\theta}, \theta)$, given P.

For the Gaussian mixture model, with a Dirichlet-Normal-Wishart prior,

$$\begin{aligned} Q(\tilde{\theta}, \theta) \;&=\; \textstyle\sum_{k=1}^m \sum_{j=1}^n p_k^j \left(\log \tilde{w}_k + \log N(x^j \,|\, \tilde{b}^k, \tilde{R}^k)\right) + fl(\tilde{\theta} \,|\, \dot{\theta}) \\ fl(\tilde{\theta} \,|\, \dot{\theta}) \;&=\; \log D(\tilde{w} \,|\, \dot{y}) + \textstyle\sum_{k=1}^m \log NW(\tilde{b}^k, \tilde{R}^k \,|\, \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k) \end{aligned}$$

Lagrange optimality conditions give a simple analytical solutions for the M-step:

$$y = P\mathbf{1} \;\;,\;\; \tilde{w}_k = \left(y_k + \dot{y}_k - 1\right) / \left(n - m + \textstyle\sum_{k=1}^m \dot{y}_k\right)$$

$$u^k = \tfrac{1}{y_k} \textstyle\sum_{j=1}^n p_k^j x^j \;\;,\;\; S^k = \textstyle\sum_{j=1}^n p_k^j (x^j - \tilde{b}^k) \otimes (x^j - \tilde{b}^k)'$$

$$\tilde{b}^k = \dfrac{\dot{n}_k \dot{u}^k + y_k u^k}{\dot{n}_k + y_k} \;\;,\;\; \tilde{V}^k \;=\; \dfrac{S^k + \dot{n}_k (\tilde{b}^k - \dot{u}^k) \otimes (\tilde{b}^k - \dot{u}^k)' + \dot{S}^k}{y_k + \dot{e}_k - d}$$

In more general (non-Gaussian) mixture models, if an analytical solution for the M-step is not available, a robust local optimization algorithm can be used, see for example Birgin et al. (2004). The EM is only a local optimizer, but the MCMC provides plenty of good starting points, so we have the basic elements for a global optimizer. To avoid using many starting points going to a same local maximum, we can filter the (ranked by the posteriori) top portion of the MCMC output using a clustering algorithm, and select a starting point from each cluster. For better efficiency, or more complex problems, the Stochastic EM or SEM algorithm can be used to provide starting points near each important local maximum, see Celeux et al. (1996), Pflug (1996) and Spall (2003).

MODEL SELECTION AND COMPARATIVE RESULTS

The problem under study is to determine the number of components (or classes) in a population, given a sample X drawn from that population. Each component k is assumed to follow a multivariate Normal distribution, whose mean vector b^k and variance matrix V^k must also be estimated.

In the FBST formulation of the problem, the base model has m components, and the hypothesis to be tested is the constraint of having $m-1$ components, i.e., components m and $m-1$ are identical. The FBST selects the m component model, rejecting H, if the evidence against the hypothesis is above a critical level, $\overline{\mathrm{Ev}}(H) > c$, and selects the $m-1$ component model, accepting H, otherwise. In order to determine the number of components, we apply the FBST in the base model with $m = 2, 3, \ldots$ components, and stop the process at the lowest m such that the hypotheses is accepted, m_f. The elected model has $m_f - 1$ components.

Several methods can be used to choose the critical level c. Empirical power analysis, see Stern and Zacks (2002), Lauretto et al. (2003), and sensitivity analysis, Stern (2004), require calibration procedures. Loss functions, Madruga et al. (2001), require decision theoretical interpretations. Application of these methods will be discussed in forthcoming papers. Following an anonymous referee suggestion, we proceed with a traditional power analysis. This is a form of the Rule of Parsimony, or Occam's Razor: Accept H, the smaller model, unless there is strong evidence not to do so.

We use approximate (asymptotic) critical levels corresponding to the standard Fisher confidence level of $1-\alpha$ for $\alpha = 0.01$. For example (see section 1), at the $m = 3$ base model, $t = 17$ and $h = 11$, giving $c = 0.53$.

When implementing the FBST one has to be careful with trapping states on the MCMC. These typically are states where one component has a small number of sample points, that become (nearly) collinear, resulting in a singular posterior. A standard way to avoid this inconvenience is to use flat or minimally informative priors, instead of non-informative priors, see Robert (1996). We used as flat prior parameters: $\dot{y} = \mathbf{1}$, $\dot{n} = 1$, $\dot{u} = u$, $\dot{e} = 3$, $\dot{S} = (1/n)S$. Robert (1996) uses, with similar effects, $\dot{e} = 6$, $\dot{S} = (1.5/n)S$.

In this work we compare the FBST performance with Mclust, a software for model-based cluster analysis, see Banfield and Raftery (1993) and Fraley and Raftery (1999). Mclust is available at the authors' internet site as an easy to use and ready to run software package, that has been extensively and successfully used in many applications. Also, Mclust has no extra parameters that need to be adjusted or calibrated to the specific application. These characteristics motivated our choice of Mclust for a first comparison with the FBST. Forthcoming articles will include other well published methods, based on Dirichlet processes, jump-diffusion and birth-death MCMC.

In Mclust, the variance structure and the number of components are selected via Bayesian Information Criterion (BIC), see Schwarz (1978): $BIC = 2\Lambda - \kappa \log(n)$, where Λ is the maximum model log-likelihood, κ its number of parameters, and n the sample size. BIC is a regularization criteria, weighting the model fit against the number of parameters. A larger BIC score indicates stronger evidence for the corresponding model.

Our numerical experiments are based on the *Old Faithful* dataset, see Stephens (1997), which consists of 272 eruptions observations of the Old Faithful geyser in the Yellow-

stone National Park. Each observation has the eruption duration and waiting time before the next eruption. The problem is to decide how many classes of eruptions there exist. Old Faithful is a standard dataset for experiments in the area, allowing our results to be easily reproduced, but our general conclusions have been confirmed in several randomly generated datasets.

Two numerical experiments on simulated data were performed, using parameters θ^* and $\widehat{\theta}$, the maximum likelihood estimators for 2 and 3 component models in the original dataset. In the first experiment, our interest was to analyze the overestimate and underestimate rates on the number of components, for FBST and Mclust. We used Mclust library to generate a random collection of 500 datasets with 272 points each using parameter θ^* and a second collection of 500 datasets with 272 points each using parameter $\widehat{\theta}$. Table 1 shows the number of datasets according to the estimated number of components by FBST and Mclust. Each column corresponds to one of the collections, at θ^* and $\widehat{\theta}$, and each row represents the estimated number of components.

In the second numerical experiment we examine the FBST and Mclust choice between the 2 and 3 component models, as the sample size n increases. For each $n \in \{200, 300, 400, 500, 600\}$, we simulated two collections of 500 datasets with n points each, one using the parameter θ^*, and the other using parameter $\widehat{\theta}$. Table 2 shows the number of missclassifications for FBST and Mclust, at each of the 10 collections.

These (preliminary) results corroborate the authors' previous findings, indicating that the FBST is a robust Bayesian sharp hypothesis test, and a promissing tool for model selection, deserving further development and investigation.

Finally, let us point out a related topic for further research: The problem of discriminating between models consists of determining which of m alternative models, $f_k(x, \psi_k)$, more adequately fits or describes a given dataset. In general the parameters ψ_k have distinct dimensions, and the models f_k have distinct (unrelated) functional forms. In this case it is usual to call them "separate" models (or hypotheses). Atkinson (1970), although in a very different theoretical framework, was the first to analyse this problem using a mixture formulation, $f(x \mid \theta) = \sum_{k=1}^{m} w_k f_k(x, \psi_k)$. The general theory for mixture models presented in this article can be adapted to analyse the problem of discriminating between separate hypotheses. This is the subject of the authors' ongoing research with C.A.B.Pereira and B.B.Pereira, to be presented in forthcoming articles.

The authors are grateful for the support of CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo.

Estimated components	FBST θ^*	FBST $\widehat{\theta}$	Mclust θ^*	Mclust $\widehat{\theta}$
1	0	0	0	0
2	498	187	500	280
3	2	288	0	218
4	0	25	0	2
–	–	–	–	–

Table 1: Datasets (in 500), according to estimated number of components.

Dataset Size	FBST θ^*	FBST $\widehat{\theta}$	Mclust θ^*	Mclust $\widehat{\theta}$
200	4	356	0	390
300	2	82	0	235
400	1	47	0	156
500	5	3	0	69
600	6	0	0	31

Table 2: Missclassifications (in 500), according to dataset size.

REFERENCES

A.C.Atkinson (1970). A Method for Discriminating Between Models. *J. Royal Stat. Soc. B*, 32, 323-354.

J.D.Banfield,A.E.Raftery (1993). Model Based Gaussian and nonGaussian Clustering. *Biometrics*,803-21.

C.H.Bennett (1976). Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* 22, 245-268.

E.G.Birgin, R.Castillo, J.M.Martinez (2004). Numerical comparison of Augmented Lagrangian algorithms for nonconvex problems. to appear in *Computational Optimization and Applications.*

W.Borges,J.M.Stern (2005). *On the Truth Value of Complex Hypotheses.* Tech.Rep.MAC-IME-USP-05-5.

G.Celeux, G.Govaert (1995). Gaussian Parsimonious Clustering Models. *Pattern Recog.* 28, 781-793.

G.Celeux, D.Chauveau, J.Diebolt (1996). On Stochastic Versions of the EM Algorithm. An Experimental Study in the mixture Case. *Journal of Statistical Computation and Simulation*, 55, 287–314.

M.H.DeGroot (1970). *Optimal Statistical Decisions*. NY: McGraw-Hill.

A.P.Dempster, N.M.Laird, D.B.Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society B.* 39, 1-38.

C.Fraley, A.E.Raftery (1999). Mclust: Software for Model-Based Cluster Analysis. *J. Classif.*,16,297-306.

W.R.Gilks, S.Richardson, D.J.Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* NY: CRC.

O.Häggström (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge University Press.

M.E.Johnson (1987). Multivariate Statistical Simulation. NY: Wiley.

M.Lauretto, C.A.B.Pereira, J.M.Stern, S.Zacks (2003). Comparing Parameters of Two Bivariate Normal Distributions Using the Invariant FBST. *Brazilian Journal of Probability and Statistics*, 17, 147-168.

M.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*,10,291–299.

M.R.Madruga, C.A.B.Pereira, J.M.Stern (2003). Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference*, 117,185–198.

X.L.Meng, W.H.Wong (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, 6, 831-860.

D.Ormoneit, V.Tresp (1995). Improved Gaussian Mixtures Density Estimates Using Bayesian Penalty Terms and Network Averaging. *Advances in Neural Information Processing Systems 8*, 542–548. MIT.

C.A.B.Pereira, J.M.Stern, (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.

C.A.B.Pereira, J.M.Stern, (2001). Model Selection: Full Bayesian Approach. *Environmetrics*,12,559–68.

G.C.Pflug (1996). *Optimization of Stochastic Models.* Boston: Kluwer.

C.P.Robert (1996). Mixture of Distributions: Inference and Estimation. in Gilks et al. (1996).

S.Russel (1988). Machine Learning: The EM Algorithm. Unpublished note.

G.Schwarz (1978). Estimating the Dimension of a Model. *Ann. Stat.*, 6, 461-464.

J.C.Spall (2003). *Introduction to Stochastic Search and Optimization.* Hoboken: Wiley.

M.Stephens (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Oxford University.

J.M.Stern (1992). Simulated Annealing with a Temperature Dependent Penalty Function. *ORSA Journal on Computing*, 4, 311-319.

J.M.Stern (2003). Significance Tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse. Laptec'03, *Frontiers in Artificial Intelligence and its Applications*, 101, 139–147.

J.M.Stern (2004a). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *Lecture Notes Artificial Intelligence*, 3171, 134–143.

J.M.Stern (2004b). Uninformative Reference Sensitivity in Possibilistic Sharp Hypotheses Tests. MaxEnt 2004, *American Institute of Physics Proceedings*, 735, 581–588.

J.M.Stern (2005). Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Proc. 3rd Conference on the Foundations of Information Science, FIS-2005*, 1–23.

J.M.Stern, S.Zacks (2002). Testing the Independence of Poisson Variates under the Holgate Bivariate Distribution. The Power of a New Evidence Test. *Statistical and Probability Letters*, 60, 313–320.

Testing Significance in Bayesian Classifiers

Marcelo Lauretto and Julio M. Stern
BIOINFO and Computer Science Dept., São Paulo University

Abstract. The Fully Bayesian Significance Test (FBST) is a coherent Bayesian significance test for sharp hypotheses. This paper explores the FBST as a model selection tool for general mixture models, and gives some computational experiments for Multinomial-Dirichlet-Normal-Wishart models.

1 FBST and Model Selection

The Fully Bayesian Significance Test (FBST) is presented by Pereira and Stern, [21], as a coherent Bayesian significance test. The FBST is intuitive and has a geometric characterization. In this article the parameter space, Θ, is a subset of R^n, and the hypothesis is defined as a further restricted subset defined by vector valued inequality and equality constraints: $H : \theta \in \Theta_H$ where $\Theta_H = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}$. For simplicity, we often use H for Θ_H. We are interested in precise hypotheses, with $\dim(\Theta_0) < \dim(\Theta)$. $f(\theta)$ is the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps: The optimization step consists of finding f^*, the maximum (supremum) of the posterior under the null hypothesis. The integration step consists of integrating the posterior density over the Tangential Set, T where the posterior is higher than anywhere in the hypothesis, i.e.,

$$\begin{aligned} \text{Ev}(H) &= \Pr(\theta \in T \mid x) = \int_T f(\theta) d\theta \ , \text{ where} \\ T &= \{\theta \in \Theta : f(\theta) > f^*\} \quad \text{and} \quad f^* = \sup_H f(\theta) \end{aligned}$$

$\text{Ev}(H)$ is the evidence against H, and $\overline{\text{Ev}}(H) = 1 - \text{Ev}(H)$ is the evidence supporting (or in favour of) H. Intuitively, if $\text{Ev}(H)$ is "large", T is "heavy", and the hypothesis set is in a region of "low" posterior density, meaning a "strong" evidence against H.

Several FBST applications and examples, efficient computational implementation, interpretations, and comparisons with other techniques for testing sharp hypotheses, can be found in the authors' papers in the reference list.

2 Dirichlet-Normal-Wishart Mixtures

In a d-dimensional multivariate finite mixture model with m components (or classes), and sample size n, any given sample x^j is of class k with probability w_k; the weights, w_k, give the probability that a new observation is of class k. A sample j of class $k = c(j)$ is distributed with density $f(x^j \mid \psi_k)$.

This paragraph defines some general matrix notation. Let $r : s : t$ indicate either the vector $[r, r + s, r + 2s, \ldots t]$ or the corresponding index range from r to t with step s; $r : t$ is a

short hand for $r : 1 : t$. A matrix array has a superscript index, like $S^1 \ldots S^m$. So $S^k_{h,i}$ is the h-row, i-column element of matrix S^k. We may write a rectangular matrix, X, with the row (or shorter range) index subscript, and the column (or longer range) index superscript. So x_i, x^j, and x^j_i are row i, column j, and element (i, j) of matrix X. $\mathbf{0}$ and $\mathbf{1}$ are matrices of zeros and ones which dimensions are given by the context. $V > 0$ is a positive definite matrix. In this paper, let h, i be indices in the range $1 : d$, k in $1 : m$, and j in $1 : n$.

The classifications z^j_k are boolean variables indicating whether or not x^j is of class k, i.e. $z^j_k = 1$ iff $c(j) = k$. Z is not observed, being therefore named latent variable or missing data. Conditioning on the missing data, we get:

$$\begin{aligned} f(x^j \,|\, \theta) &= \sum\nolimits_{k=1}^{m} f(x^j \,|\, \theta, z^j_k) f(z^j_k \,|\, \theta) = \sum\nolimits_{k=1}^{m} w_k f(x^j \,|\, \psi_k) \\ f(X \,|\, \theta) &= \prod\nolimits_{j=1}^{n} f(x^j \,|\, \theta) = \prod\nolimits_{j=1}^{n} \sum\nolimits_{k=1}^{m} w_k f(x^j \,|\, \psi_k) \end{aligned}$$

Given the mixture parameters, θ, and the observed data, X, the conditional classification probabilities, $P = f(Z \,|\, X, \theta)$, are:

$$p^j_k = f(z^j_k \,|\, x^j, \theta) = \frac{f(z^j_k, x^j \,|\, \theta)}{f(x^j \,|\, \theta)} = \frac{w_k f(x^j \,|\, \psi_k)}{\sum_{k=1}^{m} w_k f(x^j \,|\, \psi_k)}$$

We use y_k for the number of samples of class k, i.e. $y_k = \sum_j z^j_k$, or $y = Z\mathbf{1}$. The likelihood for the "completed" data, X, Z, is:

$$f(X, Z \,|\, \theta) = \prod\nolimits_{j=1}^{n} f(x^j \,|\, \psi_{c(j)}) f(z^j_k \,|\, \theta) = \prod\nolimits_{k=1}^{m} \big({w_k}^{y_k} \prod\nolimits_{j \,|\, c(j)=k} f(x^j \,|\, \psi_k) \big)$$

We will see in the following sections that considering the missing data Z, and the conditional classification probabilities P, is the key for successfully solving the numerical integration and optimization steps of the FBST. In this article we will focus on Gaussian finite mixture models, where $f(x^j \,|\, \psi_k) = N(x^j \,|\, b^k, R^k)$, a normal density with mean b^k and variance matrix V^k, or precision $R^k = (V^k)^{-1}$. Next we specialize the theory of general mixture models to the Dirichlet-Normal-Wishart case.

Consider the random matrix X^j_i, i in $1 : d$, j in $1 : n$, $n > d$, where each column contains a sample element from a d-multivariate normal distribution with parameters b (mean) and V (covariance), or $R = V^{-1}$ (precision). Let u and S denote the statistics:

$$u = (1/n) \sum\nolimits_{j=1}^{n} x^j = (1/n)\, X\mathbf{1} \;\;,\;\; S = \sum\nolimits_{j=1}^{n} (x^j - b) \otimes (x^j - b)' = (X - b)(X - b)'$$

The random vector u has normal distribution with mean b and precision nR. The random matrix S has Wishart distribution with n degrees of freedom and precision matrix R. The Normal, Wishart and Normal-Wishart pdfs have expressions:

$$N(u \,|\, n, b, R) = (\tfrac{n}{2\pi})^{d/2} |R|^{1/2} \exp\big(-(n/2)(u - b)' R (u - b) \big)$$

$$W(S \,|\, e, R) = c^{-1} \, |S|^{(e-d-1)/2} \exp\big(-(1/2)\mathrm{tr}(S\, R) \big)$$

with normalization constant $c = |R|^{-e/2} \, 2^{ed/2} \, \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma((e - i + 1)/2)$.

Now consider the matrix X as above, with unknown mean b and unknown precision matrix R, and the statistic

$$S = \sum\nolimits_{j=1}^{n} (x^j - u) \otimes (x^j - u)' = (X - u)(X - u)'$$

The conjugate family of priors for multivariate normal distributions is the Normal-Wishart, see [6]. Take as prior distribution for the precision matrix R the wishart distribution with $\dot{e} > d-1$ degrees of freedom and precision matrix $\dot{S}$ and, given R, take as prior for b a multivariate normal with mean $\dot{u}$ and precision $\dot{n}R$, i.e. let us take the Normal-Wishart prior $NW(b, R \mid \dot{n}, \dot{e}, \dot{u}, \dot{S})$. Then, the posterior distribution for R is a Wishart distribution with $\ddot{e}$ degrees of freedom and precision $\ddot{S}$, and the posterior for b, given R, is k-Normal with mean $\ddot{u}$ and precision $\ddot{n}R$, i.e., we have the Normal-Wishart posterior:

$$\begin{aligned} NW(b, R \mid \ddot{n}, \ddot{e}, \ddot{u}, \ddot{S}) &= W(R \mid \ddot{e}, \ddot{S})\ N(b \mid \ddot{n}, \ddot{u}, R) \\ \ddot{n} &= \dot{n} + n \ , \ \ddot{e} = \dot{e} + n \ , \ \ddot{u} = (nu + \dot{n}\dot{u})/\ddot{n} \\ \ddot{S} &= S + \dot{S} + (n\dot{n}/\ddot{n})(u - \dot{u}) \otimes (u - \dot{u})' \end{aligned}$$

All covariance and precision matrices are supposed to be positive definite, and proper priors have $\dot{e} \geq d$, and $\dot{n} \geq 1$. Non-informative Normal-Wishart improper priors are given by $\dot{n} = 0$, $\dot{u} = 0$, $\dot{e} = 0$, $\dot{S} = 0$, i.e. we take a Wishart with 0 degrees of freedom as prior for R, and a constant prior for b, see [6]. Then, the posterior for R is a Wishart with n degrees of freedom and precision S, and the posterior for b, given R, is d-Normal with mean u and precision nR.

The conjugate prior for a multinomial distribution is a Dirichlet distribution:

$$\begin{aligned} M(y \mid n, w) &= \left(n! / y_1! \dots y_m!\right)\ {w_1}^{y_1} \dots {w_m}^{y_m} \\ D(w \mid y) &= \left(\Gamma(y_1 + \dots + y_k) / \Gamma(y_1) \dots \Gamma(y_k)\right) \prod_{k=1}^{m} {w_k}^{y_k - 1} \end{aligned}$$

with $w > \mathbf{0}$ and $w\mathbf{1} = 1$. Prior information given by $\dot{y}$, and observation y, result in the posterior parameter $\ddot{y} = \dot{y} + y$. A non-informative prior is given by $\dot{y} = \mathbf{1}$.

Finally, we can write the posterior and completed posterior for the model as:

$$f(\theta \mid X, \dot{\theta}) = f(X \mid \theta) f(\theta \mid \dot{\theta})$$

$$f(X \mid \theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} p_k^j w_k N(x^j \mid b^k, R^k)$$

$$f(\theta \mid \dot{\theta}) = D(w \mid \dot{y}) \prod_{k=1}^{m} NW(b^k, R^k \mid \dot{n}_k, \dot{e}_k, \dot{u}^k, \dot{S}^k)$$

$$p_k^j = w_k N(x^j \mid b^k, R^k) \ / \ \sum_{k=1}^{m} w_k N(x^j \mid b^k, R^k)$$

$$f(\theta \mid X, Z, \dot{\theta}) = f(\theta \mid X, Z) f(\theta \mid \dot{\theta}) = D(w \mid \ddot{y}) \prod_{k=1}^{m} NW(b^k, R^k \mid \ddot{n}_k, \ddot{e}_k, \ddot{u}^k, \ddot{S}^k)$$

$$y = Z\mathbf{1} \ , \ \ddot{y} = \dot{y} + y \ , \ \ddot{n} = \dot{n} + y \ , \ \ddot{e} = \dot{e} + y$$

$$u^k = (1/y_k) \sum_{j=1}^{n} z_k^j x^j \ , \ S^k = \sum_{j=1}^{n} z_k^j (x^j - u^k) \otimes (x^j - u^k)'$$

$$\ddot{u}^k = (1/\ddot{y}_k)(\dot{n}_k \dot{u}^k + y_k u^k) \ , \ \ddot{S}^k = S^k + \dot{S}^k + (\dot{n}_k y_k / \ddot{n}_k)(u^k - \dot{u}^k) \otimes (u^k - \dot{u}^k)'$$

3 Gibbs Sampling, Integration and Optimization

In order to integrate a function over the posterior measure, we use an ergodic Markov Chain. The form of the Chain below is known as Gibbs sampling, and its use for numerical integration is known as Markov Chain Monte Carlo, or MCMC.

Given θ, we can compute P. Given P, $f(z^j \mid p^j)$ is a simple multinomial distribution. Given the latent variables, Z, we have simple conditional posterior density expressions for the mixture parameters:

$$f(w \mid Z, \dot{y}) = D(w \mid \ddot{y}) \ , \ \ f(R^k \mid X, Z, \dot{e}_k, \dot{S}^k) = W(R \mid \ddot{e}_k, \ddot{S}^k)$$

$$f(b^k \mid X, Z, R^k, \dot{n}_k, \dot{u}^k) = N(b \mid \ddot{n}_k, \ddot{u}^k, R^k)$$

Gibbs sampling is nothing but the MCMC generated by cyclically updating variables Z, θ, and P, by drawing θ and Z from the above distributions, see [10], [11]. A uniform generator is all what is needed to the multinomial variate. A Dirichlet variate w can be drawn using a gamma generator with shape and scale parameters α and β, $G(\alpha, \beta)$, see [9]. Johnson [12] describes a simple procedure to generate the Cholesky factor of a Wishart variate $W = U'U$ with n degrees of freedom, from the Cholesky factorization of the covariance $V = R^{-1} = C'C$, and a chi-square generator: a) $g_k = G(y_k, 1)$; b) $w_k = g_k \ / \ \sum_{k=1}^{m} g_k$; c) for $i < j$, $B_{i,j} = N(0, 1)$; d) $B_{i,i} = \sqrt{\chi^2(n - i + 1)}$; and e) $U = BC$. All subsequent matrix computations proceed directly from the Cholesky factors, [13].

Given a mixture model, we obtain an equivalent model renumbering the components $1 : m$ by a permutation $\sigma([1 : m])$. This symmetry must be broken in order to have an identifiable model, see [27]. Let us assume there is an order criterion that can be used when numbering the components. If the components are not in the correct order, Label Switching is the operation of finding permutation $\sigma([1 : m])$ and renumbering the components, so that the order criterion is satisfied. If we want to look consistently at the classifications produced during a MCMC run, we must enforce a label switching to break all non-identifiability symmetries. For example, in the Dirichlet-Normal-Mixture model, we could choose to order the components (switch labels) according to the the rank given by: 1) A given linear combination of the vector means, $c' * b^k$; 2) The variance determinant $|V^k|$. The choice of a good label switching criterion should consider not only the model structure and the data, but also the semantics and interpretation of the model.

The semantics and interpretation of the model may also dictate that some states, like certain configurations of the latent variables Z, are either meaningless or invalid, and shall not be considered as possible solutions. The MCMC can be adapted to deal with forbidden states by implementing rejection rules, that prevent the chain from entering the forbidden regions of the complete and/or incomplete state space, see [3], [19].

The EM algorithm optimizes the log-posterior function $fl(X \mid \theta) + fl(\theta \mid \dot{\theta})$, see [7], [20], [25]. The EM is derived from the conditional log-likelihood, and the Jensen inequality: If $w, y > \mathbf{0}, w'\mathbf{1} = 1$ then $\log w'y \geq w' \log y$. Let θ and $\tilde{\theta}$ be our current and next estimate of the MAP (Maximum a Posteriori), and $p_k^j = f(z_k^j \mid x^j, \theta)$ the conditional classification probabilities. At each iteration, the log-posterior improvement is:

$$\delta(\tilde{\theta},\theta\,|\,X,\dot{\theta}) \;=\; fl(\tilde{\theta}\,|\,X,\dot{\theta}) - fl(\theta\,|\,X,\dot{\theta}) \;=\; \delta(\tilde{\theta},\theta\,|\,X) + \delta(\tilde{\theta},\theta\,|\,\dot{\theta})$$

$$\delta(\tilde{\theta},\theta\,|\,\dot{\theta}) \;=\; fl(\tilde{\theta}\,|\,\dot{\theta}) - fl(\theta\,|\,\dot{\theta})$$

$$\delta(\tilde{\theta},\theta\,|\,X) \;=\; fl(X\,|\,\tilde{\theta}) - fl(X\,|\,\theta) \;=\; \sum\nolimits_j \delta(\tilde{\theta},\theta\,|\,x^j)$$

$$\delta(\tilde{\theta},\theta\,|\,x^j) \;=\; fl(x^j\,|\,\tilde{\theta}) - fl(x^j\,|\,\theta) \;=\; \log\sum\nolimits_k \tilde{w}_k f(x^j\,|\,\tilde{\psi}_k) \;- fl(x^j\,|\,\theta) \;=$$

$$=\; \log\sum\nolimits_k \frac{p_k^j}{p_k^j}\frac{\tilde{w}_k f(x^j\,|\,\tilde{\psi}_k)}{f(x^j\,|\,\theta)} \;\;\geq\;\; \Delta(\tilde{\theta},\theta\,|\,x^j) \;=\; \sum\nolimits_k p_k^j \log\frac{\tilde{w}_k f(x^j\,|\,\tilde{\psi}_k)}{p_k^j f(x^j\,|\,\theta)}$$

Hence, $\Delta(\tilde{\theta},\theta\,|\,X,\dot{\theta}) = \Delta(\tilde{\theta},\theta\,|\,X) + \delta(\tilde{\theta},\theta\,|\,\dot{\theta})$, is a lower bound to $\delta(\tilde{\theta},\theta\,|\,X,\dot{\theta})$. Also $\Delta(\theta,\theta\,|\,X,\dot{\theta}) = \delta(\theta,\theta\,|\,X,\dot{\theta}) = 0$. So, under mild differentiability conditions, both surfaces are tangent, assuring convergence of EM to the nearest local maximum. But maximizing $\Delta(\tilde{\theta},\theta\,|\,X,\dot{\theta})$ over $\tilde{\theta}$ is the same as maximizing

$$Q(\tilde{\theta},\theta) \;\;=\;\; \sum\nolimits_{k,j} p_k^j \log\Big(\tilde{w}_k f(x^j\,|\,\tilde{\psi}_k)\Big) + fl(\tilde{\theta}\,|\,\dot{\theta})$$

and each iteration of the EM algorithm breaks down in two steps:
E-step: Compute $P = E(Z\,|\,X,\theta)$.
M-step: Optimize $Q(\tilde{\theta},\theta)$, given P.

For the Gaussian mixture model, with a Dirichlet-Normal-Wishart prior,

$$\begin{aligned} Q(\tilde{\theta},\theta) \;\;&=\;\; \sum\nolimits_{k=1}^{m}\sum\nolimits_{j=1}^{n} p_k^j\big(\log\tilde{w}_k + \log N(x^j\,|\,\tilde{b}^k,\tilde{R}^k)\big) + fl(\tilde{\theta}\,|\,\dot{\theta}) \\ fl(\tilde{\theta}\,|\,\dot{\theta}) \;\;&=\;\; \log D(\tilde{w}\,|\,\dot{y}) \;+ \sum\nolimits_{k=1}^{m} \log NW(\tilde{b}^k,\tilde{R}^k\,|\,\dot{n}_k,\dot{e}_k,\dot{u}^k,\dot{S}^k) \end{aligned}$$

Lagrange optimality conditions give a simple analytical solutions for the M-step:

$$y = P\mathbf{1} \;\;,\;\; \tilde{w}_k = (y_k + \dot{y}_k - 1) \,/\, \Big(n - m + \sum\nolimits_{k=1}^{m}\dot{y}_k\Big)$$

$$u^k = \tfrac{1}{y_k}\sum\nolimits_{j=1}^{n} p_k^j x^j \;\;,\;\; S^k = \sum\nolimits_{j=1}^{n} p_k^j (x^j - \tilde{b}^k)\otimes(x^j - \tilde{b}^k)'$$

$$\tilde{b}^k = \frac{\dot{n}_k\dot{u}^k + y_k u^k}{\dot{n}_k + y_k} \;\;,\;\; \tilde{V}^k \;\;=\;\; \frac{S^k + \dot{n}_k(\tilde{b}^k - \dot{u}^k)\otimes(\tilde{b}^k - \dot{u}^k)' + \dot{S}^k}{y_k + \dot{e}_k - d}$$

In more general (non-Gaussian) mixture models, if an analytical solution for the M-step is not available, a robust local optimization algorithm can be used, for example [18]. The EM is only a local optimizer, but the MCMC provides plenty of good starting points, so we have the basic elements for a global optimizer. To avoid using many starting points going to a same local maximum, we can filter the (ranked by the posteriori) top portion of the MCMC output using a clustering algorithm, and select a starting point from each cluster. For better efficiency, or more complex problems, the Stochastic EM or SEM algorithm can be used to provide starting points near each important local maximum, see [5], [23], [26], [29].

4 Experimental Tests and Final Remarks

The test case used in this study is given by a sample X assumed to follow a mixture of bivariate normal distributions with unknown parameters, including the number of components. X is the *Iris virginica* data set, with sepal and petal length of 50 specimens (1 discarded outlier). The botanical problem consists of determining whether or not there are two distinct subspecies in the population, [1], [8], [15]. Figure 1 presents the dataset and posterior density level curves for the parameters, θ^* and $\widehat{\theta}$, optimized for the 1 and 2 component models.

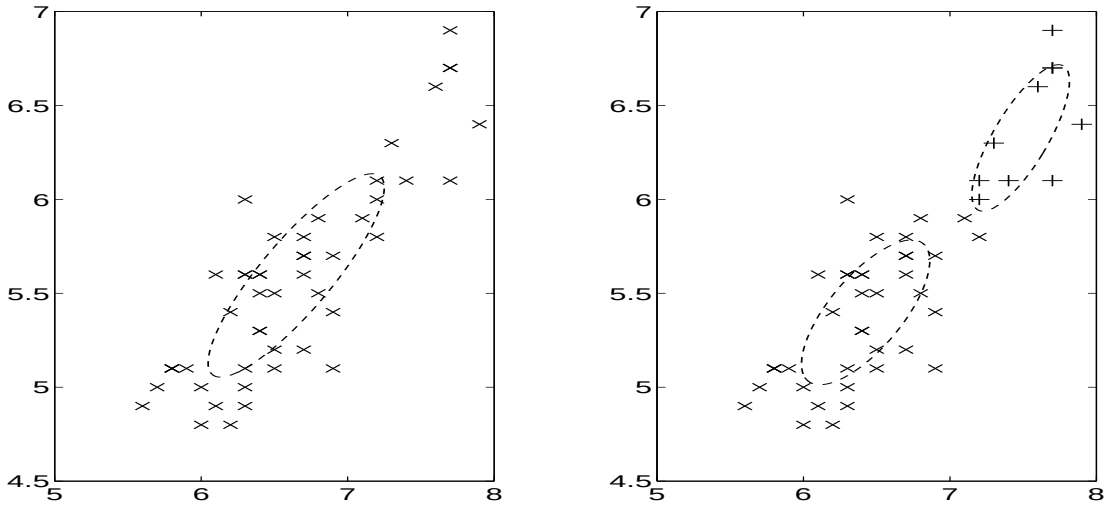


Figure1: Iris virginica data and models with one (left) and two (right) components

In the FBST formulation of the problem, the 2 components is the base model, and the hypothesis to be tested is the constraint of having only 1 component. When implementing the FBST one has to be careful with trapping states on the MCMC. These typically are states where one component has a small number of sample points, that become (nearly) collinear, resulting in a singular posterior. This problem is particularly serious with the Iris dataset because of the small precision, only 2 significant digits, of the measurements. A standard way to avoid this inconvenience is to use flat or minimally informative priors, instead of non-informative priors, see [24].

We used as flat prior parameters: $\dot{y} = \mathbf{1}$, $\dot{n} = 1$, $\dot{u} = u$, $\dot{e} = 3$, $\dot{S} = (1/n)S$. Robert [24] uses, with similar effects, $\dot{e} = 6$, $\dot{S} = (1.5/n)S$.

The FBST selects the 2 component model, rejecting H, if the evidence against the hypothesis is above a given threshold, $\mathrm{Ev}(H) > \tau$, and selects the 1 component model, accepting H, otherwise. The threshold τ is chosen by empirical power analysis, see [14], [17] and [32]. Let θ^* and $\widehat{\theta}$ represent the constrained (1 component) and unconstrained (2 component) maximum a posteriori (MAP) parameters optimized to the Iris dataset. Next, generate two collections of t simulated datasets of size n, the first collection at θ^*, and the second at $\widehat{\theta}$. $\alpha(\tau)$ and $\beta(\tau)$, the empirical type 1 and type 2 statistical errors, are the rejection rate in the first collection and the acceptance rate in the second collection. A small, $t = 500$, calibration run sets the threshold τ so to minimize the total error, $(\alpha(\tau) + \beta(\tau))/2$. Other methods like sensitivity analysis, see [29], [30], [31], and loss functions, see [16], could also be used.

Biernacki and Govaert [4] studied similar mixture problems and compared several selection criteria, pointing as the best overall performers: AIC - Akaike Information Criterion, AIC3 - Bozdogan's modified AIC, and BIC - Schwartz' Bayesian Information Criterion. These are regularization criteria, weighting the model fit against the number of parameters, see [22]. If λ is the model log-likelihood, κ its number of parameters, and n the sample size, then,

$$AIC = -2\lambda + 2\kappa \ , \ \ AIC3 = -2\lambda + 3\kappa \text{ and } \ BIC = -2\lambda + \kappa \log(n) \ .$$

Figure 2 show α, β, and the total error $(\alpha + \beta)/2$. The FBST outperforms all the regularization criteria. For small samples, BIC is very biased, always selecting the 1 component model. AIC is the second best criterion, caching up with the FBST for sample sizes larger than $n = 150$.

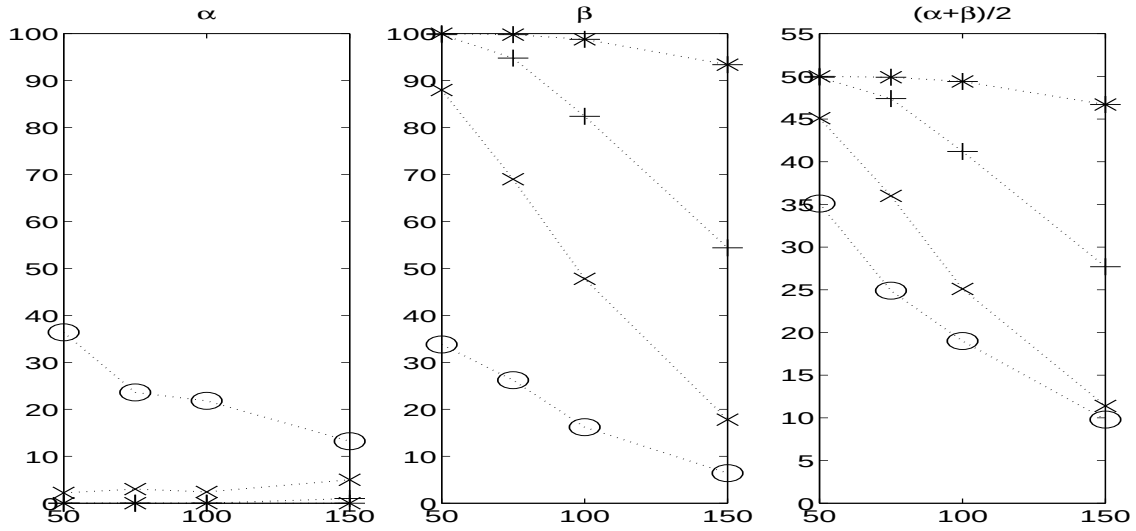


Figure 2: Criteria O= FBST, X= AIC, += AIC3, *= BIC,
Type 1, 2 and total error rates for different sample sizes

Finally, let us point out a related topic for further research: The problem of discriminating between models consists of determining which of m alternative models, $f_k(x, \psi_k)$, more adequately fits or describes a given dataset. In general the parameters ψ_k have distinct dimensions, and the models f_k have distinct (unrelated) functional forms. In this case it is usual to call them "separate" models (or hypotheses). Atkinson [2], although in a very different theoretical framework, was the first to analyse this problem using a mixture formulation,

$$f(x \,|\, \theta) = \sum\nolimits_{k=1}^{m} w_k f_k(x, \psi_k) \ .$$

The general theory for mixture models presented in this article can be adapted to analyse the problem of discriminating between separate hypotheses. This is the subject of the authors' ongoing research with Carlos Alberto de Bragança Pereira and Basílio de Bragança Pereira, to be presented in forthcoming articles.

The authors are grateful for the support of CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPESP - Fundação de Apoio à Pesquisa do Estado de São Paulo.

References

[1] E.Anderson (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society,* 59, 2-5.

[2] A.C.Atkinson (1970). A Method for Discriminating Between Models. *J. Royal Stat. Soc. B,* 32, 323-354.

[3] C.H.Bennett (1976). Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* 22, 245-268.

[4] C.Biernacki G.Govaert (1998). Choosing Models in Model-based Clustering and Discriminant Analysis. Technical Report INRIA-3509-1998.

[5] G.Celeux, D.Chauveau, J.Diebolt (1996). On Stochastic Versions of the EM Algorithm. An Experimental Study in the mixture Case. *Journal of Statistical Computation and Simulation,* 55, 287–314.

[6] M.H.DeGroot (1970). *Optimal Statistical Decisions*. NY: McGraw-Hill.

[7] A.P.Dempster, N.M.Laird, D.B.Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Statistical Society B.* 39, 1-38.

[8] R.A.Fisher (1936). Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*,7,179–188.

[9] J.E.Gentle (1998). *Random Number Generator and Monte Carlo Methods.* NY: Springer.

[10] W.R.Gilks, S.Richardson, D.J.Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* NY: CRC

[11] O.Häggström (2002). *Finite Markov Chains and Algorithmic Applications.* Cambridge Univ.

[12] M.E.Johnson (1987). Multivariate Statistical Simulation. NY: Wiley.

[13] M.C.Jones (1985). Generating Inverse Wishart Matrices. *Comm. Statist. Simula. Computa.* 14, 511–514.

[14] M.Lauretto, C.A.B.Pereira, J.M.Stern, S.Zacks (2003). Comparing Parameters of Two Bivariate Normal Distributions Using the Invariant FBST. *Brazilian Journal of Probability and Statistics,* 17, 147-168.

[15] G.McLachlan, D.Peel (2000). *Finite Mixture Models.* NY: Wiley.

[16] M.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*,10,291–299.

[17] M.R.Madruga, C.A.B.Pereira, J.M.Stern (2003). Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference,* 117,185–198.

[18] J.M.Martinez (2000). BOX-QUACAN and the Implementation of Augmented Lagrangian Algorithms for Minimization with Inequality Constraints. *Computational and Applied Mathematics.* 19, 31-56.

[19] X.L.Meng, W.H.Wong (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica,* 6, 831-860.

[20] D.Ormoneit, V.Tresp (1995). Improved Gaussian Mixtures Density Estimates Using Bayesian Penalty Terms and Network Averaging. *Advances in Neural Information Processing Systems 8*, 542–548. MIT.

[21] C.A.B.Pereira, J.M.Stern, (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.

[22] C.A.B.Pereira, J.M.Stern, (2001). Model Selection: Full Bayesian Approach. *Environmetrics*, 12, 559–568.

[23] G.C.Pflug (1996). *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*. Boston: Kluwer.

[24] C.P.Robert (1996). Mixture of Distributions: Inference and Estimation. in Gilks (1996).

[25] S.Russel (1988). Machine Learning: The EM Algorithm. Unpublished note.

[26] J.C.Spall (2003). *Introduction to Stochastic Search and Optimization.* Hoboken: Wiley.

[27] M.Stephens (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Oxford Univ.

[28] J.M.Stern (1992). Simulated Annealing with a Temperature Dependent Penalty Function. *ORSA Journal on Computing,* 4, 311-319.

[29] J.M.Stern (2003). Significance Tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse. Laptec'03, *Frontiers in Artificial Intelligence and its Applications*, 101, 139–147.

[30] J.M.Stern (2004a). Paraconsistent Sensitivity Analysis for Bayesian Significance Tests. SBIA'04, *Lecture Notes Artificial Intelligence,* 3171, 134–143.

[31] J.M.Stern (2004b). Uninformative Reference Sensitivity in Possibilistic Sharp Hypotheses Tests. MaxEnt 2004, *American Institute of Physics Proceedings*, 735, 581–588.

[32] J.M.Stern, S.Zacks (2002). Testing the Independence of Poisson Variates under the Holgate Bivariate Distribution. The Power of a New Evidence Test. *Statistical and Probability Letters*, 60, 313–320.

The Problem of Separate Hypotheses via Mixture Models

Marcelo de Souza Lauretto[*,†], Silvio R. de Faria Jr.[*], Basilio B. Pereira[**], Carlos A. B. Pereira[*] and Julio M. Stern[*,†]

[*] *University of Sao Paulo, Institute of Mathematics and Statistics*
[†] *lauretto@ime.usp.br jstern@ime.usp.br*
[**] *University of Rio de Janeiro, Medical School and COPPE*

Abstract. This article describes the Full Bayesian Significance Test for the problem of separate hypotheses. Numerical experiments are performed for the Gompertz vs. Weibull life span test.

Keywords: FBST, Life span, Model selection, Separate hypotheses, Significance tests, Reliability.

INTRODUCTION

An important problem in statistics inference consists of deciding which of m alternative models, $f_k(x,\psi_k)$, more adequately fits a given dataset. When the candidate models f_k have distinct (unrelated) functional forms, it is usual to call them "separate" models (or hypotheses). Many discriminate models have been developed, which counterpoise a (null) model $f_1(x,\psi_1)$ against one alternative model $f_2(x,\psi_2)$, providing a measure of evidence in data favoring model 1 over model 2 [1,18]. However, these methods are not capable of give a straight answer when neither candidate model individually describes well the data. Non-parametric tests (e.g. Goodness-of-fit and Kolmogorov-Smirnov), on the other hand, have a comparatively slow convergence rate.

In this article we analyze this problem in the context of mixture models, see [14]. The basic distribution of this statistical model is a weighted sum of two or more candidate pdf's. Deciding if the data comes from a specific distribution is to test if the other distributions weights equal 0. Under this formulation, if neither model describes adequately the data, the test is capable of give a direct answer – a high evidence against all candidate models. As a numerical example we use a classical problem in reliability analysis, the Gompertz vs. Weibull life span, see [11,12].

The Fully Bayesian Significance Test (FBST) is presented by Pereira and Stern [19] as a coherent Bayesian significance test. The FBST is intuitive and has a geometric characterization. In this article the parameter space, Θ, is a subset of R^n, and the hypothesis is defined as a further restricted subset defined by vector valued inequality and equality constraints: $H : \theta \in \Theta_H$ where $\Theta_H = \{\theta \in \Theta \,|\, g(\theta) \leq 0 \wedge h(\theta) = 0\}$. For simplicity, we often use H for Θ_H. We are interested in precise hypotheses, with $\dim(H) < \dim(\Theta)$. $f(\theta)$ is the posterior probability density function.

The computation of the evidence measure used on the FBST is performed in two steps: The optimization step consists of finding f^* and $\widehat{f}$, the constrained (over H) and unconstrained maxima of the posterior. The integration step consists of integrating the

posterior density over the Tangential Set, $\overline{T}$ where the posterior is higher than anywhere in H, i.e., $\overline{T} = \{\theta \in \Theta : f(\theta) > f^*\}$, $f^* = \max_H f(\theta) = f(\theta^*)$, $\widehat{f} = \max_\Theta f(\theta) = f(\widehat{\theta})$, $\overline{\mathrm{Ev}}(H) = \Pr(\theta \in \overline{T} \,|\, x) = \int_{\overline{T}} f(\theta) d\theta$.

$\overline{\mathrm{Ev}}(H)$ is the evidence against H, and $\mathrm{Ev}(H) = 1 - \overline{\mathrm{Ev}}(H)$ is the evidence supporting (or in favor of) H. Intuitively, if $\overline{\mathrm{Ev}}(H)$ is "large", $\overline{T}$ is "heavy", and the hypothesis set is in a region of "low" posterior density, meaning a "strong" evidence against H.

Let us consider the cumulative distribution of the evidence value against the hypothesis, $\overline{V}(\tau) = \Pr(\overline{\mathrm{Ev}} \leq \tau)$, given θ^0, the true value of the parameter. Under appropriate regularity conditions, for increasing sample size, $n \to \infty$, we can state the following:
- If H is false, $\theta^0 \notin H$, then $\overline{\mathrm{Ev}}$ converges (in probability) to one, that is, $\overline{V}(\tau) \to \delta(1)$.
- If H is true, $\theta^0 \in H$, then $\overline{V}(\tau)$, the confidence level, is approximated by the function $\overline{W}(t,h,\tau) = \mathrm{Chi2}\left(t-h, \mathrm{Chi2}^{-1}(t,c)\right)$, where $t = \dim(\Theta)$, $h = \dim(H)$ and $\mathrm{Chi2}(k,x)$ is the cumulative chi-square distribution with k degrees of freedom.

Hence, for large n, to reject H with a level of significance δ, we set $\tau = \overline{W}^{-1}(t,h,1-\delta)$, i.e. set τ such that $\overline{W}(t,h,\tau) = 1-\delta$.

Several FBST applications and examples, efficient computational implementation, interpretations, and comparisons with other techniques for testing sharp hypotheses, can be found in the authors' papers in the reference list. For a FBST review see the on line document [21].

WEIBULL AND GOMPERTZ DISTRIBUTIONS

In this paper we analyze the Gompertz vs. Weibull life span model selection problem. For the importance and interpretation of this problem see [11].

The Weibull hazard and probability density functions, for a failure time $x \geq 0$, given the shape and characteristic life (or scale) parameters, $\beta > 0, \gamma > 0$, are:

$$h_W(x \,|\, \beta, \gamma) = \beta x^{\beta-1}/\gamma^\beta \;, \;\; f_W(x \,|\, \beta, \gamma) = (\beta x^{\beta-1}/\gamma^\beta)\exp(-(x/\gamma)^\beta) \;.$$

The Gompertz hazard and probability density functions, for a failure time $x \geq 0$, given the parameters, $\alpha > 1, \lambda > 0$, are:

$$h_G(x \,|\, \alpha, \lambda) = \lambda \alpha^x \;, \;\; f_G(x \,|\, \alpha, \lambda) = \lambda \alpha^x \exp(-(\alpha^x - 1)\lambda / \log \alpha) \;.$$

The Gompertz distribution exhibits a strong nonlinear correlation between the parameters α and λ, see Figure 1A. This correlation explains the *compensation law of mortality*, which states that higher values for the parameter α are compensated by lower values of parameter λ in different populations of a given species: $ln(\lambda) = ln(M) - B\alpha$, where B and M are universal species-specific invariants, see [11]. As a result, the Gompertz density in its original form is not log-concave. As we shall discuss later, we use adaptive samplers for the parameters, which depend on the shape of the density function – preferably log-concave distributions. In order to separate the parameters α and λ, diminishing this nonlinear dependence and enhancing the shape of density function for sampling, we adopt the reparameterization $u = 1/\log \alpha$ and $v = \log(\log \alpha)/\lambda)$, suggested by Meeker and Escobar [17], see Figure 1B.

The log-likelihoods of Weibull and (reparameterized) Gompertz models and their respective gradients (used for maximum likelihood estimation) are:

$$\begin{aligned}
L_W(\beta,\gamma|X) &= n\log\beta - n\beta\log\gamma + (\beta+1)\sum_j \log x - \sum_j (x_j/\gamma)^\beta \ , \\
dL_W/d\beta &= n/\beta - n\log\gamma + \sum_j \log x_j - \sum_j (x_j/\gamma)^\beta \log(x_j/\gamma) \ , \\
dL_W/d\gamma &= -n\beta/\gamma + \beta/\gamma \sum_j (x_j/\gamma)^\beta \ , \\
L_G(u,v|X) &= -n\log u - nv + \sum_j x_j/u + n/\exp(v) - \sum_j \exp(x_j/u - v) \ , \\
dL_G/du &= -n/u - \sum_j x_j/u^2 + \sum_j x_j/u^2 \exp(x_j/u - v) \ , \\
dL_G/dv &= -n - n/\exp(v) + \sum_j \exp(x_j/u - v) \ .
\end{aligned}$$

MIXTURES OF SEPARATE MODELS

Given a dataset $X = \{x_1, x_2 \ldots x_n\}$ and distinct alternative probability densities, $f_1(X\,|\,\psi_1), f_2(X\,|\,\psi_1), \ldots, f_m(X\,|\,\psi_m)$, where ψ_k are (vector) parameters, the problem of interest is to measure the evidence in favour of each model for fitting the dataset. In this paper, the consider a general model including all candidate distributions, where the choice of a specific distribution is a special case. The origin of this model comes in the work of Cox [7], who suggested that, in the presence of two alternative models, the p.d.f of data could be taken proportional to

$$f(x\,|\,w,\psi_1,\psi_2) \propto f_1(x\,|\,\psi_1)^{w_1} f_2(x\,|\,\psi_2)^{w_2} \ , \ \ w > 0\,|\,w\mathbf{1} = w_1 + w_2 = 1 \ ,$$

Then, deciding if the model 1 is adequate to describe the data is to test the hypothesis $H_1 : w_1 = 1$ against the hypothesis $w \neq 1$. Atkinson [2] developed this idea for some distributions of the exponential class, writing the density as

$$f(x\,|\,w,\psi_1,\psi_2) = \frac{f_1(x\,|\,\psi_1)^{w_1} f_2(x\,|\,\psi_2)^{w_2}}{\int f_1(y\,|\,\psi_1)^{w_1} f_2(y\,|\,\psi_2)^{w_2} dy} \ .$$

In this paper, we consider that the p.d.f. of data is a convex linear combination of the fixed candidate densities: denoting $\theta = [w, \psi_1, \ldots \psi_m]$,

$$f(x\,|\,\theta) = w_1 f_1(x\,|\,\psi_1) + \ldots + w_m f_m(x\,|\,\psi_m) \ , \ \ w \geq 0\,|\,w\mathbf{1} = 1 \ .$$

The likelihood then is

$$f(X\,|\,\theta) = \prod_{j=1}^n \sum_{k=1}^m w_k f_k(x_j\,|\,\psi_k) \ .$$

Here it is important to remember some key concepts of mixture models. In mixture analysis for unsupervised classification, we assume that the data come from one or more subpopulations (classes), distributed under distinct densities. The evidence in favor of the existence of more than one subpopulation will be higher if some subsets of data are more adequately fitted by a particular component of the mixture, where other subsets are

better fitted by another components. In order to detect this situation, the mixture model must be able to infer the (probability of) data classifications. The real classifications are considered non observable and, for this reason, called *hidden* or *latent* variables. The problem of deciding if one single candidate distribution fits adequately the data is analogous to decide the number of components in a traditional mixture model, and the behavior of the system will be also similar: if the candidate model does not fit well the data, some observed points may be better described by a particular component of mixture, where the remaining will be better fitted by other components.

A sample j of class $k = c(j)$ is distributed with density $f_k(x_j \,|\, \psi_k)$. The boolean classification matrix Z indicates whether or not x_j is of class k, i.e. $z_j^k = 1$ iff $c(j) = k$. Conditioning on the latent variables we can rewrite:

$$\begin{aligned} f(x_j \,|\, \theta) &= \sum_{k=1}^{m} f_k(x_j \,|\, \theta, z_k^j) f(z_j^k \,|\, \theta) = \sum_{k=1}^{m} w_k f_k(x_j \,|\, \psi_k) \,, \\ f(X \,|\, \theta) &= \prod_{j=1}^{n} f(x_j \,|\, \theta) = \prod_{j=1}^{n} \sum_{k=1}^{m} w_k f_k(x_j \,|\, \psi_k) \,. \end{aligned}$$

Given the mixture parameters, θ, and the observed data, X, the conditional classification probability matrix, $P = f(Z \,|\, X, \theta)$, is given by:

$$p_k^j = f(z_j^k \,|\, x_j, \theta) = \frac{f_k(z_j^k, x_j \,|\, \theta)}{f(x_j \,|\, \theta)} = \frac{w_k f_k(x_j \,|\, \psi_k)}{\sum_{k=1}^{m} w_k f(x_j \,|\, \psi_k)} \,.$$

We use y_k for the number of samples of class k, i.e. $y_k = \sum_j z_j^k$, or $y = Z\mathbf{1}$.

The density for the "completed" data, X, Z, is:

$$f(X, Z \,|\, \theta) = \prod_{j=1}^{n} f_{\psi_{c(j)}}(x_j \,|\, \psi_{c(j)}) f(z_j^k \,|\, \theta) = \prod_{k=1}^{m} \left({w_k}^{y_k} \prod_{j \,|\, c(j)=k} f_k(x_j \,|\, \psi_k) \right) \,.$$

In the remaining of this section we discuss the FBST formulation for the Weibull vs. Gompertz mixture model. The conjugate prior for a multinomial distribution is a Dirichlet distribution:

$$\begin{aligned} M(y \,|\, n, w) &= n! \big/ (y_1! \ldots y_m!) \; {w_1}^{y_1} \ldots {w_m}^{y_m} \,, \\ D(w \,|\, y) &= \Gamma(y_1 + \ldots + y_k) \big/ (\Gamma(y_1) \ldots \Gamma(y_k)) \; \prod_{k=1}^{m} {w_k}^{y_k - 1} \,, \end{aligned}$$

with $w > \mathbf{0}$ and $w\mathbf{1} = 1$. Prior information given by $\dot{y}$, and observation y, result in the posterior parameter $\ddot{y} = \dot{y} + y$. Here we take the non-informative prior given by $\dot{y} = \mathbf{1}$. We also consider a improper uniform prior for (β, γ, u, v). Therefore, the posteriori is

$$\begin{aligned} f(\theta \,|\, X) &\propto f(X \,|\, \theta) = \prod_{j=1}^{n} \left(p_j^1 w_1 f_W(x_j \,|\, \beta, \gamma) + p_j^2 w_2 f_G(x_j \,|\, \alpha, \lambda) \right) \,, \\ p_j^1 &= \frac{w_1 f_W(x_j \,|\, \beta, \gamma)}{w_1 f_W(x_j \,|\, \beta, \gamma) + w_2 f_G(x_j \,|\, \alpha, \lambda)} \,, \;\; p_j^2 = 1 - p_j^1 \,. \end{aligned}$$

The hypotheses of interest are $H_1 : w_1 = 1 \wedge w_2 = 0$ and $H_2 : w_1 = 0 \wedge w_2 = 1$. The FBST procedure for testing $H_k, k = 1, 2$ consists of two steps:

- Estimate the maximum of the log-likelihood L_k^* under H_k, which corresponds to the maximum log-likelihood under the corresponding single component distribution.

- Estimate the e-value supporting the hypothesis H_k, that is, the ratio

$$\mathrm{Ev}(H_k) = \frac{\int_{T_k} f(\theta \,|\, X) d\theta}{\int_{\Theta} f(\theta \,|\, X) d\theta} \ , \ T_k = \{\theta \in \Theta \,|\, L(\theta) \leq L_k^*\} \ .$$

Notice that since the likelihood normalization constant is the same for both numerator and denominator, so it is cancelled and can therefore be ignored in the computational procedure. For the optimization step, we used the `Algencan-Tango` solver, which source code and detailed description are freely distributed (see internet link at the reference), see [4,5].

In order to perform the integration over the posterior measure, we used a Gibbs sampling Markov Chain Monte Carlo algorithm, MCMC. Given the current vector parameter θ^i, we compute P. Given P, we draw Z from $f(z_j \,|\, p_j)$, a simple multinomial distribution. Given the latent variables, Z, we separate the samples of classes 1 and 2. In the Weibull component, we draw a parameter value $[\beta^{i+1}, \gamma^{i+1}]$ with density proportional to the partial likelihood $\prod_{j|c(j)=1} f_W(x_j \,|\, \beta, \gamma)$. The same idea is applied to draw the Gompertz parameters $[\alpha^{i+1}, \lambda^{i+1}]$. Given $\ddot{y} = Z\mathbf{1} + \dot{y}$, we can draw a new weight vector $[w_1^{i+1}, w_2^{i+1}]$ using a Dirichlet distribution $D(w \,|\, \ddot{y}_1, \ddot{y}_2)$. At the end of iteration (i), we have a new vector parameter $\theta^{i+1} = [w_1^{i+1}, w_2^{i+1}, \beta^{i+1}, \gamma^{i+1}, \alpha^{i+1}, \lambda^{i+1}]$, and can begin iteration $(i+1)$.

We do not know a direct method to draw the parameters from the Weibull or Gompertz likelihood. For this purpose we used the adaptive sampler `HITRO`, see [13,20,22]. `HITRO` combines the multivariate Ratio-of-Uniforms method with the Hit-and-Run sampler. The Ratio-of-Uniforms transformation maps the region below the p.d.f f, i.e. $G(f) = \{(x,y) : 0 < y < f(x)\}$ into the region

$$A(f) = A_{r,m}(f) = \left\{ (u,v) : 0 < v < f\left(\frac{u}{v^r} + m\right)^{1/(rn+1)} \right\}$$

by means of the transformation

$$(u,v) \mapsto (x,\ y) = \left(\frac{u}{v^r} + m,\ v^{rn+1}\right) \ .$$

The vector m must be a point near the mode (in our implementation, we set m as the mode). The method relies on the theorem that, if (u,v) is uniformly distributed over $A(f)$, then $x = u/v^r + m$ has probability density function $f(x)/\int f(z)dz$. The Hit-and-run sampler is used for generating points (u,v) uniformly over $A(f)$.

NUMERICAL EXPERIMENTS AND FINAL REMARKS

We run some numerical experiments in order to evaluate the FBST performance on our problem of separate models. The experiments were based on the IBGE data bank for the mortality of Brazilian male population in the year of 2005, available on line at *http://www.ibge.gov.br /home/estatistica/populacao/tabuadevida/2005/default.shtm*. We used the mortality rate table from ages 5 to 80, hence avoiding the early infancy or burn-in period, see [3,17].

The experiments were based on simulated data, drawn from four distributions, the parameters of which have always been chosen to provide the best fit to the IBGE data bank. The distributions fitted were: (1)-Weibull, (2)-Gompertz, (3)-Gamma, and (4)-Beta (rescaled), see Figure 2. Our main interest was to measure the convergence rate of correct decisions, concerning the acceptance / rejection of the Weibull vs. Gompertz hypotheses, when using the FBST on the mixture model. Of course, in cases (1) and (2) we want to accept the correct hypothesis and reject the false one, whereas in cases (3) and (4) we want to reject them both.

As acceptance / rejection threshold, we adopted the critical level τ according to criterion presented in section 1, with a significance level of 5%. Since the mixture model and the restricted model have 5 and 2 degrees of freedom, respectively, we have $\tau = \overline{W}^{-1}(5,2,0.95) = 0.83$. Therefore, we reject H if $\overline{\mathrm{Ev}}(H) > 0.83$, or equivalently if $\mathrm{Ev}(H) < 0.17$. Using each of the four fitted distributions we generated 500 samples of size $n = 30$, 50, 75, 100, 150, 200, 300, 400 and 500.

We have compared the performance of the FBST with the Kolmogorov–Smirnov (KS) test, [9]. In this test, the goodness of fit measure is taken to be the Kolmogorov distance $D_n^* = D(F_n, F^*) = sup_x |F_n(x) - F^*(x \mid \theta)|$, where F_n denotes the sample (empirical) distribution and F^* denotes the theoretical distribution to be tested. Due to difficulty in estimate θ which minimizes $D(F_n, F^*)$, it is usually adopted the maximum likelihood estimator for θ. Kolmogorov and Smirnov demonstrated in 1930's that, if the null hypothesis $F(X) = F^*(X \mid \theta)$, then $lim_{n\to\infty} Pr(\sqrt{n} D_n^* \leq t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2 t^2)$. The distribution at the right side of this equation allows one to compute the significance (p-value) of D_n^*. For a meaningful comparison, we also used a 5% significance level.

The hole batch of 500 simulations for each of the 4 cases and 9 sample sizes, took about 2 day of computation on a Intel Pentium server, or about 10 seconds per test. Computing time was dominated by *Hitro*, a flexible and robust but generic subroutine. Hence, its substitution by a tailor made and more efficient sampler could enhance the program computational performance.

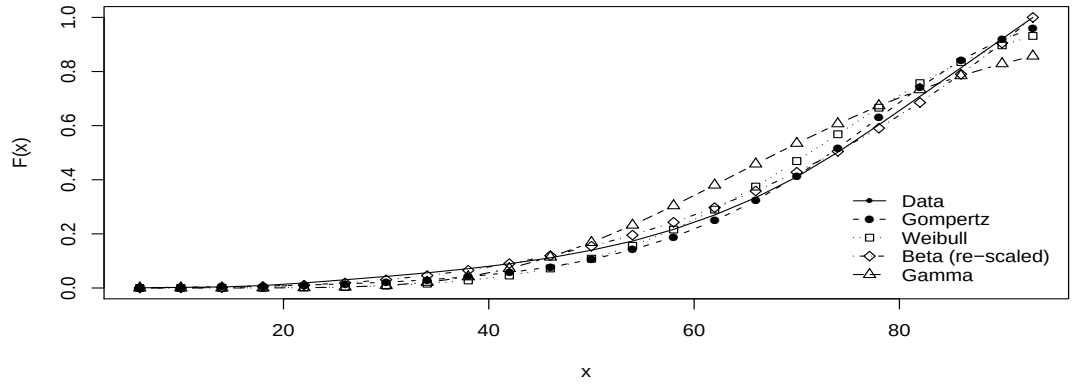
Figure 3 summarized the correct decision rates in the numerical simulations. The Weibull distribution can approximate very well a Gamma distribution. This explains the relatively slow convergence in the decision to reject the Weibull hypothesis in the simulations from the Gamma.

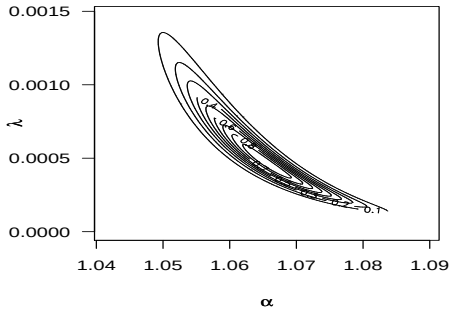
As expected, the FBST had a good performance. Moreover its implementation is straightforward, following the guidelines presented in [19,21]. It would be interesting to replace the Kolmogorov-Smirnov benchmark with a parametric alternative, like some form of jump MCMC. However, as far as the authors know, none is available at this time. The authors intend to collaborate with other research groups in order to develop and implement such algorithms.

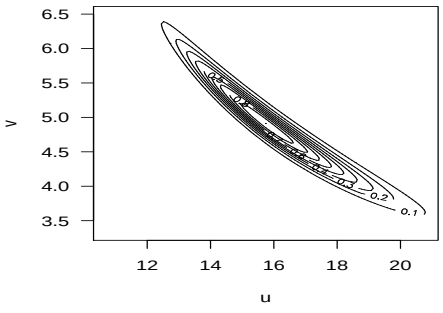
Acknowledgments: The authors are grateful for the support of CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo.

REFERENCES

1. M.A.Araujo, B.B.Pereira (2007). A Comparison of Bayes Factors for Separated Models: Some Simulation Results. *Communications in Statistics-Simulation and Computation* 36(2), 297–309.
2. A.C.Atkinson (1970). A Method for Discriminating Between Models. *J. R. Statist Soc. B* 32, 323–354.
3. R.E.Barlow, F.Prochan (1981). *Statistical Theory of Reliability and Life Testing Probability Models.* Silver Spring: To Begin With.
4. E.G.Birgin, J.M.Martnez and M.Raydan (2000). Nonmonotone Spectral Projected Gradient Methods on Convex Sets. *SIAM Journal on Optimization*, 10, 1196-1211. Software and documentation available at *www.ime.usp.br/∼egbirgin/tango/* .
5. E.G.Birgin , J.M.Martnez (2002). Large-scale Active-Set Box-Constrained Optimization Method with Spectral Projected Gradients. *Computational Optimization and Applications*, 23, 101-125.
6. W.Borges, J.M.Stern (2006). Evidence and Compositionality. p.307-315 in J.Lawry et al., *Soft Methods for Integrated Uncertainty Modelling.* NY: Springer.
7. D.R.Cox (1962). Further Results on Tests of Separated Families of Hypotheses. *J. R. Statist Soc. B* 24, 406-423.
8. D.R.Cox and D.Oakes (1984). *Analysis of Survival Data*. Monographs on Statistics and Applied Probability, Chapman & Hall.
9. M.H.DeGroot (1986). *Probability and Statistics.* Addison–Wesley.
10. B.Dodson (1994). *Weibull Analysis.* Milwaukee: ASQC Quality Press.
11. L.A.Gavrilov and N.S.Gavrilova (1991). *The Biology of Life Span: A Quantitative Approach*. New York: Harwood Academic Publisher.
12. L.A.Gavrilov and N.S.Gavrilova (2001). The Reliability Theory of Aging and Longevity. *J. Theor. Biol.* 213, 527–545.
13. R.Karawatzki, J.Leydold, K.Pötzelberger (2005). Automatic Markov Chain Monte Carlo Procedures for Sampling from Multivariate Distributions. Department of Statistics and Mathematics Wirtschaftsuniversität Wien Research Report Series. Report 27, December 2005. Software available at *http://statistik.wu-wien.ac.at/arvag/software.html*.
14. M.S.Lauretto, J.M.Stern (2005). FBST for Mixture Model Selection. Maxent'2005, *AIP Conf. Proc.* 803, 121–128.
15. M.R.Madruga, L.G.Esteves, S.Wechsler (2001). On the Bayesianity of Pereira-Stern Tests. *Test*, 10, 291–299.
16. M.R.Madruga, C.A.B.Pereira, J.M.Stern (2003). Bayesian Evidence Test for Precise Hypotheses. *Journal of Statistical Planning and Inference,* 117, 185–198.
17. W.Q.Meeker and L.A.Escobar (1998). Statistical Methods for Reliability Data. Wiley Series in Probability and Statistics.
18. B.B.Pereira (2005). Separate Families of Hypotheses. in P.Armitage, T.Colton (eds.) *Encyclopedia of Biostatistics* (2nd.ed.). 7, 4881-4886. NY: Wiley.
19. C.A.B.Pereira, J.M.Stern, (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy Journal*, 1, 69–80.
20. R.L.Smith (1984). Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research 32*, 1296–1308.
21. J.M.Stern (2007). Cognitive Constructivism, Eigen-Solutions, and Sharp Statistical Hypotheses. *Cybernetics and Human Knowing,* 14,1, 9-36.
24. I.Vaduva (1984). Computer Generation of Random Vectors Based on Transformation of Uniformly Distributed Vectors. In *Probability Theory, Proc. 7th Conf.*, Brasov/Rom, 589–598.
25. J.C.Wakefield, A.E.Gelfand, A.F.Smith (1991). Efficient Generation of Random Variates via Ratio-of-Uniforms Method. *Statist. Comput.* 1,2, 129-133.







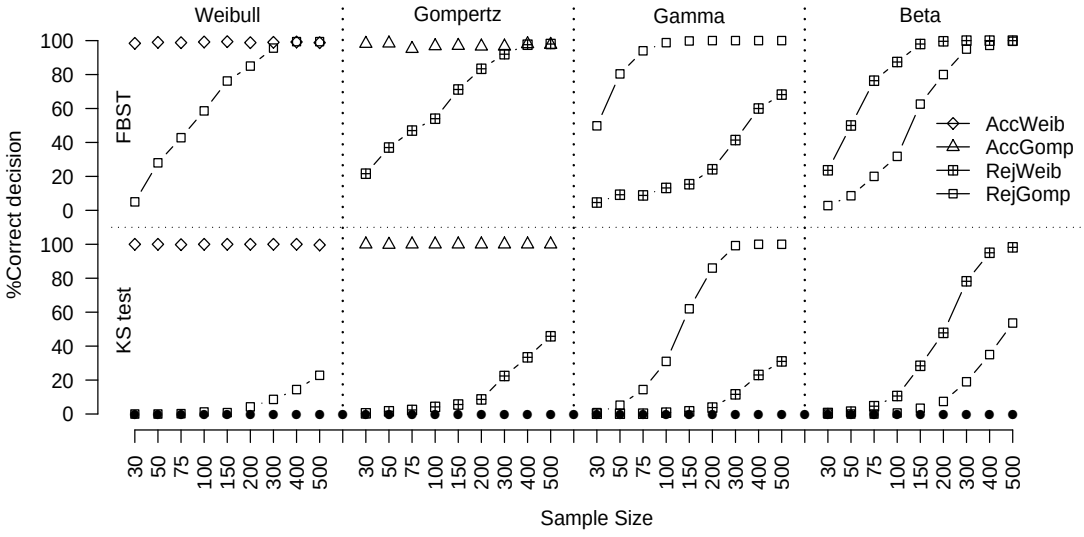


Figure 1: IBGE Brazilian mortality rates and fitted distributions.
Figure 2A,B: Contour plots for Gompertz density and reparameterization.
Figure 3: Correct decision rates on numerical simulations.