Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Departamento de Estatística

MODELO COM MISTURA DE MULTINOMIAIS APLICADO À IDENTIFICAÇÃO DE PROTEÍNAS SIMILARES

Ricardo Galante Coimbra

Orientador: Prof. Dr. Luis Aparecido Milan

Dissertação apresentada à CPG do Departamento de Estatística da Universidade Federal de São Carlos como parte dos requisitos necessários para obtenção do título de Mestre em Estatística

São Carlos / SP
Fevereiro – 2005

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C679mm

Coimbra, Ricardo Galante.
Modelo com mistura de multinomiais aplicado à identificação de proteínas similares / Ricardo Galante Coimbra. -- São Carlos : UFSCar, 2005.
82 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2005.

1. Estatística - análise. 2. Mistura de distribuições . 3. Variável latente. 4. Gibbs sampling. 5. DIC. 6. Fator de Bayes. I. Título.

CDD: 519.5 (20ª)

“Um dia é preciso parar de sonhar e, de algum modo, partir.”

Amyr Klink

“Eu dedico este trabalho aos meus pais,
Domingos Coimbra *in memorian* e
Célia Galante Coimbra.”

Agradecimentos

Ao *Senhor Jesus* por me dar à oportunidade e a benção de concluir este trabalho.

À minha querida *mãe* que esteve ao meu lado em todos os momentos, orando por mim, me incentivando, me confortando e me ensinando a olhar a vida com os olhos da esperança.

Às minhas queridas irmãs *Tânia Galante Coimbra* e *Denise Galante Coimbra* pelo apoio e incentivo que foram fundamentais para a conclusão deste trabalho.

À minha linda namorada *Thaissa Lopes de Melo* pelo amor e companheirismo em todos os momentos e pela ajuda ímpar em bioquímica e genética.

A toda minha *família* pela oração constante e pelo apoio nesta jornada.

Aos meus grandes amigos *Alex Cadei*, *Eliel Robles*, *Giovani Bianchini*, *Ricieri Nascimento* e *Vivian Davies* pela amizade, força e apoio.

Aos meus *colegas* de mestrado e graduação.

Aos *funcionários* do departamento de estatística

Aos meus *professores* do mestrado e graduação em especial ao Professor Dr. *José Galvão Leite*, Professora Dra. *Teresa Cristina Martins Dias* e Professor Dr. *Carlos Alberto Ribeiro Diniz*.

Ao meu orientador *Luís Milan* que, em todos os momentos, acreditou em mim, me orientou e me ensinou com sabedoria e dignidade.

Resumo

As proteínas são moléculas importantes das células, pois participam desde a construção das estruturas celulares até a transmissão de informações genéticas entre gerações.

Uma proteína pode ser caracterizada pela sua função, sendo que esta função é determinada pela seqüência de aminoácidos que compõe a sua estrutura. Determinar a função protéica é importante quando, por exemplo, se pesquisa a cura de doenças ou se pesquisa a fabricação de novos medicamentos.

Neste trabalho utilizamos uma metodologia *bayesiana* de inferência estatística para inferir sobre o modelo com mistura de distribuições multinomiais e variáveis latentes para identificar proteínas com funções similares.

Verificamos a performance da modelagem proposta em separar em grupos as proteínas com funções similares através de simulação.

Ao final fazemos uma aplicação a um conjunto de dados reais.

Abstract

The proteins are important molecules from the cells, whereas they take part since the construction of cell's framing until the transmission of the genetic information between the generations.

A protein can be characterized by its function and its function is determined by the sequence of amino acids that determines its structure.

To determine the protein's function is important, for instance, in a research about the cure of diseases or searching for new drugs.

In this research we use a *bayesian* statistical methodology with mixture of multinomial and latent variables to identify proteins with similar function.

We use simulations to verify the performance of the statistical model for identifying the similar proteins.

At the end we apply the modeling to a real data set.

Sumário

Capítulo 1

Introdução

1.1 Preliminares

Modelos que utilizam mistura de distribuições têm sido empregados em diversas áreas do conhecimento. Contudo, a primeira aplicação documentada do que hoje é chamado de misturas de distribuições vem do início do século XIX. Em 1890, o Professor W.R. Weldon levou ao estatístico Karl Pearson questões sobre um experimento no qual estava trabalhando. Os dados deste experimento eram oriundos de medidas corporais de siris, onde estas medidas eram razões entre tamanho da cabeça e comprimento do corpo inteiro para 1000 siris. Observando graficamente os dados, Pearson notou um comportamento não usual, pois o gráfico sendo bimodal, sugeria uma união entre duas distribuições. Weldon então, sugeriu que a razão para esta assimetria poderia estar no fato que entre os 1000 siris coletados havia duas espécies distintas, mas quando foi feita a coleta não se discriminou estas espécies. Pearson, analisando os dados, propôs que a distribuição das medidas em questão poderiam ser modeladas através de uma soma ponderada de duas distribuições normais, onde estas ponderações seriam as proporções de siris vindas das duas espécies em questão. Matematicamente, Pearson (1894), sugeriu a seguinte distribuição, $f(x) = pN(\mu_1, \sigma_1) + (1-p)N(\mu_2, \sigma_2)$, sendo que p representa a proporção de indivíduos na população para uma das espécies de siris seguindo uma distribuição normal

com média μ_1 e desvio padrão σ_1 e, $(1-p)$ é a proporção de indivíduos na população para a outra espécie de siris seguindo uma distribuição normal com média μ_2 e desvio padrão σ_2.

Um outro exemplo de aplicação de misturas finitas de distribuições é encontrado em estudos relacionados a esquizofrenia, que é uma doença da personalidade. Levine (1981), coletou uma amostra onde foi verificada a idade na qual a pessoa apresentou os primeiros sintomas de esquizofrenia, e outra amostra onde foi verificada a idade da pessoa na primeira consulta a um médico devido a problemas esquizofrênicos. Neste estudo, foram detectadas diferenças de comportamento entre pessoas do sexo masculino e do sexo feminino sendo proposto um modelo utilizando misturas de distribuições considerando o sexo do indivíduo.

Existem aplicações de misturas de distribuições, como nos casos citados acima onde se conhece *de antemão,* o número de componentes da mistura, mas existem aplicações onde este número é desconhecido. Um exemplo disto, citado em Titterington *et al.* (1987), está na sedimentologia, que é o estudo dos produtos dos processos físicos, químicos e biológicos atuantes na superfície da terra. Para se modelar o tamanho dos grãos de areia teríamos uma distribuição com mistura onde deveríamos levar em consideração a constituição e a relação entre os diferentes minerais que constituem a areia. Neste caso, temos uma distribuição com mistura onde o número de componentes é desconhecido.

Uma área onde a modelagem com distribuições mistas apresenta resultados satisfatórios é na biologia molecular que pode ser visto nos trabalhos de Brown *et al. (*1993), Karplus (1995), e Sjölander *et al.* (1996). Nesta área, ao se modelar seqüências de aminoácidos, devemos levar em consideração as características químicas inerentes à estas moléculas e desta forma a utilização de modelagem com distribuições mistas é bastante conveniente.

Nos últimos anos, o uso de mistura de distribuições utilizando inferência bayesiana vem crescendo significativamente, como é relatado, por exemplo, em Robert e Soubiran (1993), Diebolt e Robert (1994), Roeder e Wasserman (1995), Mengersen e Robert (1996),

Robert (1996), e isto, é devido em parte ao desenvolvimento teórico de métodos eficientes de simulação estocástica.

1.2 Proposta

Segundo Alberts *et al.*(2002), as proteínas são os principais constituintes das células determinando não somente a estrutura das mesmas, mas também as suas funções. As proteínas são construídas à partir da união de aminoácidos sendo que normalmente são encontrados 20 tipos de aminoácidos em sua formação. Os aminoácidos possuem características diversas e por conseqüência produzem proteínas diversas. Por sua vez, os aminoácidos são codificados por 4 bases nitrogenadas durante o processo de tradução nos ribossomos das células. Assim, podemos observar as proteínas através dos aminoácidos que as compõem e através das bases nitrogenadas que codificam estes aminoácidos. De uma forma geral, quando proteínas são compostas por aminoácidos similares elas possuem funções similares.

Nesta dissertação fazemos uma aplicação de mistura finita de distribuições utilizando inferência bayesiana em proteínas. O nosso interesse é propor um modelo estatístico que ao comparar duas ou mais proteínas seja capaz de agrupá-las em relação às suas funções. A aplicabilidade deste modelo é grande pois, em muitos casos, não se sabe ou se sabe pouco em relação à função de uma proteína e este modelo poderá ser uma alternativa para indicar esta função. Isto pode ser feito da seguinte forma: são observadas, através do modelo, grupos de proteínas que possuam características conhecidas e que de antemão já se saiba qual ou quais são suas funções. Agregada a esta amostra, está a proteína cuja função é desconhecida. Ao final deste processo, o grupo em que esta proteína for alocado nos dará uma indicação a respeito de sua função.

1.3 Apresentação dos capítulos

Os capítulos desta dissertação estão organizados de forma que, no capítulo 2, fazemos uma breve explanação sobre alguns aspectos relacionados à proteínas. Neste capítulo abordamos de maneira sucinta como uma proteína pode ser vista através dos aminoácidos que a constitui e também, como esta proteína pode ser vista através das bases nitrogenadas que codificam estes aminoácidos. No capítulo 3 exibimos um resumo contendo definições e exemplos sobre o processo de Bernoulli, distribuição binomial e distribuição multinomial. Neste capítulo adotamos uma abordagem elementar em função do caráter interdisciplinar desta pesquisa. No quarto capítulo apresentamos o modelo proposto nesta dissertação, ou seja, o modelo com mistura de distribuições multinomiais apresentando suas características. No capítulo 5, considerando a presença de duas componentes ou de dois grupos protéicos no modelo, mostramos duas aplicações do modelo proposto com dados simulados apresentando seus respectivos resultados. No capítulo 6 exibimos uma análise de performance onde verificamos a eficiência do modelo para diferentes comprimentos de seqüências de aminoácidos e de bases nitrogenadas. Esta verificação é feita através de dados simulados sendo que, utilizamos nesta simulação os dados do capítulo 5. Para fazermos esta análise, verificamos a proporção de acertos com que o modelo separa as proteínas em relação aos grupos protéicos. Apresentamos dois exemplos explicando detalhadamente esta análise. No capítulo 7, para detectarmos o número de componentes no modelo de mistura, modelamos os dados simulados para vários números de grupos protéicos e utilizamos duas técnicas de seleção de modelos conhecidas como fator de *bayes* e DIC, *deviance information criterion.* A seleção de modelos tem como objetivo identificar o número de componentes, ou de grupos funcionais protéicos presentes na amostra analisada, e verificar se o modelo consegue captar a presença deste número. Desta forma generalizamos a utilização do método para mais de dois grupos funcionais de proteínas na amostra. No capítulo 8 apresentamos os resultados obtidos, as discussões e as conclusões deste trabalho.

Capítulo 2

Proteínas

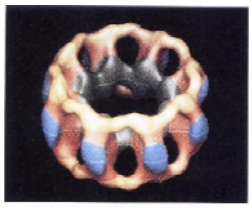
2.1 Introdução

Apresentamos de maneira sucinta neste capítulo alguns aspectos relacionados às proteínas citando alguns fundamentos básicos que são importantes no entendimento do modelo estatístico empregado na dissertação.

Proteínas são caracterizadas como as moléculas trabalhadoras em uma célula, que executam um programa de atividades codificadas pelos genes. Este programa, requer um esforço coordenado de diferentes tipos de proteínas, como construção de estruturas na célula ou processamento de materiais através de reações químicas intra e extracelulares. Estas funções são de tal maneira entrelaçadas na fisiologia das células que seria difícil encontrar alguma atividade celular que não envolvesse proteínas. A palavra proteína é derivada da palavra grega *proteios* que significa "primário" e foi dada por Jöns J. Berzelius em 1838 no sentido de enfatizar a importância desta classe de moléculas. Em praticamente todo processo biológico as proteínas desempenham um papel chave.

As proteínas podem ser agrupadas em famílias nas quais cada membro apresenta uma conformação tridimensional que é similar a todos os outros membros da família. Esta conformação tridimensional fornece à proteína a sua função. Desta forma podemos dizer que dentro de cada família ou grupo protéico os membros possuem funções similares. As

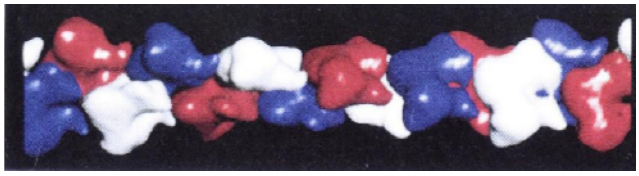
estruturas exibidas na Figura 1 e 2 ilustram como o *design* e a função protéica estão relacionados.



Fonte: Molecular Cell Biology

Figura 1: Complexo de Proteínas

A Figura 1 mostra um complexo de várias proteínas situadas na membrana do núcleo celular que atuam como um canal pelo qual moléculas entram e saem do citoplasma para o núcleo celular.



Fonte: Molecular Cell Biology

Figura 2: Cytoskeleton

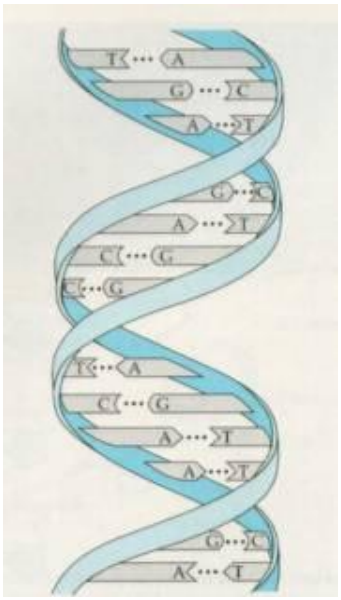
Na Figura 2 temos a representação da proteína *cytoskeleton*, que tem por função reforçar a estrutura do citoplasma assim como o aço reforça a estrutura de um concreto. Tanto na Figura 1 como na Figura 2, temos exemplos de como a estrutura tridimensional e a função protéica são interligadas.

Como o intuito desta dissertação é propor um modelo estatístico que seja capaz de identificar proteínas semelhantes através das moléculas que as formam, é importante que

analizemos mais detalhadamente estas moléculas. Desta forma, vamos abordar sucintamente o tema relacionado ao DNA e RNA.

2.2 DNA e RNA

O DNA é uma molécula que está presente no núcleo das células dos organismos vivos. Ela é conhecida como "molécula da vida", por entre outras coisas conseguir se auto-duplicar, ou seja, essa molécula tem a capacidade de gerar uma cópia idêntica de si mesma. Para isso ela utiliza recursos presentes na própria célula, como proteínas, energia etc. O DNA é formado de duas fitas de nucleotídeos, ligadas uma a outra através de pontes de hidrogênio entre as bases nitrogenadas de cada fita. Cada nucleotídeo é composto por um açúcar que, no caso do DNA, é a *desoxirribose*, um grupo fosfato e por uma base nitrogenada que pode ser *adenina*, *citosina*, *guanina* ou *timina* que se unem em uma ordem determinada, ou seja, *adenina* com *timina,* e *citosina* com *guanina.*



Fonte: www.ib.usp.br/ evolucao/inic/dna2.jpg

Figura 3: Ligações no DNA

Na Figura 3 temos a representação da molécula de DNA mostrando como as bases nitrogenadas se ligam em uma seqüência determinada.

2.2.1 Transcrição

Quando uma determinada proteína é necessitada pela célula, a seqüência de nucleotídeos da parte apropriada da molécula de DNA é transcrita. De um modo geral, a transcrição consiste em "copiar" a parte requerida da seqüência de DNA em uma seqüência de RNA. Basicamente, o RNA é formado por apenas uma fita de nucleotídeos, sendo que ele possui a *ribose* como açúcar e suas bases nitrogenadas são *citosina, guanina, adenina* e *uracila*. Durante este processo, o pareamento dos nucleotídeos de RNA na cadeia de DNA, segue um padrão determinado. A *adenina* se pareia com *uracila*, e a *citosina* com *guanina*. A grande maioria dos genes presentes no DNA das células determina a seqüência de aminoácidos das proteínas, e as moléculas de RNA, que são copiadas a partir desses genes, são coletivamente chamadas de RNA *mensageiro*.

2.2.2 Tradução

Feita a transcrição, este RNA vai para o citoplasma da célula onde será incorporado pelo *ribossomo*. No *ribossomo* é feita a tradução que consiste na síntese de uma proteína a partir de informações contidas na molécula de RNA *mensageiro*. Neste processo, ocorre que a cada grupo de 3 bases nitrogenadas, que são chamadas de *códons,* se ligará um *anticódon*, que é uma trinca de bases que paream com os *códons*, presente num RNA que transporta aminoácidos. De uma forma geral, podemos dizer que os aminoácidos são codificados a partir de três bases nitrogenadas, ou *códons*, sendo que este sistema é conhecido como *código genético.* Por exemplo, o aminoácido *Methionine* é codificado por *atg*, ou seja, *adenina, timina* e *guanina*.

2.2.3 Aminoácidos e Bases Nitrogenadas

As proteínas são construídas a partir de um repertório de 20 aminoácidos que são as unidades básicas estruturais das proteínas. Como letras em um alfabeto, que ao se unirem formam palavras, proteínas são construídas a partir de uniões de aminoácidos.

As proteínas são funcionalmente diversas pelo fato de suas unidades básicas serem quimicamente diversas. Os aminoácidos podem ser classificados de acordo com suas características químicas em cinco grupos que são denominados como: *não polares, aromáticos, polares, carregados negativamente* e *carregados positivamente*. A Tabela 1 mostra esta classificação e mostra também os *códons* que codificam os respectivos aminoácidos.

Tabela 1: Classificação dos aminoácidos e código genético

Classificação	Aminoácido	Nomes Abreviados		Códons
Não Polares	Glycine	Gly	G	GGT GGC GGA GGG
	Alanine	Ala	A	GCT GCC GCA GCG
	Valine	Val	V	GTT GTC GTA GTG
	Leucine	Leu	L	TTA TTG CTT CTC CTA CTG
	Isoleucine	Ile	I	ATT ATC ATA
	Proline	Pro	P	CCT CCC CCA CCG
Aromáticos	Phenylalanine	Phe	F	TTT TTC
	Tyrosine	Tyr	Y	TAT TAC
	Tryptophan	Trp	W	TGG
Polares	Serine	Ser	S	TCT TCC TCA TCG AGT AGC
	Threonine	Thr	T	ACT ACC ACA ACG
	Cysteine	Cys	C	TGT TGC
	Methionine	Met	M	ATG
	Asparagine	Asn	N	AAT AAC
	Glutamine	Gln	Q	CAA CAG
Carregados -	Aspartane	Asp	D	GAT GAC
	Glutamate	Glu	E	GAA GAG
Carregados +	Lysine	Lys	K	AAA AAG
	Arginine	Arg	R	CGT CGC CGA CGG AGA AGG
	Histidine	His	H	CAT CAC

Fonte: Principles of Biochemistry

Observando a Tabela 1 temos que mais de um *códon* pode determinar o mesmo aminoácido, sendo que apenas os aminoácidos *Methionine* e *Tryptophan* são determinados por

um único *códon*.

Uma proteína pode então ser observada através dos aminoácidos que a forma e através das bases nitrogenadas que codificam estes aminoácidos. Como exemplo citamos a proteína g*alactin* encontrada na *Drosophila melanogaster*. Esta proteína quando observada através dos aminoácidos que a compõe é dada por **mlkltvllitllviaktgwtrekfsiqvneg-ntfgalvkaepftkindgyyffgteslnw**, onde cada letra, listada na Tabela 1, representa os aminoácidos que juntamente formam a g*alactin*. Quando esta mesma proteína é observada através das bases ela é, em parte, dada por: **atgctgaagcttacggttctactaattacattgctg-gtcatagcaaaaac**... . Temos que *a* representa *adenina*, *t* é *timina*, *g* representa *guanina* e *c* é *citosina*.

Analisar as seqüências de aminoácidos traz ricas informações a respeito da relação existente entre as proteínas. Uma seqüência de aminoácidos de interesse pode ser comparada com outras seqüências conhecidas para verificação de similaridades, ou seja, verificar se estas proteínas pertencem a uma mesma família e desta forma possuam funções similares. Citamos como exemplo as proteínas *mioglobina* e *hemoglobina* que pertencem a família *globina* e possuem a função de transportar oxigênio sendo que a primeira proteína transporta oxigênio nos músculos e a segunda no sangue. Outro exemplo de proteínas que pertencem a uma mesma família são as proteínas *chymotripisina* e a *tripsina* que são da família *protease serina*, que são tipos de proteínas chamadas de *enzimas*. Pode-se também comparar a mesma proteína em diferentes espécies para tentar obter informações evolucionárias entre elas. Segundo Brown (1999), a comparação da proteína *serum albumins* em primatas indicou que os homens e os macacos africanos divergem evolucionariamente entre si cerca de cinco milhões de anos e não trinta milhões de anos como se pensava anteriormente. Estas são algumas das informações que o estudo entre seqüências de aminoácidos e de bases que codificam estes aminoácidos pode trazer.

Referências

Neste capítulo, para abordarmos os temas mencionados foram utilizados Lehninger (1993), Stryer (1995), Brown (1999), e Alberts (2002) como referências bibliográficas.

Capítulo 3

Modelo multinomial

3.1 Introdução

Apresentamos neste capítulo alguns modelos probabilísticos que são utilizados, entre outras coisas, para descrever vários fenômenos ou situações encontradas na natureza. Estes modelos são expressos por distribuições de probabilidade que dependem de um ou mais parâmetros. Exibimos algumas destas distribuições considerando modelos em que as variáveis aleatórias que os expressam são discretas, ou seja, assumem um conjunto finito ou infinito enumerável de valores. Inicialmente, definimos o processo de Bernoulli para introduzirmos a distribuição binomial e em seguida, apresentamos a sua generalização para mais de dois eventos, ou seja, a distribuição multinomial.

3.2 Processo de Bernoulli

Um experimento no qual se observa um entre dois possíveis resultados, ou eventos, é chamado de ensaio de Bernoulli e um experimento consistindo de uma série de ensaios independentes e idênticos de Bernoulli é conhecido como processo de Bernoulli, Winkler e Hays (1975). Em geral, um dos dois eventos possíveis em um ensaio de Bernoulli é chamado de sucesso e o outro de fracasso. Denotamos por θ a probabilidade de sucesso e

portanto $1-\theta$ é a probabilidade de fracasso.

Um processo de Bernoulli é dito:

(a) **dicotômico**: pois se observa um entre dois possíveis resultados

(b) **independente**: o experimento consiste de uma série de ensaios de Bernoulli independentes

(c) **estacionário**: a probabilidade de sucesso, θ, não se altera de um ensaio para outro.

3.2.1 Exemplo

Suponha que uma moeda seja lançada n vezes. Desta maneira, para cada um destes n lançamentos, temos dois possíveis eventos, face da moeda cara ou face da moeda coroa. Seja θ a probabilidade da face da moeda ser cara, que chamamos aqui de sucesso, e $1-\theta$ a probabilidade da face da moeda ser coroa, que chamamos de fracasso. Seja X uma função que associa a cada resultado possível um número, X=1, correspondendo a cara e, X=0 correspondendo a coroa. Portanto temos que X é uma variável aleatória e sua distribuição de probabilidades é dada na Tabela 2.

Tabela 2: Distribuição de probabilidade

Eventos	x	P(X=x)
cara	1	θ
coroa	0	$1-\theta$

Considere que esta moeda seja lançada 3 vezes e que haja o interesse em se encontrar a probabilidade da seguinte seqüência de eventos (cara,cara,coroa). A probabilidade de que a primeira observação seja cara é θ, ou P(X=1) =θ. Por definição, temos que se A e B são dois eventos quaisquer independentes, então P(A$\cap$B)=P(A)P(B). Como o segundo lançamento da moeda é independente do primeiro lançamento, a segunda observação é independente da primeira e desta forma, temos que P(cara $\cap$ cara) = P(cara)P(cara)

$= \theta\theta = \theta^2$.

A sureedeldgdgh gh txh d sulp hlud h d vhgxqgd revhuydçõhv vhmdp fdud h d whufhlud revhuydçãr vhmdl frurd é gdgr sru P(fdud∩fdud∩frurd,=P(fdud,P(fdud,P(frurd,= $\theta^2(1-\theta)$.

Sh wyévvhp rv lqwhuhvvh qd sureedeldgdgh gd vht üêqfld grv hyhqwr v (frurd,fdud,fdud,, whuídp rv txh hvwd sureedeldgdgh vheld gdgd sru P(frurd∩fdud∩fdud, = $\theta^2(1-\theta)$, rx vhmdl, lqghs hqghqwh gd rughp hp txh ds duhfhp dv idfhv gd p rhgd d sureedeldgdgh sdud r hyhqwr whqgr gxdv fdudv h xp d frurd vheá vhp suh gdgd sru $\theta^2(1-\theta)$.

I rup ddzdqgr rv uhvxowdgrv ylwrv qr hxhp sor 3.2.1 whp rv txh, hp xp surfhvvr gh Bhuqrxod, d sureedeldgdgh gh txdotxhu vht üêqfld gh vxfhvvrv h iudfdvvrv hp n lqghs hqghqwhv hqvdlrv gh Bhuqrxod ghs hqgh gr qúp hur gh vxfhvvrv, gr qúp hur gh iudfdvvrv h gd sureedeldgdgh θ d hch dvvrfldgd vhqgr hvwd sureedeldgdgh gdgd sru

$$\theta^x(1-\theta)^{n-x}, \qquad (3.1,$$

rqgh x uhsuhvhqwd r qúp hur gh vxfhvvrv.

3.3 Distribuição binomial

Nd vhçãr dqwhulru whp rv r fdvr rqgh r remhwlyr é ghwhup lqdu d sureedeldgdgh gh xp d vht üêqfld gh n hqvdlrv gh Bhuqrxod frp x vxfhvvrv h $n-x$ iudfdvvrv. Ep p xlwrv fdvrv, sruép , r lqwhuhvvh qãr hvwá qd sureedeldgdgh gh xp d sduwlfxodu vht üêqfld, p dv vlp , qd sureedeldgdgh gr qúp hur gh vxfhvvrv suhvhqwhv, lqghs hqghqwhp hqwh gd rughp frp txh dv revhuydçõhv hvwãr glvs rvwdv qhvwd vht üêqfld.

3.3.1 Exemplo

Sxs rqkd txh xp d p rhgd vhmdl odlqçdgd 5 yhzhv. Crqvlghuh txh 1 uhsuhvhqwh d idfh gd p rhgd *cara* h 0 uhsuhvhqwh d idfh gd p rhgd *coroa*. Dhvwd irup d whuídp rv xp wrwdo gh $2^5 = 32$ vht üêqfldv gh 0'v h 1'v s rvvíyhlv frp r sru hxhp sor, 00001, 01011, 11110,... .

Crqwxgr, vxs rqkd txh r qrvvr lqwhuhvvh qãr hvwhmd hp wrgrv rv hyhqwrv h vlp qrv hyhqwrv rqgh ds duhçdp 2 fdudv frp r p rvwad d Wdehod 3.

Wdehod 3: Eyhqwrv s rvvíyhlv frp 2 fdudv

Eyhqwr 1: $(1,1,0,0,0)$,
Eyhqwr 2: $(1,0,1,0,0)$,
Eyhqwr 3: $(1,0,0,1,0)$,
Eyhqwr 4: $(1,0,0,0,1)$,
Eyhqwr 5: $(0,1,0,0,1)$,
Eyhqwr 6: $(0,1,1,0,0)$,
Eyhqwr 7: $(0,1,0,1,0)$,
Eyhqwr 8: $(0,0,1,1,0)$,
Eyhqwr 9: $(0,0,1,0,1)$ h
Eyhqwr 10: $(0,0,0,1,1)$.

Dhvwd p dqhlud whuhp rv $\binom{5}{2} = \frac{5!}{2!3!} = 10$ hyhqwrv s rvvíyhlv. Pru ghfiqlçãr, vh grlv hyhqwrv A h B vãr p xwxdp hqwh hxfoxghqwhv, hqwãr P(A∪B,=P(A,. P(B,. Crp r ylp rv hp (3.1,, s dud fdgd xp grv hyhqwrv gd Wdehod 3, d s urededldgdgh gd vht üêqfld frp hxdwdp hqwh 2 fdudv é lgxdo d $\theta^2(1-\theta)^3$. Crp r rv hyhqwrv 1 d 10 vãr p xwxdp hqwh hxfoxghqwhv, whp rv txh d s urededldgdgh gh 2 vxfhvvr v hp 5 odqçdp hqwr v é gdgr s ru $\theta^2(1-\theta)^3 + \theta^2(1-\theta)^3 + \ldots + \theta^2(1-\theta)^3 = \binom{5}{2}\theta^2(1-\theta)^3$.

Ep dp r vwadv rulxqgdv gh xp s urfhvvr gh Bhuqrxod frp s urededldgdgh gh vxfhvvr θ, d s urededldgdgh gh vh re vhuydu hxdwdp hqwh X vxfhvvr v hp n hqvdlr v lqghs hqghqwhv vhgxh xp d glvwulexlçãr elqr p ldo txh, dqddwfdp hqwh é ghvfulwd frp r

$$P(X = x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad (3.2,$$

s dud $0 \leq x \leq n$.

Ep xp d glvwulexlçãr elqrp ldo, r qúp hur hvshudgr gh vxfhvvrvs dud d yduláyhoddhdwóuld X é gdgr s ru

$$E(X) = n\theta, \tag{3.3,}$$

h d ydulâqfld é gdgd s ru

$$V(X) = n\theta(1-\theta). \tag{3.4,}$$

Nhvvh hxhp sor, fdvr d p rhgd vhmd *não viciada*, vh $n = 5$ whuhp rv $E(X) = 2,5$ h $V(X) = 1,25$, rqgh X uhsuhvhqwd r qúp hur gh fdudvh n é lgxdodr qúp hur gh odqçdp hqwrv.

3.3.2 Exemplo

Nd dqkd gh surgxçãr gh xp d iáeulfd, hp frqglçõhvqrup dlv gh ixqflrqdp hqwr fdgd xp d gdvshçdvsrgh vhu frqvlghudgd frp r surgxzlgd lqghshqghqwh gdvghp dlv. Sh uhwudup rv xp d dp rwwd, frp uhsrvlçãr, gh n shçdvgd dqkd gh surgxçãr h fkdp dup rv gh θ d iudçãr gh shçdv ghihlwxrvdv txh vãr surgxzlgdv, hqwãr X, r qúp hur gh shçdv ghihlwxrvdv qd dp rwwd, é xp d yduláyhoddhdwóuld frp glvwulexlçãr elqrp ldofrp sduâp hwur v n h θ.

Sxsrqkd txh whqkdp rvd lqirup dçãr gh txh xp d iudçãr gh 0,2 gdvshçdvvãr surgxzlgdv frp xp ghihlwr hvshflfifr. Ar frchwdup rv 5 shçdvfrp uhsrvlçãr d sureded ldgdgh gh txh whqkdp rv 2 shçdvghihlwxrvdv é fdcfxodgd frp r

$$P(X = 2|n = 5, \theta = 0,2) = \binom{5}{2}(0,2)^2(0,8)^{5-2} = 0,2.$$

Awudyév gh (3.3, h (3.4, whp rvtxh dr frchwdup rv 5 shçdvddhdwruldp hqwh frp uhsrvlçãr whuídp rvhp p égld

$$E(X) = 5 \times 0,2 = 1,$$

rx vhmd, 1 shçd ghihlwxrvd frp xp d ydulâqfld gh

$$V(X) = 5 \times 0,2 \times 0,8 = 0,8.$$

3.4 Distribuição multinomial

A distribuição binomial pode ser generalizada para situações onde ocorram mais de dois eventos. Esta generalização é conhecida como distribuição multinomial.

Formalmente, para definirmos a distribuição multinomial, considere k eventos mutuamente exclusivos com probabilidades $\theta_1, \theta_2, ..., \theta_k$ de ocorrerem. Por definição, dois eventos A e B são mutuamente exclusivos se A$\cap$B=$\emptyset$. Se n observações são feitas de forma independente e aleatória, então a probabilidade de que ocorra exatamente x_1 eventos 1, x_2 eventos 2, ..., x_k eventos k segue uma distribuição multinomial que analiticamente é descrita como

$$P(X = (x_1, x_2, ..., x_k)|n, \theta = (\theta_1, \theta_2, ..., \theta_k)) = \frac{n!}{x_1!x_2!...x_k!}\theta_1^{x_1}\theta_2^{x_2}...\theta_k^{x_k} = n!\prod_{l=1}^{k}\frac{\theta_l^{x_l}}{x_l!}, \quad (3.5)$$

onde $\sum_{l=1}^{k} x_l = n$, e $\sum_{l}^{k} \theta_l = 1$.

A seguir temos algumas propriedades da distribuição multinomial.

Em n experimentos o número esperado para o componente l o vetor aleatório $X = (X_1, X_2, ..., X_k)$, para $l = 1, 2, ..., k$, é dado por

$$E(X_l) = n\theta_l, \quad (3.6)$$

e a variância associada à componente l do vetor aleatório $X = (X_1, X_2, ..., X_k)$, para $l = 1, 2, ..., k$, é dado por

$$V(X_l) = n\theta_l(1 - \theta_l). \quad (3.7)$$

A covariância entre as componentes de f é dada por

$$Cov(\mathrm{f}_{l}\mathrm{cf}_{j}) = -n\mathrm{w}_{l}\mathrm{w}_{j}\mathrm{c}\ \text{para}\ , \neq j;\ ,\mathrm{c}j = 1\mathrm{c}2\mathrm{c}...\mathrm{c}k\ \ \mathrm{e}, \tag{3.8}$$

a correlação linear é dada por

$$1_{lj} = \frac{Cov(\mathrm{f}_{l}\mathrm{cf}_{j})}{\sqrt{\mathrm{T}(\mathrm{f}_{l})\mathrm{T}(\mathrm{f}_{j})}}\ \text{para}\ i \neq j;\ ,\mathrm{c}j = 1\mathrm{c}2\mathrm{c}...k \tag{3.9}$$

onde 1_{lj} é o coeficiente de correlação linear entre f_{l} e f_{j}.

Observe o exemplo ilustrando as informações dadas acima.

3.4.1 Exemplo

Através de uma pesquisa feita em uma cidade a respeito da preferência popular sobre marcas de café sabe-se que 80% dos consumidores preferem a marca de café A, 18% dos consumidores preferem a marca de café B, 1% preferem a marca de café C e 1% dos consumidores preferem a marca de café D.

Suponha que se queira saber qual é a probabilidade em uma amostra aleatória e com reposição de 11 consumidores de que 5 consumidores irão preferir a marca de café A, 3 irão preferir a marca B, 2 irão preferir a marca C e, 1 consumidor irá preferir a marca D.

Temos que a distribuição do vetor aleatório $\mathrm{f} = (\mathrm{f}_{1}\mathrm{cf}_{2}\mathrm{cf}_{3}\mathrm{cf}_{4})\mathrm{c}$ segue uma distribuição multinomial com parâmetro n =11 e w= $(\mathrm{w}_{1} = 0\mathrm{c}80\mathrm{cw}_{2} = 0\mathrm{c}18\mathrm{cw}_{3} = 0\mathrm{c}01\mathrm{cw}_{4} = 0\mathrm{c}01)$ e através de (3.5) temos $P(\mathrm{f} = (\mathrm{f}_{1}\mathrm{cf}_{2}\mathrm{cf}_{3}\mathrm{cf}_{4})|n = 11\mathrm{cw} = (\mathrm{w}_{1} = 0\mathrm{c}80\mathrm{cw}_{2} = 0\mathrm{c}18\mathrm{cw}_{3} = 0\mathrm{c}01\mathrm{cw}_{4} = 0\mathrm{c}01)) = 0\mathrm{c}000053$

Através de (3.6) e (3.7), caso tivéssemos uma amostra de 100 consumidores teríamos

$$.\ (\mathrm{f}_{1}) = 100(0\mathrm{c}80) = 80\mathrm{c}$$

ou seja, 80 consumidores em média preferindo a marca A sendo que a variância de f_{1} é dada por

$$\mathrm{T}(\mathrm{f}_{1}) = 100(0\mathrm{c}80)(0\mathrm{c}20) = 16\mathrm{c}$$

teríamos também uma esperança para f_{2} igual a

$$.\ (\mathrm{f}_{2}) = 100(0\mathrm{c}18) = 18\mathrm{c}$$

e uma variância para f_{2} igual a

$$100(0\mathrm{c}18)(0\mathrm{c}82) = 14\mathrm{c}76 \cong 15\mathrm{c}$$

e assim por diante.

Na Tabela 4, estão exibidas as covariâncias entre as componentes do vetor aleatório f , calculadas através de (3.8).

Tabela 4: Covariância entre as componentes do vetor aleatório f

Componentes de f	X_1	X_2	X_3	X_4
X_1	-	-14,40	-0,80	-0,80
X_2	-14,40	-	-0,18	-0,18
X_3	-0,80	-0,18	-	-0,01
X_4	-0,80	-0,18	-0,01	-

Observando a Tabela 4 temos que entre as componentes do vetor aleatório f , para f_{1} e f_{3} existe uma covariância de -0,80, já entre f_{2} e f_{4} existe uma covariância de -0,18.

Através de (3.9) temos que a correlação entre as componentes do vetor aleatório f_{1}, f_{1} e f_{2} é dada por $1_{12} = \frac{-14,40}{\sqrt{(16)(15)}} = -0\mathrm{c}92\mathrm{c}$ e entre f_{2} e f_{4} temos um coeficiente de correlação $1_{24} = \frac{-0,18}{\sqrt{(15)(1)}} = -0\mathrm{c}04$.

Referências

Neste capítulo, para abordarmos os temas mencionados foram utilizados Winkler, *et al.* (1975), e Casella *et al.* (1990) como referências bibliográficas.

Capítulo 4

Modelo com mistura

4.1 Introdução

No primeiro capítulo apresentamos exemplos gerais de aplicações de modelos com distribuições mistas sem nos atermos a questões teóricas que formulam estas distribuições. Neste capítulo descrevemos o modelo com distribuições mistas, apresentamos a função de verossimilhança dos parâmetros e exibimos uma maneira para estes parâmetros serem estimados utilizando uma abordagem bayesiana. Em seguida, mostramos uma estratégia para se construir uma distribuição *a priori* informativa partindo de informações disponíveis.

4.2 Fundamentos

Os modelos estatísticos formulados partindo de misturas de distribuições são aplicados quando as observações são heterogêneas, ou seja, quando as observações são provenientes de uma população particionada em vários grupos ou componentes, sendo que, não se sabe à qual destas componentes pertence cada observação em particular. Desta forma os modelos com mistura apresentam características que expressam tanto a heterogeneidade das observações quanto a pluralidade expressa pelo número de componentes que particionam a população.

Para definirmos um modelo com mistura de distribuições, considere um vetor aleatório f = (f 1cf 2c...cf k) assumindo valores no conjunto dos números reais R^k. Se a distribuição de f cpode ser representada por uma função de densidade de probabilidade, no caso do vetor f assumir valores contínuos, ou função de probabilidade no caso do vetor assumir valores discretos, da forma

$$\mathrm{s}(x) = p_1\mathrm{s}_1(x) + p_2\mathrm{s}_2(x) + ... + p_M\mathrm{s}_M(x)\mathrm{c} \tag{4.1}$$

onde $\sum_{j=1}^{M} p_j = 1\mathrm{c}p_j \ 7 \ 0\mathrm{c}\,\mathrm{s}_j(.) \geq 0$ e $\int_R \mathrm{s}_j(x)dx = 1$ com $j = 1\mathrm{c}2\mathrm{c}...\mathrm{c}'$, então podemos afirmar que f = (f 1cf 2c...cf k) possui uma distribuição com mistura sendo que s (.), definida em (4.1) é uma função densidade com mistura finita.

Por convenção temos que em (4.1) os parâmetro $p_1\mathrm{c}p_2\mathrm{c}...\mathrm{c}p_M\mathrm{c}$ são chamados de proporções da mistura e $\mathrm{s}_1(.)\mathrm{c}\mathrm{s}_2(.)\mathrm{c}...\mathrm{c}\mathrm{s}_M(.)$ são chamadas densidades componentes da mistura e representam quaisquer distribuição.

Note que (4.1) define uma função densidade de probabilidade pois

$$\mathrm{s}(x) \geq 0\mathrm{c}$$

pois $p_j \ 7 \ 0$ e $\mathrm{s}_j(x) \geq 0$ para $j = 1\mathrm{c}2\mathrm{c}...\mathrm{c}'$ ce

$$\int_R \mathrm{s}(x)dx = 1.$$

Considerando $\mathrm{s}_j(x) = \mathrm{s}_j(x|\mathrm{w}_j)$ para $j = 1\mathrm{c}2\mathrm{c}...\mathrm{c}'$ cou seja, as densidades componentes da mistura $\mathrm{s}_j(x)$ pertencendo a uma mesma família, temos que a função de densidade com mistura finita é expressa como

$$\mathrm{s}(x|\mathrm{wc}p) = \sum_{j=1}^{M} p_j\mathrm{s}(x|\mathrm{w}_j)\mathrm{c} \tag{4.2}$$

e podemos dizer que em (4.2) a função densidade é expressa como uma soma ponderada das j densidades componentes sendo que, esta ponderação é feita pelas proporções da

mistura p_j, onde $\sum_{j=1}^{M} p_j = 1$. Assim, podemos pensar nesta ponderação feita por p_j como sendo a probabilidade associada a cada componente da mistura de produzir as observações f $= (\mathrm{f}_{1}\mathrm{cf}_{2}\mathrm{c...cf}_{k})$.

Neste trabalho, as observações são as moléculas que compõem as proteínas, ou seja, os aminoácidos e as bases que codificam estes aminoácidos. As componentes que particionam a população são as famílias de proteínas chamadas também de grupos protéicos.

4.3 Função de verossimilhança

Procuramos verificar como a modelagem estatística com mistura de distribuições desempenha o papel de identificar e separar proteínas semelhantes observando-as através dos aminoácidos que as constitui e também através das bases nitrogenadas que codificam estes aminoácidos. Para observarmos uma proteína, verificamos e contamos o número de vezes que cada molécula aparece na seqüência que forma a proteína. Observe por exemplo 2 proteínas sendo vistas através da seqüência de bases nitrogenadas como

proteína 1: **atggtgcacctgactcctga** e proteína 2: **aacctcaagggcacctttgc,**

neste caso, verificamos o número de vezes que apareceram as bases adenina (**a**), timina (**t**), citosina (**c**) e guanina (**g**) e desta forma temos o vetor de observações como mostra a Tabela 5.

Tabela 5: Proporções para cada base nitrogenada

	adenina	timina	citosina	guanina
proteína 1	4	5	6	5
proteína 2	5	4	7	4

Temos então o seguinte vetor de observações: f $= [(4\mathrm{c}5\mathrm{c}6\mathrm{c}5)\mathrm{c}(5\mathrm{c}4\mathrm{c}7\mathrm{c}4)]$

Dado que o vetor f $_i = (x_{i1}\mathrm{c}x_{i2}\mathrm{c...c}x_{ik})\mathrm{c}$ para $i = 1\mathrm{c}2\mathrm{c...c}n$, foi observado a função de verossimilhança de we p, construída a partir do modelo (4.2), é dada por

$$\mathrm{u}\,(\mathrm{wc}p|x) = \prod_{i=1}^{n} \sum_{j=1}^{M} p_j \mathrm{s}\,(x_i|\mathrm{w}_j), \tag{4.3}$$

e pode ser pensada como uma função dos parâmetros quando os dados estão fixos.

4.3.1 Variáveis latentes

Segundo Gelman *et al.* (1995), o procedimento básico a ser feito quando se trabalha com mistura de distribuições é utilizar um algoritmo baseado em uma amostra ampliada que tem por objetivo classificar as observações em relação às componentes. Para ampliar a amostra, agrega-se ao vetor de observações f $_i = (x_{i1}\mathrm{c}x_{i2}\mathrm{c}...\mathrm{c}x_{ik})\mathrm{c}$ para $i = 1\mathrm{c}2\mathrm{c}...\mathrm{c}n\mathrm{c}$ vetores $Z_i = (Z_{i1}\mathrm{c}Z_{i2}\mathrm{c}...\mathrm{c}Z_{iM})$ de variáveis aleatórias indicadoras não observáveis, chamadas de variáveis latentes. Estas variáveis se comportam da seguinte forma, $Z_{ij} = 1$ se a i-ésima observação é proveniente da j-ésima componente da mistura e $Z_{ij} = 0$ caso contrário observando a seguinte restrição, $\sum_{j=1}^{M} Z_{ij} = 1$, que faz com que o vetor Z_i indique que a observação f $_i$ seja atribuída a somente uma das componentes da mistura.

Temos então que o vetor Z_i segue uma distribuição multinomial

$$Z_i|x\,\mathrm{cwc}p \sim \text{'} \quad u,tino3\ i{=}(1; q_i = (q_{i1}\mathrm{c}q_{i2}\mathrm{c}...\mathrm{c}q_{iM}))\mathrm{c} \tag{4.4}$$

sendo que segundo Gilks *et al.* (1996), cada componente q_i, para $i = 1\mathrm{c}2\mathrm{c}...\mathrm{c}n$, em (4.5) é

$$q_{ij} = \frac{p_j \mathrm{s}\,(x_i|\mathrm{w}_j)}{\sum_{r=1}^{M} p_r \mathrm{s}\,(x_i|\mathrm{w}_r)}\mathrm{c} \tag{4.5}$$

para $i = 1\mathrm{c}2\mathrm{c}...\mathrm{c}n$ e $j = 1\mathrm{c}2\mathrm{c}...\mathrm{c}$' .

Segundo Diebolt *et al.* (1994), a classificação feita das variáveis observáveis em relação às componentes com a inclusão das variáveis latentes, atua como uma estrutura oculta do modelo que pode ser visto como dados perdidos. A inclusão das variáveis latentes Z_i, para $i = 1\mathrm{c}2\mathrm{c}...\mathrm{c}n$, além de classificar as observações contribui para simplificação da função de verossimilhança, pois após a inclusão de Z_i a função de verossimilhança(4.3) passa a ser

descrita como

$$\mathrm{u}\left(\mathrm{wc}p|x\mathrm{c}2\right) = \prod_{j=1}^{M}\prod_{i=1}^{n}[p_j\mathrm{s}\left(x_i|\mathrm{w}_j\right)]^{Z_{ij}}. \tag{4.6}$$

Vamos abordar dois aspectos relacionados aos parâmetros de interesse no modelo (4.6).

O primeiro aspecto diz respeito a quantidade de componentes que particionam a população. Quando existe a informação a respeito da quantidade de componentes então não existe problema em relação a dimensionalidade do espaço paramétrico, mas quando o número ' de componentes é desconhecido, devemos utilizar procedimentos como por exemplo, técnicas de seleção de modelos para estimar o valor de ' . Neste trabalho consideramos o número de componentes desconhecido. Inicialmente propomos o método considerando o número de componentes ' conhecido, posteriormente estimamos o valor de ' mais adequado utilizando métodos de seleção de modelos.

O segundo aspecto é referente à questão da estimação dos parâmetros desconhecidos p e w. Uma possível abordagem é a utilização de um processo via máxima verossimilhança para obter as estimativas dos parâmetros. Neste tipo de abordagem, o objetivo é encontrar valores das estimativas dos parâmetros que maximizem a função de verossimilhança. Freqüentemente, por facilidade analítica, tenta-se maximizar o logaritmo da verossimilhança, ou seja, tenta-se maximizar o $\log\left[\mathrm{u}\left(\mathrm{wc}p|x\mathrm{c}2\right)\right]$. Dependendo da forma de $\mathrm{u}\left(\mathrm{wc}p|x\mathrm{c}2\right)$ o problema de maximização pode se tornar bastante complexo exigindo técnicas mais elaboradas. Uma destas técnicas é o algoritmo *EM, (Expectation-Maximization).* O algoritmo *EM* é um método para encontrar estimadores de máxima verossimilhança dos parâmetros de modelos que possuem uma estrutura com variáveis latentes ou dados *missing.*

4.4 Abordagem bayesiana

Uma alternativa é estimar os parâmetro através de métodos *MCMC, (Markov Chain Monte Carlo)*, utilizando uma abordagem bayesiana.

O procedimento padrão para modelagem bayesiana é feito combinando uma infor-

mação prévia sobre os parâmetro de interesse, chamada de distribuição *a priori,* à uma função de verossimilhança produzindo desta forma, uma distribuição *a posteriori.* O mecanismo usado para combinar estas informações é dado por um resultado que fornece o relacionamento entre várias probabilidades condicionais chamada de *teorema de Bayes.* A versão simples do teorema de Bayes para dois eventos A e B pertencentes a um espaço amostral R é dado como

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\overline{A})P(B|\overline{A})} \tag{4.7}$$

onde $\overline{A}$ representa o complemento de A, ou seja, são todos os pontos amostrais que estão em R e não estão em A.

Em (4.7) o *teorema de Bayes* é visto em termos de eventos mas é possível interpretá-lo em termos de distribuições condicionais de variáveis aleatórias discretas ou contínuas.

Suponha que haja interesse em fazer inferências a respeito de um parâmetro w que assume apenas a possíveis valores como $w_1, w_2, \ldots, w_J$ e que as informações a respeito de w possam ser sumarizadas através de uma distribuição de probabilidades $P(w = w_j)$. Esta distribuição de probabilidades é chamada de distribuição *a priori* de we, através do *teorema de Bayes* (4.7), combinamos esta distribuição com a função de verossimilhança $u(w_j|x)$ obtendo a distribuição *a posteriori* como

$$P(w_j|x) = \frac{P(w_j)u(w_j|x)}{\sum_{r=1}^{J} P(w_r)u(w_r|x)}. \tag{4.8}$$

Se o parâmetro de interesse w assume valores contínuos as distribuições *a priori* e *a posteriori* são representadas por funções de densidades e o *teorema de Bayes* é descrito como

$$s(w|x) = \frac{s(w)u(w|x)}{\int s(w)u(w|x)dw}. \tag{4.9}$$

Utilizando esta abordagem podemos estimar os parâmetros através dos métodos de

simulação baseados em cadeias de *Markov* chamados de *MCMC.* A idéia básica destes métodos é obter amostras da distribuição *a posteriori.* Estas amostras são obtidas por meio das distribuições *a posteriori* condicionais utilizando cadeias de *Markov,* cuja distribuição estacionária é a distribuição *a posteriori* de interesse.

Metropolis *et al* (1953) desenvolveram um algoritmo que atendia esta idéia básica, sendo que, posteriormente, foi generalizado por Hastings (1970) e desta forma este algoritmo passou a ser conhecido na literatura como algoritmo *Metropolis-Hastings.* A proposta do algoritmo *Metropolis-Hastings* é atingir a distribuição estacionária através de simulações sucessivas de uma distribuição conveniente, aceitando ou rejeitando o valor gerado segundo uma probabilidade específica. Um caso particular de *Metropolis-Hastings* é o algoritmo *Gibbs-Sampling.* Neste algoritmo, a probabilidade de aceitação do valor gerado é 1 e a distribuição estacionária é atingida através de simulações sucessivas feitas diretamente das distribuições *a posteriori* condicionais.

Utilizamos nesta dissertação uma abordagem bayesiana e para estimação dos parâmetros do modelo o algoritmo *Gibbs-Sampling.*

4.5 Algoritmo *Gibbs-Sampling*

O algoritmo *Gibbs-Sampling* foi proposto inicialmente por Geman e Geman (1984) com o intuito de utilizá-lo em problemas de reconstrução de imagens. Foram Gelfand e Smith (1990) que mostraram como o algoritmo poderia ser utilizado para obter amostras de distribuições *a posteriori* e como conseqüentemente ser usado para resolver problemas inerentes à inferência bayesiana.

Segundo Besag (1974) este algoritmo é baseado no fato de que se a distribuição condicional $P(w|x)$ for positiva em $\Theta_1 \times \Theta_2 \times ... \times \Theta_M$c onde Θ_j representa o espaço paramétrico da distribuição de w_j para $j = 1c2c...c'$, então ela é unicamente determinada pelas distribuições condicionais completas $P(w_j|xcw_{(-j)})$, $j = 1c2c...c'$, onde w= $(w_1cw_2c...cw_M)$ e $w_{(-j)}$ representa o vetor wsem a j-ésima componente, isto é, $w_{(-j)} =$

$(w_1 c...cw_{j-1} cw_{j+1} c...cw_M)$. Desta forma pode-se dizer que o algoritmo é um esquema markoviano que requer a amostragem destas distribuições condicionais.

Descrição do algoritmo

Considere o vetor de parâmetros w= $(w_1 cw_2 c...cw_M)$ e as observações f $= (f_1 cf_2 c...cf_k)$. Considere também que se deseja estimar os parâmetros de uma distribuição conjunta *a posteriori* dada por $P(w|x)$ e que ' seja conhecido.

Passo 1:

Seja o vetor $w^{(0)} = (w_1^{(0)} cw_2^{(0)} c...cw_M^{(0)})$ de valores iniciais tais que $P(w^{(0)}|x)$ 7 0.

Passo 2:

Gerar $w_1^{(1)}$ de $P(w_1|x cw_2^{(0)} cw_3^{(0)} cw_4^{(0)} cw_5^{(0)} c...cw_M^{(0)})c$

gerar $w_2^{(1)}$ de $P(w_2|x cw_1^{(1)} cw_3^{(0)} cw_4^{(0)} cw_5^{(0)} c...cw_M^{(0)})c$

gerar $w_3^{(1)}$ de $P(w_3|x cw_1^{(1)} cw_2^{(1)} cw_4^{(0)} cw_5^{(0)} c...cw_M^{(0)})c$

gerar $w_4^{(1)}$ de $P(w_4|x cw_1^{(1)} cw_2^{(1)} cw_3^{(1)} cw_5^{(0)} c...cw_M^{(0)})c$

$\vdots$

gerar $w_M^{(1)}$ de $P(w_M|x cw_1^{(1)} cw_2^{(1)} cw_3^{(1)} cw_4^{(1)} c...cw_{M-1}^{(1)})$.

Ao se completar a primeira iteração se completa a transição de $w^{(0)} = (w_1^{(0)} cw_2^{(0)} c...cw_M^{(0)})$ para $w^{(1)} = (w_1^{(1)} cw_2^{(1)} c...cw_M^{(1)})$.

O esquema anterior é repetido com $w^{(1)}$ anteriormente obtido, como um vetor inicial e obtêm-se um novo vetor $w^{(2)} = (w_1^{(2)} cw_2^{(2)} c...cw_M^{(2)})$.

Repete-se o **Passo 2** A vezes descartando as 3 iterações iniciais chamadas de *burn in*, produzindo assim um vetor $w^{(0)} cw^{(1)} cw^{(2)} c...cw^{(T)}$.

A sucessão $w^{(0)} cw^{(1)} cw^{(2)} c...cw^{(T)}$ é uma realização de uma cadeia de *Markov* sendo que quando T tende ao infinito, A $\longrightarrow \infty$, o vetor $(w_1^{(T)} cw_2^{(T)} c...cw_M^{(T)})$ converge em distribuição para um vetor aleatório cuja função de densidade conjunta é $P(w|x)$.

Em particular, $w_j^{(T)} c$ converge em distribuição para uma quantidade aleatória cuja densidade é $P(w_j|x)$ que é a densidade marginal *a posteriori* de w_j para $j = 1c2c...c'$ e

$$\frac{1}{A}\sum_{i=1}^{T} z(w^{(i)}) \longrightarrow . \ [z(w)|x] \ \text{(quase certamente)}$$

quando $A \longrightarrow \infty$ para qualquer função $z(.)$ onde . $[z(w)|x]$ representa o valor esperado de $z(w)$ em relação à distribuição *a posteriori* $P(w|x)$.

Uma questão relevante a ser abordada diz respeito a convergência da cadeia de *Markov* para o estado de equilíbrio. Na literatura existem vários métodos propostos para verificar esta convergência. Entre eles citamos Gelfand *et al.* (1990), Geweke (1992), Gelman e Rubin (1992), Raftery e Lewis (1992) e Ritter e Tanner (1992). No entanto, segundo Cowles e Carlin (1996) não há um método que se possa dizer ser o melhor ou o mais eficiente que os outros.

Nesta dissertação o método utilizado para verificar a convergência da cadeia de *Markov* utilizando o algoritmo *Gibbs-Sampling* é o de Gelman e Rubin (1992). Neste método trabalhamos com múltiplas cadeias partindo de pontos iniciais dispersos e, comparamos a variância dentro das seqüências com a variância entre as seqüências. No caso de convergência, esperamos que a variância dentro das cadeias seja muito semelhante e que a variância entre as cadeias seja muito pequena.

4.5.1 Algoritmo *Gibbs Sampling* com variáveis latentes

Nesta seção apresentamos o algoritmo *Gibbs Sampling* para modelos com mistura que possuem a estrutura de verossimilhança exibida em (4.6). Este algoritmo foi proposto por Diebolt e Robert (1994), com o objetivo de estimar os parâmetros de interesse.

Seja $w = (w_1, w_2, ..., w_M)$ e $p = (p_1, p_2, ..., p_M)$ os parâmetros do modelo e as observações $f = (f_1, f_2, ..., f_k)$, considere $w^{(0)} = (w_1^{(0)}, w_2^{(0)}, ..., w_M^{(0)})$ e $p^{(0)} = (p_1^{(0)}, p_2^{(0)}, ..., p_M^{(0)})$ vetores de valores iniciais tais que $P(w^{(0)}|x, p) > 0$ e $P(p^{(0)}|x, w) > 0$ e que $\ldots$ seja conhecido.

Faça $t = 0$.

Passo 1:

Gerar $Z_i^{(t+1)}$ de uma distribuição multinomial com parâmetros $(1, q_i^{(t)})$ onde q_i é dado

(4.5) para $i = 1, 2, ..., n$.

Passo 2:

Gerar $w_1^{(t+1)}$ de $P(w_1|x, w_2^{(t)}, w_3^{(t)}, w_4^{(t)}, ..., w_M^{(t)}, Z_i^{(t+1)}, p_1^{(t)}, p_2^{(t)}, ..., p_M^{(t)})$,

gerar $p_1^{(t+1)}$ de $P(p_1|x, w_1^{(t+1)}, w_2^{(t)}, w_3^{(t)}, w_4^{(t)}, ..., w_M^{(t)}, Z_i^{(t+1)}, p_2^{(t)}, ..., p_M^{(t)})$,

gerar $w_2^{(t+1)}$ de $P(w_2|x, w_1^{(t+1)}, w_3^{(t)}, w_4^{(t)}, ..., w_M^{(t)}, Z_i^{(t+1)}, p_1^{(t+1)}, p_2^{(t)}, ..., p_M^{(t)})$,

$\vdots$

gerar $w_M^{(t+1)}$ de $P(w_M|x, w_1^{(t+1)}, w_2^{(t+1)}, w_3^{(t+1)}, ..., w_{M-1}^{(t+1)}, Z_i^{(t+1)}, p_1^{(t+1)}, p_2^{(t+1)}, ..., p_M^{(t)})$,

gerar $p_M^{(t+1)}$ de $P(w_M|x, w_1^{(t+1)}, w_2^{(t+1)}, w_3^{(t+1)}, ..., w_M^{(t+1)}, Z_i^{(t+1)}, p_1^{(t+1)}, p_2^{(t+1)}, ..., p_{M-1}^{(t+1)})$

Faça $t = t + 1$.

Após um grande número de iterações dos passos 1 e 2, descartadas as 3 iterações iniciais e verificado a convergência da cadeia, os vetores gerados

$$(w_1^{(t+1)}, ..., w_M^{(t+1)}) \text{ e } (p_1^{(t+1)}, ..., p_M^{(t+1)})$$

são gerados da distribuição desejada.

4.5.2 Algoritmo *Gibbs Sampling* para mistura de multinomiais

Nesta seção apresentamos o algoritmo *Gibbs-Sampling* sendo aplicado ao modelo com mistura de multinomiais empregado nesta dissertação. Substituindo em (4.6) $s(x_i|w_j)$ pelo modelo multinomial com parâmetro N e $w_j = (w_{j1}, ..., w_{jk})$ para $j = 1, 2, ..., '$ componentes,

$$u(w, p|x, 2) = \prod_{j=1}^{M} \prod_{i=1}^{n} [p_j s(x_i|w_j)]^{Z_{ij}}$$

considerando N conhecido, passamos a ter a função de verossimilhança de w e p baseada em uma amostra ampliada sendo dada por

$$u(w, p|x, 2) \propto \prod_{i=1}^{n} \left[(p_1(w_{11}^{x_{i1}} ... w_{1k}^{x_{ik}}))^{Z_{i1}} (p_2(w_{21}^{x_{i1}} ... w_{2k}^{x_{ik}}))^{Z_{i2}} ... (p_M(w_{M1}^{x_{i1}} ... w_{Mk}^{x_{ik}}))^{Z_{iM}} \right]$$

$$= \prod_{i=1}^{n} \left[p_1^{Z_{i1}} (w_{11}^{x_{i1}Z_{i1}} ... w_{1k}^{x_{ik}Z_{i1}}) p_2^{Z_{i2}} (w_{21}^{x_{i1}Z_{i2}} ... w_{2k}^{x_{ik}Z_{i2}}) ... p_M^{Z_{iM}} (w_{M1}^{x_{i1}Z_{iM}} ... w_{Mk}^{x_{ik}Z_{iM}}) \right]$$

$$= \prod_{i=1}^{n} \left[p_1^{Z_{i1}} p_2^{Z_{i2}} ... p_M^{Z_{iM}} (w_{11}^{x_{i1}Z_{i1}} ... w_{1k}^{x_{ik}Z_{i1}}) (w_{21}^{x_{i1}Z_{i2}} ... w_{2k}^{x_{ik}Z_{i2}}) ... (w_{M1}^{x_{i1}Z_{iM}} ... w_{Mk}^{x_{ik}Z_{iM}}) \right]$$

e desta forma obtemos

$$u(w|p,x) \propto \prod_{j=1}^{M} p_j^{\sum_{i=1}^{n} Z_{ij}} \prod_{j=1}^{M} \prod_{l=1}^{k} w_{jl}^{\sum_{i=1}^{n} x_{il} Z_{ij}}. \tag{4.10}$$

Distribuição *a priori* para p e w

Consideramos as distribuições *a priori* para os parâmetros do modelo (4.10) a distribuição *Dirichlet*. Desta forma, dados $\alpha_{jl} > 0$ temos que a distribuição *a priori* para w é dada por

$$P(w) = \prod_{j=1}^{M} \frac{\Gamma(\alpha_{j1} + \alpha_{j2} + ... + \alpha_{jk})}{\Gamma(\alpha_{j1})\Gamma(\alpha_{j2})...\Gamma(\alpha_{jk})} \prod_{l=1}^{k} w_{jl}^{\alpha_{jl}-1}$$

$$P(w) = \prod_{j=1}^{M} \frac{\Gamma(\sum_{l=1}^{k} \alpha_{jl})}{\Gamma(\alpha_{j1})\Gamma(\alpha_{j2})...\Gamma(\alpha_{jk})} \prod_{l=1}^{k} w_{jl}^{\alpha_{jl}-1} \tag{4.11}$$

onde os pontos $w_{j1}, w_{j2}, ..., w_{jk}$ são tais que $\sum_{l=1}^{k} w_{jl} = 1, 0 < w_{j1}, w_{j2}, ..., w_{jk} < 1$ para $j = 1, 2, ..., ,$ e $, = 1, 2, ..., k$.

De maneira análoga, temos que a distribuição *a priori* para p, sendo $\alpha_j > 0$ é dada por

$$P(p) = \frac{\Gamma(\sum_{j=1}^{M} \alpha_j)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_M)} \prod_{j=1}^{M} p_j^{\beta_j - 1} \tag{4.12}$$

onde $\sum_{j=1}^{M} p_j = 1, 0 < p_j < 1$ para $j = 1, 2, ...,$.

Distribuição *a posteriori* para p e w

Considerando, *a priori,* $w_j = (w_{j1}, ..., w_{jk})$ para j=1,2,..., ' e p independentes, a distribuição *a posteriori* é dada por

$$P(w, p|x, 2) = \frac{P(w)P(p)u(w, p|x, 2)}{\int P(w)P(p)u(w, p|x, 2)dwdp},$$

ou

$$P(w, p|x, 2) \propto P(w)P(p)u(w, p|x, 2). \tag{4.13}$$

Substituindo em (4.13) a função de verossimilhança baseada nos dados ampliados, (4.10), e as distribuições *a priori* de w (4.11) e p (4.12) temos

$$P(w, p|x, 2) \propto \prod_{j=1}^{M} p_j^{\beta_j - 1} \prod_{j=1}^{M} \prod_{l=1}^{k} w_{jl}^{\alpha_{jl} - 1} \prod_{j=1}^{M} p_j^{\sum_{i=1}^{n} Z_{ij}} \prod_{j=1}^{M} \prod_{l=1}^{k} w_{jl}^{\sum_{i=1}^{n} x_{il} Z_{ij}},$$

que pode se reescrita como

$$P(w, p|x, 2) \propto \prod_{j=1}^{M} p_j^{\left[\sum_{i=1}^{n} Z_{ij}\right] + \beta_j - 1} \prod_{j=1}^{M} \prod_{l=1}^{k} w_{jl}^{\left[\sum_{i=1}^{n} x_{il} Z_{ij}\right] + \alpha_{jl} - 1}. \tag{4.14}$$

Distribuições condicionais

A distribuição *a posteriori* condicional de w_j é dada por

$$w_j|p, x, 2 \sim Dirichlet\left((\sum_{i=1}^{n} x_{i1} Z_{ij}) + \alpha_{j1}, ..., (\sum_{i=1}^{n} x_{ik} Z_{ij}) + \alpha_{jk}\right), \tag{4.15}$$

para $j = 1, 2, ..., '$.

A distribuição *a posteriori* condicional de p é

$$p|w, x, 2 \sim Dirichlet\left(\sum_{i=1}^{n} Z_{i1} + \alpha_1, ..., \sum_{i=1}^{n} Z_{iM} + \alpha_M\right). \tag{4.16}$$

A distribuição *a posteriori* condicional de Z_i é dada por

$$Z_i|x\text{cwc}p \sim \text{'} \quad u,tino3\ i=,(1;q_{i1}\text{c...c}q_{iM})\text{c} \tag{4.17}$$

sendo que

$$q_{ij} = \frac{p_j\text{s}\,(x_i|\text{w}_j)}{\sum_{r=1}^{M} p_r\text{s}\,(x_i|\text{w}_r)}\text{c}\ 1 \leq i \leq n \text{ e } 1 \leq j \leq \text{'} \quad . \tag{4.18}$$

Desta forma o algoritmo *Gibbs-Sampling* utilizado em distribuições com mistura de multinomiais considerando ' conhecido é descrito como:

Passo 1:

Fazer $t = 0$. Considerar valores iniciais $\text{w}_j^t = (\text{w}_{j1}^t\text{c...cw}_{jk}^t)$ e $p^t = (p_\perp^t\text{c...c}p_M^t)$.

Passo 2:

Gerar $Z_i^{t+\perp}$ de uma distribuição multinomial com parâmetros $(1; q_{i\perp}^{t+\perp}\text{c}q_{i2}^{t+\perp}\text{c...c}q_{iM}^{t+\perp}))$ para $i = 1\text{c}2\text{c...c}n$, onde

$$q_{ij}^{t+\perp} = \frac{p_j^t\text{s}\,(x_i|\text{w}_j)^t}{\sum_{r=1}^{M} p_r^t\text{s}\,(x_i|\text{w}_r)^t}.$$

Passo 3:

Gerar $\text{w}_j^{t+\perp}|p^t\text{cf}\ \text{c}Z^{t+\perp}$ de uma distribuição *Dirichlet* com parâmetro

$$\left((\sum_{i=\perp}^{n} x_{i\perp}Z_{ij}^{t+\perp}) + \text{k}_{j\perp}\text{c}(\sum_{i=1}^{n} x_{i2}Z_{ij}^{t+\perp}) + \text{k}_{j2}\text{...c}(\sum_{i=1}^{n} x_{ik}Z_{ij}^{t+\perp}) + \text{k}_{jk}\right)$$

e, gerar $p^{t+\perp}|\text{w}^{t+\perp}\text{cf}\ \text{c}Z^{t+\perp}$ de uma distribuição *Dirichlet* com parâmetro

$$\left(\sum_{i=1}^{n} Z_{i1}^{t+\perp} + \text{q}_\perp\text{c...c}\sum_{i=\perp}^{n} Z_{iM}^{t+\perp} + \text{q}_M\right).$$

Passo 4: Fazer $t = t + 1$.

Repetir os passos 2 a 4 um número grande de vezes até que haja convergência. Descartar os 3 valores iniciais e temos que os valores gerados de we p, podem ser considerados como uma amostra das respectivas distribuições *a posteriori.*

4.6 Escolha da distribuição *a priori* informativa

Em determinadas situações existem informações disponíveis sobre os parâmetro do modelo. A questão é como esta informação pode ser transformada em uma distribuição *a priori* para que seja devidamente utilizada.

Nesta seção, apresentamos uma maneira possível de como estabelecer uma distribuição *a priori* partindo de informações prévias sobre os parâmetro do modelo.

Considere um caso onde as observações são obtidas de uma população particionada em ’ componentes sendo que as observações f $_i$ = (f $_{i1}$cf $_{i2}$c...cf $_{ik}$), i = 1c2c...n, tenham uma distribuição multinomial com parâmetro N e w$_j$ = (w$_{j1}$cw$_{j2}$c...cw$_{jk}$), j = 1c2c...c’ . Considere também que saibamos qual é o número ’ de componentes e que dispomos de informações sobre algumas das componentes w$_j$ da mistura. Vamos supor que tenhamos a informação de que algum elemento da amostra tenha vindo de um grupo onde as proporções de cada componente da multinomial são conhecidas. Sejam estas proporções denotadas por , $_j$ = (, $_{j1}$c, $_{j2}$c...c, $_{jk}$) para j = 1c2c...c’ .

Supondo w$_j$ = (w$_{j1}$cw$_{j2}$c...cw$_{jk}$) independentes para j = 1c2c...c’ , com

$$\mathrm{w}_j \sim Dioi\mathrm{P}h,\mathrm{e}t(\mathrm{k}_{j1}\mathrm{ck}_{j2}\mathrm{c...ck}_{jk})\mathrm{c}$$

a distribuição *a priori* para os parâmetros w$_j$ segue uma distribuição *Dirichlet* definida como

$$P(\mathrm{w}) = \prod_{j=1}^{M} \frac{\Gamma(\mathrm{k}_{j1} + \mathrm{k}_{j2} + ... + \mathrm{k}_{jK})}{\Gamma(\mathrm{k}_{j1})\Gamma(\mathrm{k}_{j2})...\Gamma(\mathrm{k}_{jk})} \prod_{l=1}^{k} \mathrm{w}_{jl}^{\alpha_{jl}-1}\mathrm{c} \tag{4.19}$$

para k$_{jl}$ 7 0 onde $\sum_{l=1}^{k}$ w$_{jl}$ = 1c0 < w$_{j1}$cw$_{j2}$c...cw$_{jk}$ < 1.

Na distribuição *Dirichlet* a média de w$_{jl}$ é dada por

$$. \ (\mathrm{w}_{jl}) = \frac{\mathrm{k}_{jl}}{\sum_{l=1}^{k} \mathrm{k}_{jl}}\mathrm{c} \tag{4.20}$$

para j = 1c2c...c’ , e, a variância é dada por

$$\mathrm{T}\left(\mathrm{w}_{jl}\right)=\frac{\mathrm{k}_{jl}\left[\left(\sum_{l=1}^{k}\mathrm{k}_{jl}\right)-\mathrm{k}_{jl}\right]}{\left(\sum_{l=1}^{k}\mathrm{k}_{jl}\right)^{2}\left[\left(\sum_{l=1}^{k}\mathrm{k}_{jl}\right)+1\right]}\mathrm{c} \tag{4.21}$$

para $j=1\mathrm{c}2\mathrm{c}...\mathrm{c}'$.

Desta maneira, podemos estabelecer uma distribuição *a priori* informativa da seguinte forma, igualamos o valor esperado da proporção à média (4.20) da distribuição *Dirichlet* obtendo

$$,_{jl}=\frac{\mathrm{k}_{jl}}{\sum_{l=1}^{k}\mathrm{k}_{jl}}\mathrm{c} \tag{4.22}$$

para $j=1\mathrm{c}2\mathrm{c}...\mathrm{c}'$. Dividindo em (4.21) tanto o numerador quanto o denominador por $\sum_{l=1}^{k}\mathrm{k}_{jl}$ e substituindo (4.22) em (4.21) obtemos

$$v\text{=o}(\mathrm{w}_{jl})=,_{jl}\frac{(1-,_{jl})}{\left[\left(\sum_{l=1}^{k}\mathrm{k}_{jl}\right)+1\right]}. \tag{4.23}$$

Reescrevendo (4.23) temos

$$\sum_{l=1}^{k}\mathrm{k}_{jl}=,_{jl}\frac{(1-,_{jl})}{v\text{=o}(\mathrm{w}_{jl})}-1. \tag{4.24}$$

Através de (4.24) podemos controlar a precisão da informação disponível previamente através de $\sum_{l=1}^{k}\mathrm{k}_{jl}$, ou seja, quanto maior o valor de $\sum_{l=1}^{k}\mathrm{k}_{jl}$ menor será a variância de w_{jl} para $,=1\mathrm{c}2\mathrm{c}...\mathrm{c}k$ e $j=1\mathrm{c}2\mathrm{c}...\mathrm{c}'$. Para ilustrar esta estratégia propomos o seguinte exemplo.

4.6.1 Exemplo

Suponha que uma população de interesse esteja particionada em duas componentes como por exemplo pessoas do sexo masculino e pessoas do sexo feminino. Suponha também que os dados seguem uma distribuição multinomial com parâmetros iguais a $(N\mathrm{cw}_{j})\mathrm{c}$ sendo $\mathrm{w}_{j}=(\mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{cw}_{j3})\mathrm{c}\, j=1$ e 2.

Considere que previamente tenhamos uma expectativa de que as proporções amostrais sejam

$$, \ _1 = (0c2; 0c3; 0c5) \tag{4.25}$$

para as observações vindas da componente 1 e

$$, \ _2 = (0c1; 0c6; 0c3) \tag{4.26}$$

para as observações provenientes da componente 2, tendo uma variação de 0c1 em torno de cada uma destas proporções.

Utilizando uma distribuição $Dioi\mathcal{P}h, et$ com parâmetro $\mathrm{k}_j = (\mathrm{k}_{j1}c\mathrm{k}_{j2}c\mathrm{k}_{j3})$ como distribuição *a priori* para wc igualamos a média da distribuição $Dioi\mathcal{P}h, et$ (4.20) à cada proporção (4.25) e (4.26) como,

$$\frac{\mathrm{k}_{11}}{\sum_{l=1}^{3}\mathrm{k}_{1l}} = 0c2; \frac{\mathrm{k}_{12}}{\sum_{l=1}^{3}\mathrm{k}_{1l}} = 0c3; \frac{\mathrm{k}_{13}}{\sum_{l=1}^{3}\mathrm{k}_{1l}} = 0c5 \tag{4.27}$$

e

$$\frac{\mathrm{k}_{21}}{\sum_{l=1}^{3}\mathrm{k}_{2l}} = 0c1; \frac{\mathrm{k}_{22}}{\sum_{l=1}^{3}\mathrm{k}_{2l}} = 0c6; \frac{\mathrm{k}_{23}}{\sum_{l=1}^{3}\mathrm{k}_{2l}} = 0c3. \tag{4.28}$$

Como temos a informação de que para cada proporção a variação é de 0c1, as proporções esperadas exibidas em (4.27) estarão respectivamente nos intervalos $[0c1; 0c3]c$ $[0c2; 0c4]$ e $[0c4; 0c6]$ e as proporções esperadas exibidas em (4.28) estarão respectivamente nos intervalos$[0; 0c2]c\,[0c5; 0c7]$ e $[0c2; 0c4]$. Desta forma, a amplitude da variação esperada é de 0c2.

Fazendo o quociente $\frac{0,2}{4} = 0c05$ temos aproximadamente o valor do desvio padrão amostral. E fazendo $(0c05)^2 = 0c0025$ temos aproximadamente o valor da a variância amostral. Substituindo o valor da variância amostral e os valores das proporções dadas em (4.27) e (4.28) em (4.23) e, reescrevendo com (4.24) obtemos

$$\frac{0c2(1-0c2)}{0c0025}-1=63;\ \frac{0c3(1-0c3)}{0c0025}-1=83;\ \frac{0c5(1-0c5)}{0c0025}-1=99; \tag{4.29}$$

e

$$\frac{0c1(1-0c1)}{0c0025}-1=35;\ \frac{0c6(1-0c6)}{0c0025}-1=95;\ \frac{0c3(1-0c3)}{0c0025}-1=83. \tag{4.30}$$

Em seguida, calculamos a média para os valores encontrados, ou seja (63+83+99)/3 que é igual a 81,66 e (35+95+83)/3 que é igual a 71. Fazemos

$$\mathrm{k}_{\perp}=(0c2;0c3;0c5)\times 81c66 \text{ e obtemos } (16c33;24c49;40c83)$$

e,

$$\mathrm{k}_{2}=(0c1;0c6;0c3)\times 71 obtendo (7c1;42c6;21c3).$$

Substituindo estes valores em (4.19) temos

$$P(\mathrm{w})=\frac{\Gamma(16c33+24c49+40c83+7c1+42c6+21c3)}{\Gamma(16c33)\Gamma(24c49)\Gamma(40c83)\Gamma(7c1)\Gamma(42c6)\Gamma(21c3)}\times$$

$$\mathrm{w}_{11}^{(16,33)-1}\mathrm{w}_{12}^{(24,49)-1}\mathrm{w}_{13}^{(40,83)-1}\mathrm{w}_{21}^{(7,1)-1}\mathrm{w}_{22}^{(42,6)-1}\mathrm{w}_{23}^{(21,3)-1}\mathrm{c}$$

ou seja,

$$\mathrm{w}_{1l}\sim Dioi\mathrm{P}h,et\,(16c33;24c49;40c83) \tag{4.31}$$

e

$$\mathrm{w}_{2l}\sim Dioi\mathrm{P}h,et\,(7c1;42c6;21c3)$$

para , = 1c2c3.

Assim, estabelecemos uma distribuição *a priori* para wpartindo de informações sobre alguns dos parâmetros do modelo.

Capítulo 5

Aplicações com dados simulados

5.1 Introdução

Neste capítulo, fazemos aplicações do modelo proposto utilizando dados gerados por computador que simulam o comportamento de proteínas. Como os dados são gerados temos condições de avaliar o desempenho do método em estimar os parâmetros de interesse.

Apresentamos dois exemplos. No primeiro aplicamos o modelo de mistura em dados gerados que simulam o comportamento de proteínas sendo vistas através de bases nitrogenadas. No segundo exemplo, o modelo é aplicado em dados que também simulam o comportamento de proteínas mas, sendo vistos através de aminoácidos que formam as proteínas.

5.2 Características das observações

Apesar dos dados serem artificiais eles foram gerados a partir das proporções reais de duas proteínas que pertencem a dois *grupos protéicos* distintos. A primeira proteína, é a *hemoglobina beta* que possui a função de transporte de oxigênio no sangue e é encontrada nos seres humanos, a segunda, é uma enzima que que atua na degradação protéica chamada de *chymotrypsina* encontrada no levedo de pão.

Em cada exemplo geramos três observações multinomiais sendo que, duas delas foram geradas utilizando as proporções das moléculas contidas na proteína *chymotrypsina* e a outra observação foi gerada utilizando as proporções das moléculas contidas na proteína *hemoglobina beta.* Estas proteínas estão disponíveis no site *http://www.ncbi.nlm .nih.gov/.*

Geramos estas observações simulando a presença de duas componentes particionando a população, sendo que, estas componentes representam os *grupos protéicos* à qual cada proteína pertence.

Desta forma, as observações geradas são representadas por

$$\mathrm{f}_{\ i} = (\mathrm{f}_{\ i1}\mathrm{cf}_{\ i2}\mathrm{c...cf}_{\ ik}) \sim \text{'}\ \ u{,}tino3\ i{=\!\!\!|}(N; \mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{c...cw}_{jk})\mathrm{c}$$

para $i = 1\mathrm{c}2\mathrm{c}3$, e $j = 1\mathrm{c}2$. Vale ressaltar que N representa o comprimento das seqüências de aminoácidos e de bases nitrogenadas e que estes comprimentos são previamente conhecidos.

5.3 Modelos

Função de verossimilhança

Dado que $\mathrm{f}_{\ i}\mathrm{c}$ para $i = 1\mathrm{c}2\mathrm{c}$ e 3, foi observado, a função de verossimilhança utilizando variáveis latentes é dada por

$$\mathrm{u}\,(\mathrm{w}\mathrm{c}p|x\,\mathrm{c}2) \propto \prod_{j=1}^{2} p_j^{\sum_{i=1}^{3} Z_{ij}} \prod_{j=1}^{2}\prod_{l=1}^{k} \mathrm{w}_{jl}^{\sum_{i=1}^{3} x_{il} Z_{ij}}. \tag{5.1}$$

5.3.1 Abordagem bayesiana

Distribuição *a priori*

Nestes dois exemplos utilizamos uma abordagem bayesiana sendo que as distribuições *a priori* para os parâmetro do modelo têm distribuições *Dirichlet* não informativas. Desta

forma, a distribuição *a priori* para os parâmetro $\mathrm{w}_j = (\mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{c...cw}_{jk})$ com $\mathrm{w}_1\mathrm{c\,w}_2\mathrm{,...,w}_M$ independentes onde $\mathrm{w}_j \sim Dioi\mathrm{P}h,\mathrm{e}t(\mathrm{k}_{j1}\mathrm{ck}_{j2}\mathrm{c...ck}_{jk})$, é dada por

$$P(\mathrm{w}_j) \sim Dioi\mathrm{P}h,\mathrm{e}t(\mathrm{k}_{j1}\mathrm{ck}_{j2}\mathrm{c...ck}_{jk})\mathrm{c} \tag{5.2}$$

com $\mathrm{k}_{jl} = 1\mathrm{c}$ para $j = 1\mathrm{c}2$, e , $= 1\mathrm{c}2\mathrm{c...c}k$, e a distribuição *a priori* para os parâmetros p é dada por,

$$P(p) \sim Dioi\mathrm{P}h,\mathrm{e}t(\mathrm{q}_j)\mathrm{c} \tag{5.3}$$

com $\mathrm{q}_j = 1$. Substituindo em (4.11) e em (4.12) os valores dos parâmetros das distribuições *a priori* para we p temos que

$$P(\mathrm{w}_j) \propto (\mathrm{w}_{j1}^{1-1}\mathrm{w}_{j2}^{1-1}...\mathrm{w}_{jk}^{1-1})$$

$$P(\mathrm{w}_j) \propto 1, \tag{5.4}$$

para $j = 1\mathrm{c}2$ e também

$$P(p) \propto (p_1^{1-1}p_2^{1-1}...p_M^{1-1})$$

$$P(p) \propto 1. \tag{5.5}$$

Distribuição *a posteriori*

A distribuição *a posteriori* para $\mathrm{w}_j = (\mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{c...cw}_{jk})$ com $\mathrm{w}_1\mathrm{c\,w}_2\mathrm{,...,w}_M$ e p independentes é obtida fazendo o produto da distribuição *a priori* de w (5.4) e p (5.5) pela função de verossimilhança (5.1). Desta maneira temos que a distribuição *a posteriori* conjunta para we p é dada por

$$P(\mathrm{w}p|x\mathrm{c}2) \propto \prod_{j=1}^{2} p_j^{\sum_{i=1}^{3} Z_{ij}} \prod_{j=1}^{2} \prod_{l=1}^{k} \mathrm{w}_{jl}^{\sum_{i=1}^{3} x_{il} Z_{ij}} \tag{5.6}$$

Distribuições *a posteriori* condicionais

Dados $p\mathrm{c}x\mathrm{c}2\mathrm{c}$ temos que $\mathrm{w}_1\mathrm{c}\ \mathrm{w}_2,...,\mathrm{w}_M$ são independentes e a distribuição *a posteriori* condicional de w_j é dada por

$$\mathrm{w}_j|p\mathrm{c}x\mathrm{c}2 \sim Dio\imath P\hbar,\mathrm{e}t\left(\sum_{i=1}^{3} x_{i1}Z_{ij} + 1\mathrm{c}...\mathrm{c}\sum_{i=1}^{3} x_{ik}Z_{ij} + 1\right)\mathrm{c} \tag{5.7}$$

para $j = 1\mathrm{c}2$.

Dados $\mathrm{w}\mathrm{c}x\mathrm{c}2$ a distribuição *a posteriori* condicional para o parâmetro p é

$$p|\mathrm{w}\mathrm{c}x\mathrm{c}2 \sim Dio\imath P\hbar,\mathrm{e}t\left(4_1 + 1\mathrm{c}4_2 + 1\right)\mathrm{c} \tag{5.8}$$

onde $4_j = \sum_{i=1}^{n} Z_{ij}$. É interessante notar que S_j representa o total das observações que pertencem a componente j para $j = 1\mathrm{c}2$.

A distribuição *a posteriori* condicional de Z_i é dada como mostra (4.17) e (4.18).

5.3.2 Exemplo

Neste exemplo, foram geradas três observações multinomiais, como é explicado na seção 5.2, que simula o comportamento de proteínas observadas através das proporções de bases nitrogenadas como

$$\mathrm{f}_{\ i} \sim \text{’}\quad u,tino3\ i=,(300;\mathrm{w}_{j1}\mathrm{c}\mathrm{w}_{j2}\mathrm{c}\mathrm{w}_{j3}\mathrm{c}\mathrm{w}_{j4})\mathrm{c} \tag{5.9}$$

onde $i = 1\mathrm{c}2\mathrm{c}3$ e $j = 1\mathrm{c}2$ e 300 representa o comprimento da seqüência de bases nitrogenadas. Utilizando as observações (5.9) geramos e estimamos os parâmetros através das distribuições condicionais (5.7) e (5.8) utilizando o algoritmo *Gibbs Sampling* descrito na seção 4.5.2. Foram feitas 50000 iterações com um descarte inicial de 500 valores e com

saltos de 10 em 10 foram selecionados os valores restantes.

Os resultados obtidos são exibidos nas Tabelas 6 e 7. Temos que na primeira coluna da Tabela 6 estão representados os parâmetros que estimam as probabilidades w da distribuição multinomial. Na segunda coluna temos as proporções com que estes dados foram inicialmente gerados, ou seja, as proporções originais. Na coluna 3 e 4 temos respectivamente a média e o desvio padrão e, nas colunas 5 a 9 temos os quantis de w de 2,5%, 25%, 50%, 75% e 97,5% respectivamente. Temos que as linhas 2 a 5 da Tabela 6 representam os resumos *a posteriori* da primeira componente e, as linhas 6 a 10 representam os resumos *a posteriori* da segunda componente.

Tabela 6: Resumos *a posteriori* dos parâmetro w

parâmetro	Originais	Média	Desvio	2,5%	25%	50%	75%	97,5%
w_{11}	0,2941	0,2531	0,0177	0,2194	0,2409	0,2527	0,2650	0,2879
w_{12}	0,2527	0,2633	0,0178	0,2294	0,2510	0,2628	0,2750	0,2994
w_{13}	0,2826	0,2932	0,0187	0,2578	0,2802	0,2928	0,3058	0,3303
w_{14}	0,1706	0,1905	0,0161	0,1599	0,1794	0,1900	0,2010	0,2228
w_{21}	0,1905	0,1610	0,0213	0,1226	0,1465	0,1601	0,1743	0,2044
w_{22}	0,2413	0,1907	0,0228	0,1479	0,1750	0,1897	0,2057	0,2364
w_{23}	0,2476	0,2764	0,0256	0,2274	0,2590	0,2757	0,2934	0,3286
w_{24}	0,3206	0,3720	0,0276	0,3183	0,3529	0,3721	0,3907	0,4272

Através da Tabela 6 podemos notar que a média dos valores gerados estão próximas das proporções originais, tanto as médias da primeira componente como as médias da segunda. Notamos também que, de uma forma geral, o desvio padrão para a primeira componente é menor que para a segunda e isto se deve ao fato de termos mais observações provenientes da primeira componente.

Na Tabela 7 temos os resumos *a posteriori* para as estimativas dos parâmetros das proporções da mistura do modelo (5.6). Na segunda linha temos os resumos da estimativa para a primeira componente enquanto que na terceira linha temos os resumos da estimativa para a segunda componente. Temos que na primeira coluna estão representados os

parâmetros que estimam as ponderações p atribuídas a cada componente da mistura. Nas colunas 2 e 3 estão a média e o desvio padrão de p. Nas colunas 4 a 8 estão representadas respectivamente os quantis de 2,5%, 25%, 50%, 75% e 97,5% dos parâmetros estimados.

Tabela 7: Resumos *a posteriori* dos parâmetro p

parâmetro	Média	Desvio	2,5%	25%	50%	75%	97,5%
p_1	0,5976	0,2007	0,1915	0,4501	0,6163	0,7560	0,9271
p_2	0,4034	0,2007	0,0727	0,2440	0,3836	0,5498	0,8074

Através da Tabela 7 podemos ver que a média para a proporção da mistura p_1 foi maior que p_2. Este resultado é esperado pelo fato de termos duas observações sendo geradas da primeira componente e uma observação sendo gerada da segunda componente, ou seja, a probabilidade de uma observação ser proveniente da primeira componente é maior que a probabilidade de ser proveniente da segunda. Notamos um desvio padrão idêntico para os dois parâmetros estimados.

5.3.3 Resultados gráficos do exemplo 5.3.2

Exibimos, nesta seção, os gráficos de autocorrelação, gráficos de densidades e gráficos de convergência para algumas das estimativas dos parâmetros do modelo sendo que as demais apresentaram um comportamento semelhante.

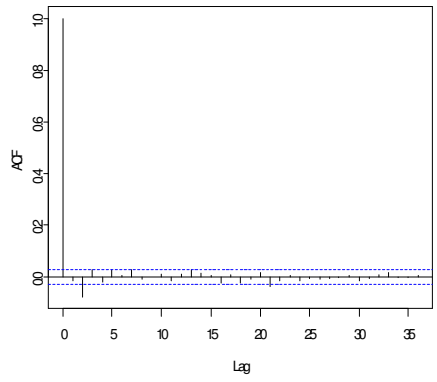


Figura 4: Autocorrelação de w_{12}

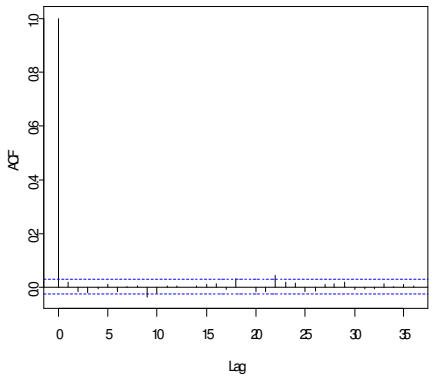


Figura 5: Autocorrelação de w_{22}

Nas Figuras 4 e 5 apresentamos os gráficos de autocorrelação de w_{12} e w_{22} respectivamente.

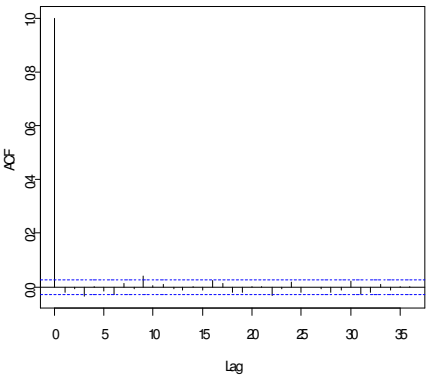


Figura 6: Autocorrelação de w_{13}

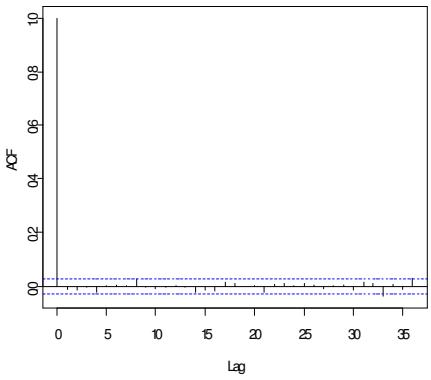


Figura 7: Autocorrelação de w_{23}

Nas Figuras 6 e 7 são exibidos os gráficos de autocorrelação de w_{13} e w_{23} respectivamente.

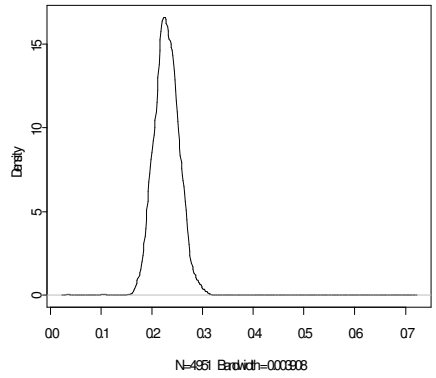


Figura 8: Densidade de w_{12}

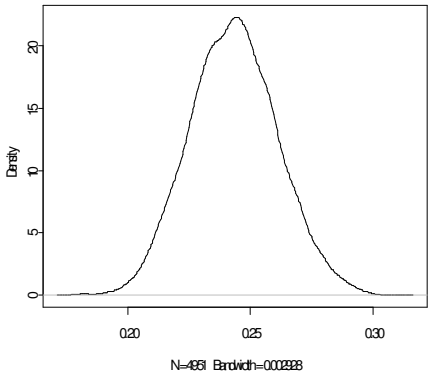


Figura 9: Densidade de w_{22}

Os gráficos das densidades das estimativas de w_{12} e w_{22} são exibidos nas Figuras 8 e 9 respectivamente.

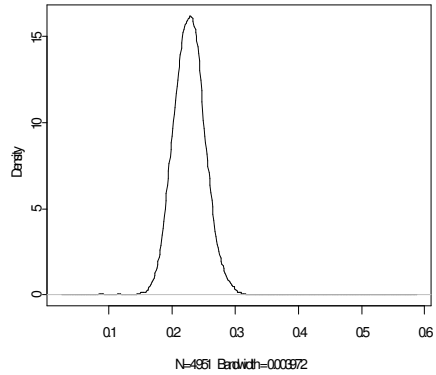


Figura 10: Densidade de w_{13}

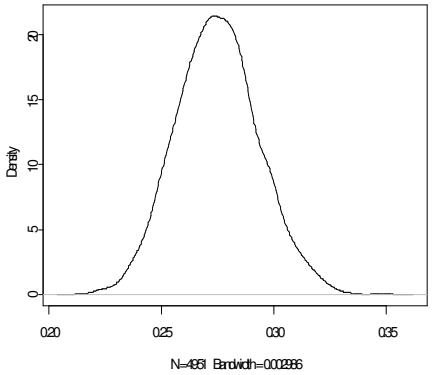


Figura 11: Densidade de w_{23}

Nas Figuras 10 e 11 estão representadas as densidades das estimativas de w_{13} e w_{23}.

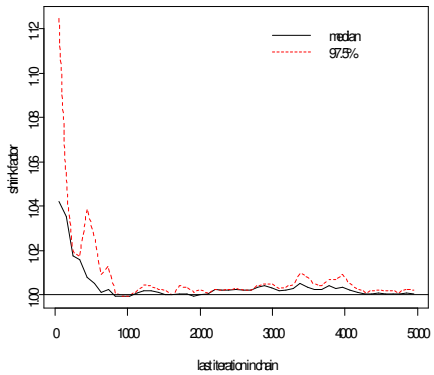


Figura 12: Convergência de w_{12}

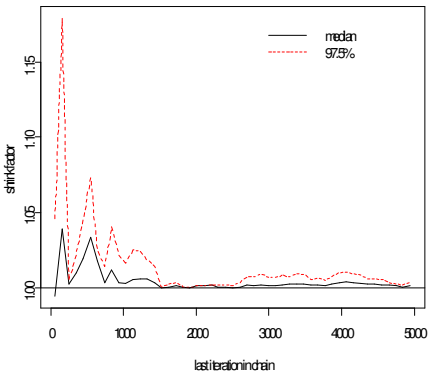


Figura 13: Convergência de w_{22}

Utilizando o critério de convergência de Gelman e Rubin (1992), apresentamos nas Figura 12 e 13 respectivamente os gráficos de convergência de w_{12} e w_{22}.

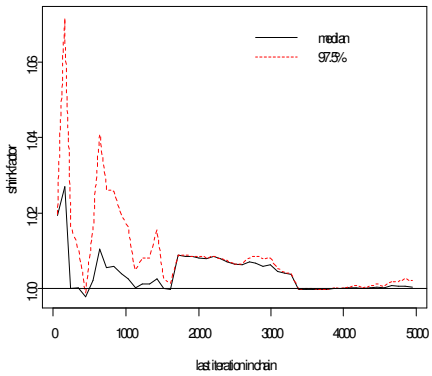


Figura 14: Convergência de w_{13}

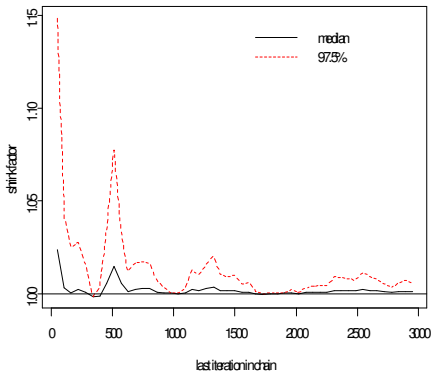


Figura 15: Convergência de w_{23}

Nas Figuras 14 e 15, apresentamos respectivamente os gráficos de convergência de w_{13} e w_{23}.

Notamos através das Figuras 4 a 15 resultados gráficos satisfatórios em relação a autocorrelação, densidades e convergência dos parâmetros estimados.

5.3.4 Exemplo

Foram geradas três observações multinomiais, segundo a seção 5.2, que simula o comportamento de proteínas observadas através das proporções de aminoácidos. Desta forma as observações são dadas como

$$f_i \sim \text{multinomial}(1000; w_{j1}, w_{j2}, w_{j3}, \ldots, w_{j18}, w_{j19}, w_{j20}) \qquad (5.10)$$

onde *i =1,2, 3, j=1,2* e 1000 representa o comprimento da seqüência de aminoácidos.

Segundo as observações (5.10), utilizando o algoritmo *Gibbs Sampling* geramos e estimamos os parâmetros através das distribuições condicionais (5.7) e (5.8). Para a geração e estimação dos parâmetros foram realizadas 50000 iterações do algoritmo, com um descarte inicial de 500 valores. Com saltos de 10 em 10, selecionamos os valores restantes.

Os resultados obtidos são exibidos nas Tabela 8 e 9. Na primeira coluna da Tabela 8 estão os parâmetros que estimam as probabilidades w da distribuição multinomial no modelo de mistura. Na segunda coluna temos as proporções com que estes dados foram inicialmente gerados e, na terceira coluna as médias para os parâmetros que estimam estas proporções. Na coluna 4 temos o desvio padrão para cada parâmetro e, nas colunas 5 a 9 temos os quantis de w de 2,5%, 25%, 50%, 75% e 97,5% respectivamente.

Tabela 8: Resumos *a posteriori* dos parâmetro w

parâmetro	Originais	Média	Desvio	2,5%	25%	50%	75%	97,5%
w_{11}	0c0462	0c0852	0c0209	0c0143	0c0776	0c0841	0c0907	0c1049
w_{12}	0c1630	0c0333	0c0172	0c0116	0c0287	0c0330	0c0374	0c0740
w_{13}	0c0332	0c0363	0c0169	0c0124	0c0318	0c0358	0c0405	0c0728
w_{14}	0c2800	0c0745	0c0187	0c0127	0c0674	0c0736	0c0797	0c0946
w_{15}	0c0104	0c0110	0c0223	0c0021	0c0043	0c0058	0c0080	0c0766
w_{16}	0c0553	0c0578	0c0170	0c0127	0c0525	0c0580	0c0635	0c0821
w_{17}	0c0215	0c0144	0c0227	0c0045	0c0077	0c0098	0c0125	0c0755
w_{18}	0c0612	0c0916	0c0231	0c0133	0c0884	0c0957	0c1026	0c1168
w_{19}	0c0033	0c0314	0c0191	0c0130	0c0271	0c0312	0c0356	0c0745
w_{110}	0c0625	0c0063	0c0218	0c0000	0c0003	0c0008	0c0118	0c0744
w_{111}	0c0345	0c1179	0c0305	0c0140	0c1166	0c1250	0c1328	0c1480
w_{112}	0c0124	0c0776	0c0194	0c0134	0c0737	0c0803	0c0866	0c1009
w_{113}	0c0189	0c0231	0c0190	0c0101	0c0166	0c0197	0c0233	0c0773
w_{114}	0c0234	0c0494	0c0168	0c0146	0c0436	0c0486	0c0537	0c0806
w_{115}	0c0586	0c0558	0c0164	0c0130	0c0508	0c0562	0c0616	0c0777
w_{116}	0c1301	0c0367	0c0165	0c0107	0c0309	0c0351	0c0396	0c0701
w_{117}	0c2811	0c0629	0c0177	0c0132	0c0580	0c0637	0c0694	0c0862
w_{118}	0c0163	0c0241	0c0193	0c0112	0c0175	0c0206	0c0242	0c0753
w_{119}	0c0189	0c0178	0c0201	0c0069	0c0112	0c0137	0c0167	0c0773
w_{120}	0c0683	0c0938	0c0240	0c0138	0c0905	0c0978	0c1046	0c1213

Continuação da Tabela 8

parâmetro	Originais	Média	Desvio	2.5%	25%	50%	75%	97.5%
w_{21}	0,0849	0,0455	0,0047	0,0366	0,0423	0,0454	0,0484	0,0555
w_{22}	0,0283	0,0209	0,0032	0,0151	0,0187	0,0208	0,0230	0,0273
w_{23}	0,0377	0,0307	0,0039	0,0235	0,0281	0,0306	0,0332	0,0386
w_{24}	0,0660	0,0247	0,0028	0,0184	0,0223	0,0245	0,0269	0,0321
w_{25}	0,0091	0,0094	0,0021	0,0056	0,0074	0,0092	0,0107	0,0140
w_{26}	0,0566	0,0531	0,0049	0,0437	0,0476	0,0530	0,0563	0,0629
w_{27}	0,0094	0,0198	0,0031	0,0141	0,0154	0,0196	0,0218	0,0264
w_{28}	0,0943	0,0590	0,0052	0,0491	0,0554	0,0589	0,0624	0,0697
w_{29}	0,0472	0,0050	0,0017	0,0024	0,0038	0,0048	0,0059	0,0087
w_{210}	0,0094	0,0674	0,0058	0,0564	0,0634	0,0673	0,0712	0,0788
w_{211}	0,1226	0,0347	0,0042	0,0272	0,0318	0,0345	0,0373	0,0431
w_{212}	0,0755	0,0124	0,0027	0,0080	0,0107	0,0123	0,0139	0,0176
w_{213}	0,0189	0,0203	0,0031	0,0146	0,0181	0,0201	0,0222	0,0267
w_{214}	0,0566	0,0238	0,0034	0,0176	0,0214	0,0236	0,0258	0,0312
w_{215}	0,0472	0,0609	0,0053	0,0509	0,0574	0,0609	0,0644	0,0718
w_{216}	0,0443	0,1177	0,0074	0,1034	0,1128	0,1177	0,1226	0,1321
w_{217}	0,0566	0,2873	0,0107	0,2678	0,2806	0,2873	0,2943	0,3072
w_{218}	0,0189	0,0179	0,0029	0,0127	0,0157	0,0177	0,0198	0,0242
w_{219}	0,0094	0,0164	0,0028	0,0114	0,0144	0,0162	0,0181	0,0223
w_{220}	0,1038	0,0733	0,0058	0,0621	0,0694	0,0731	0,0770	0,0851

Temos que ocorrem resultados nas Tabelas 6 e 8 onde os vvalores originais dos parâmetros estão fora dos intervalos exibidos. A razão para isto é decorrente do pouco número de observações feitas.

Na Tabela 9 temos os resumos *a posteriori* dos parâmetros que estimam as proporções da mistura. Temos que na primeira coluna estão representados os parâmetros que estimam as ponderações p atribuídas a cada componente da mistura. Nas colunas 2 e 3 estão

descritos a média e o desvio padrão para estes parâmetros. Nas colunas 4 a 8 estão respectivamente os quantis de 2,5%, 25%, 50%, 75% e 97,5% dos parâmetros estimados. Na segunda linha da Tabela 9 estão representados os resumos do parâmetro para a primeira componente enquanto que na terceira linha estão os resumos do parâmetro relacionados à segunda componente.

Tabela 9: Resumos *a posteriori* dos parâmetro p

Parâmetro	Média	Desvio	2,5%	25%	50%	75%	97,5%
p_1	0,4031	0,2009	0,0632	0,2475	0,3869	0,5475	0,8061
p_2	0,5969	0,2009	0,1907	0,4516	0,6131	0,7518	0,9365

Através da Tabela 9 temos que a média para o parâmetro que estima a ponderação p_1 da mistura é menor que a média para o parâmetro que estima a ponderação de p_2. Este resultado é contrário ao que se esperava pois existem mais observações provenientes da primeira componente da mistura do que da segunda componente.

5.3.5 Resultados gráficos do exemplo 5.3.4

Exibimos nesta seção os gráficos de autocorrelação, gráficos de densidade e gráficos de convergência para algumas das estimativas dos parâmetros obtidas no exemplo 5.2. As demais estimativas apresentaram um comportamento gráfico semelhante.

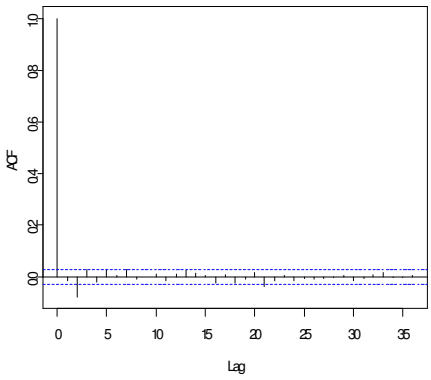


Figura 16: Autocorrelação de w_{12}

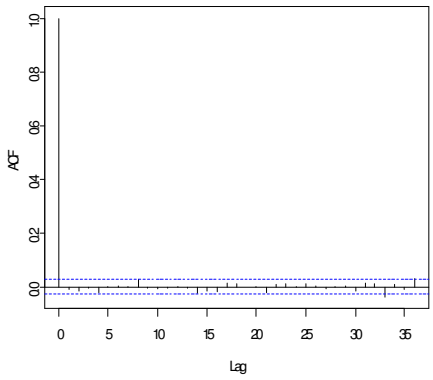


Figura 17: Autocorrelação de w_{22}

Nas Figuras 16 e 17 apresentamos os gráficos de autocorrelação de w_{11} e w_{22} respectivamente.

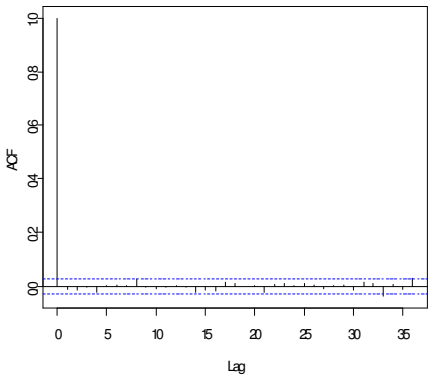


Figura 18: Autocorrelação de w_{14}

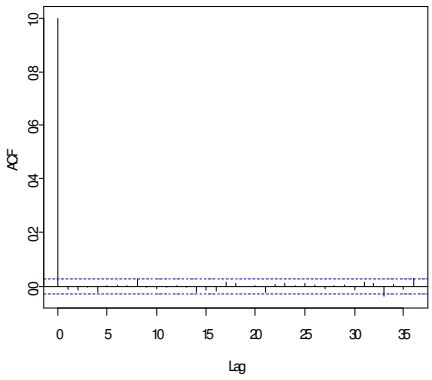


Figura 19: Autocorrelação de w_{24}

Os gráficos de autocorrelação de w_{14} e w_{24} são exibidos nas Figuras 18 e 19 respectivamente.

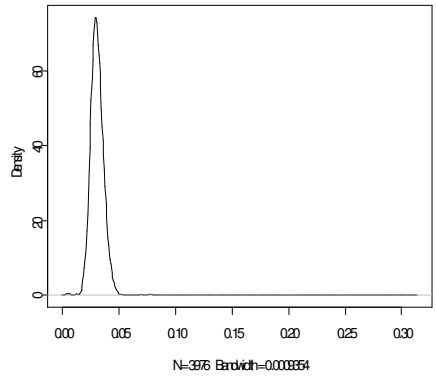


Figura 20: Densidade de w_{12}

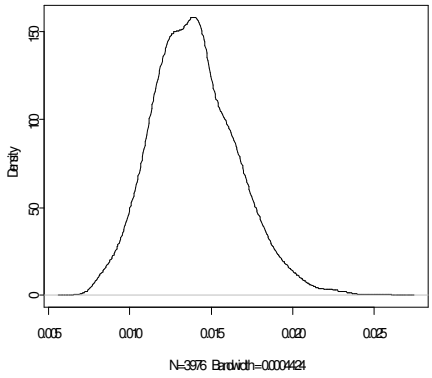


Figura 21: Densidade de w_{22}

Os gráficos de densidades das estimativas de w_{12} e w_{22} são exibidos nas Figuras 20 e 21.

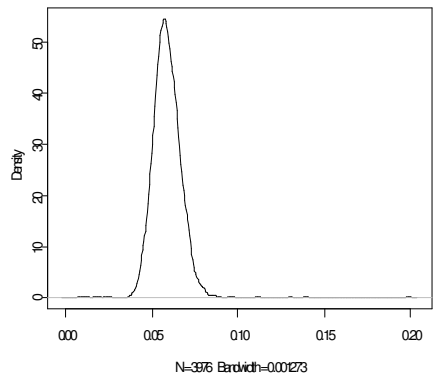


Figura 22: Densidade de w_{14}

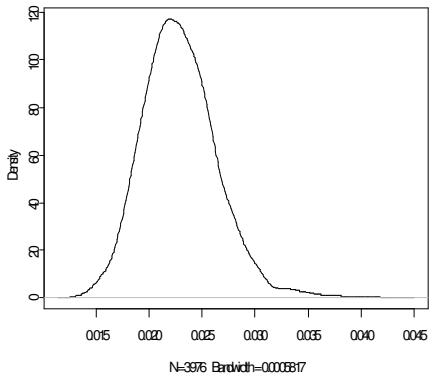


Figura 23 Densidade de w_{24}

Nas Figuras 22 e 23 temos as densidades das estimativas de w_{14} e w_{24}.

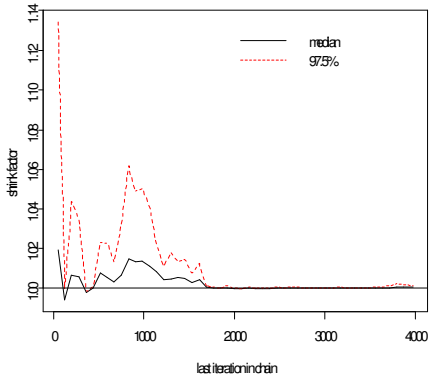


Figura 24: Convergência de w_{12}

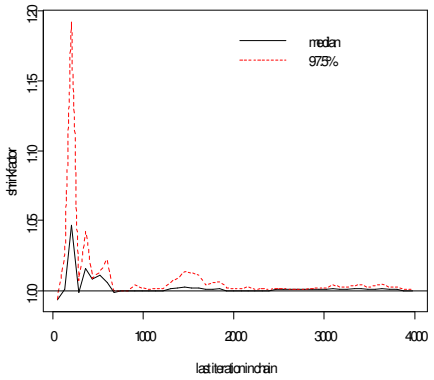


Figura 25 Convergência de w_{22}

Nas Figuras 24 e 25, exibimos os gráficos de convergência de w_{12} e w_{22}.

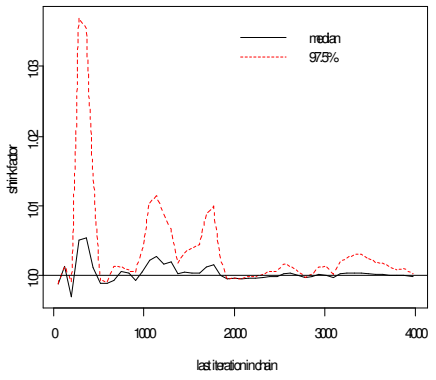


Figura 26: Convergência de w_{14}

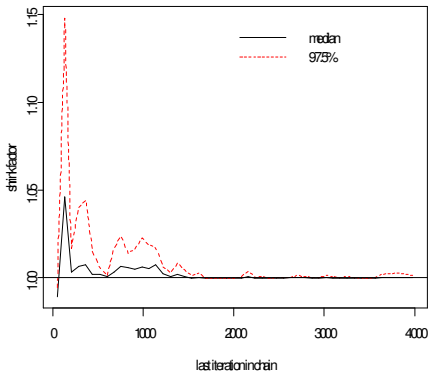


Figura 27 Convergência de w_{24}

Os gráficos de convergência de w_{14} e w_{24} estão respectivamente nas Figuras 26 e 27.

De maneira análoga ao exemplo 5.3.2, através das Figuras 16 a 27 temos resultados gráficos satisfatórios em relação a autocorrelação, densidades e convergências dos parâmetros estimados.

Capítulo 6

Análise de performance para dados simulados

6.1 Introdução

Nos capítulos anteriores expomos e exemplificamos um método estatístico que utiliza mistura finita de distribuições multinomiais, que ao verificar um grupo de proteínas seja capaz de identificar e alocar em uma mesma componente proteínas similares e separar as proteínas dissimilares em componentes distintas.

Neste capítulo, abordamos aspectos relacionados à eficiência do método empregado. A proposta aqui é verificar em que condições o modelo atua corretamente e quando ele é ineficiente. Esta verificação é feita observando dois pontos. Primeiro, a performance do método em manter juntas proteínas similares e a performance do método em separar as dissimilares. No segundo ponto, abordamos a questão da similaridade entre proteínas, ou seja, indicamos como medir quão próximas ou quão distantes duas ou mais proteínas estão uma da outra. Como medida para esta verificação propomos a divergência de *Kullback-Leibler*.

Propomos neste capítulo uma estratégia de verificar a eficiência do modelo em separar corretamente as observações em relação às componentes utilizando dados simulados. Nesta

estratégia, verificamos a proporção de acertos do modelo na separação das proteínas.

Apresentamos duas simulações para expormos esta análise de performance.

6.2 Proporção de acertos

Para verificarmos a performance do modelo em separar as observações em relação às componentes corretamente, introduzimos uma forma de indicar a eficiência do modelo empregado. Esta verificação é utilizada para dados gerados por computador considerando a presença de 2 componentes na mistura. Para estabelecermos a proporção de acertos do modelo em separar as proteínas fazemos algumas considerações.

Vamos supor que ao gerarmos os dados especifiquemos à qual componente cada observação pertence. Por exemplo, considere que tenhamos gerado 3 observações sendo que duas delas pertençam à uma componente e uma delas pertença a outra componente. Podemos sumarizar estas informações em um matriz, que chamamos de matriz A, como mostra a Tabela 10.

Tabela 10: Matriz A

	Componente 1	Componente 2
Observação 1	1	0
Observação 2	1	0
Observação 3	0	1

Na matriz A, exibida na Tabela 10, temos que as linhas representam as observações 1, 2 e 3 e as colunas as componentes 1 e 2. O número 1 indica que a observação foi gerada daquela componente e 0 caso contrário.

No capítulo 4, mostramos que as variáveis latentes $Z_i = (Z_{i1}\mathrm{c}Z_{i2}\mathrm{c}...Z_{iM})$, para $i = 1\mathrm{c}...\mathrm{c}n$, classificam cada observação em relação às componentes através de (4.17) e (4.18). Como a proporção de acertos é utilizada para dados gerados utilizando duas componentes, definimos uma matriz Z com dimensão n linhas por 2 colunas, análoga à matriz A, representando as variáveis latentes.

Definidas as matrizes A e Z, vamos mostrar como se verifica a proporção de acertos do modelo.

6.2.1 Cálculo da proporção de acertos

Passo 1: Definimos previamente uma matriz A onde temos a procedência das observações em relação às componentes como nos mostra a Tabela 10.

Passo 2: Conjuntamente ao algoritmo de estimação dos parâmetros do modelo, para cada uma das iterações utilizadas pelo algoritmo calculamos uma matriz Z de variáveis latentes..

Utilizando a primeira coluna de A e a primeira coluna de Z fazemos os cálculos

$$n_d = \sum_{i=1}^{n} |\mathrm{A}_{i1} - Z_{i1}| \mathrm{c} \tag{6.1}$$

$$i_d = 3 \acute{=} xi3\ o(n_d; n - n_d)\text{'}\ n\mathrm{c} \tag{6.2}$$

onde n é o número de observações. Em (6.1) temos a soma dos valores absolutos das diferenças da primeira coluna da matriz A com a primeira coluna da matriz Z, dada por n_d e, em (6.2) temos cálculo de i_d que é dado pelo máximo entre n_d e o número de observações menos n_d sendo que este máximo é dividido pelo número de observações n. Através de (6.1) e (6.2) temos a verificação do número de observações que foi separada corretamente em relação às componentes para cada iteração do algoritmo.

É importante ressaltar que os cálculos feitos em (6.1) e (6.2) são feitos desta forma pois estamos interessados na separação dos grupos e não necessariamente no rótulo 0 ou 1 atribuído a cada grupo. Desta maneira, para o caso onde a componente verdadeira é $\mathrm{A}_{i1} = [0\mathrm{c}1\mathrm{c}1]'$ as estimativas corretas poderiam ser $Z_{i1} = [0\mathrm{c}1\mathrm{c}1]'$ ou $Z_{i1} = [1\mathrm{c}0\mathrm{c}0]'$, para $i = 1\mathrm{c}2\mathrm{c}...c n\mathrm{c}$ pois ambas separam corretamente os grupos.

Passo 3: Finalizadas as iterações do algoritmo de estimação dos parâmetros, obtemos

a proporção de acertos denominada por P_a através de

$$P_a = \sum_{r=1}^{R} \frac{i_{d_r}}{R} \mathrm{c} \tag{6.3}$$

onde R representa o número de iterações feitas pelo algoritmo já descontados o descarte inicial e os saltos. Através de (6.3) temos o cálculo de P_a que nos fornece a proporção com que o método consegue separar corretamente as observações em relação às componentes da mistura.

Na seção seguinte, apresentamos um exemplo detalhado da utilização da proporção de acertos.

6.2.2 Exemplo

Considere uma população particionada em 2 componentes. Considere também que sejam geradas 3 observações tais que

$$\mathrm{f}_{\ i} = (\mathrm{f}_{\ i1}\mathrm{cf}_{\ i2}\mathrm{c...cf}_{\ ik}) \sim \text{'}\ \ u,tino3\ i{=},(N;\mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{c...cw}_{jk})\mathrm{c}$$

para $i = 1\mathrm{c}2\mathrm{c}3$, e $j = 1\mathrm{c}2$ onde a observação 1 é gerada de uma componente e as observações 2 e 3 são geradas de outra componente.

Seguindo os passos dados na seção 6.2.1 definimos na Tabela 11 uma matriz A indicando a procedência das observações em relação às componentes.

Tabela 11: Matriz de referência T

	Componente 1	Componente 2
Observação 1	0	1
Observação 2	1	0
Observação 3	1	0

Definimos também uma matriz Zcomo mostra a Tabela 12, com 3 linhas e 2 colunas que contêm as variáveis latentes relacionadas a cada observação, que são calculadas

durante o processo de estimação dos parâmetros.

Tabela 12: Matriz de variáveis latentes Z

	Componente 1	Componente 2
Observação 1	2_{11}	2_{12}
Observação 2	2_{21}	2_{22}
Observação 3	2_{31}	2_{32}

Vamos considerar 4 situações onde em cada uma destas situações comparamos a matriz A dada na Tabela 11 com cada um dos 4 possíveis resultados da matriz Z fornecidos pela Tabela 13.

Tabela 13: Exemplos de matriz Z

Situação 1	Situação 2	Situação 3	Situação 4
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$

Através de (6.1) e (6.2) utilizando a primeira linha da matriz A dada na Tabela 11 e as primeiras linhas das matrizes Z dadas na Tabela 13 vamos descrever os resultados obtidos para cada situação.

Situação 1:

$$n_d = |(0-1)| + |(1-0)| + |(1-1)| = 2$$

$$i_d = \frac{3 \acute{=} xi3 \ o(2; 3-2)}{3} = \frac{2}{3} = 0c66.$$

O valor de $i_d = 0.66$ representa que nesta iteração, o modelo separou corretamente 66% das observações.

Situação 2:

$$n_d = |(0-0)| + |(1-1)| + |(1-0)| = 1$$

$$i_d = \frac{3 \acute{=} xi3 \ o(1; 3-1)}{3} = \frac{2}{3} = 0\text{c}66.$$

Situação 3:

$$n_d = |(0-1)| + |(1-0)| + |(1-0)| = 3$$

$$i_d = \frac{3 \acute{=} xi3 \ o(3; 3-3)}{3} = \frac{3}{3} = 1.$$

Note que o modelo separou corretamente todas as observações, pois separou a observação 1 em uma componente e alocou as observações 2 e 3 na outra componente.

Situação 4:

$$n_d = |(0-0)| + |(1-1)| + |(1-1)| = 0$$

$$i_d = \frac{3 \acute{=} xi3 \ o(0; 3-0)}{3} = \frac{3}{3} = 1.$$

Nesta iteração o modelo separou corretamente as observações em 100% dos casos.

Para calcular a proporção de acertos do modelo, vamos considerar as quatro situações citadas. Através de (6.3), o valor de P_a é dado por

$$P_a = \frac{0\text{c}66 + 0\text{c}66 + 1 + 1}{4} = 0\text{c}83.$$

Assim, observando o valor dado por P_a temos que o modelo separou corretamente as observações em média em 83% das situações.

6.3 Divergência de *Kullback-Leibler*

Apresentamos uma maneira de medir a similaridade entre duas distribuições que simulam o comportamento de proteínas. Esta medida é chamada de divergência de *Kullback-*

Leibler.

A divergência de *Kullback-Leibler* pode ser usada para quantificar quão próximas estão duas distribuições de probabilidades $p_1(x)$ e $p_2(x)$ sendo que, a divergência entre $p_1(x)$ e $p_2(x)$ com respeito a $p_1(x)$ é definida como

$$\text{g u}(p_1(x)\text{c}p_2(x)) = \sum_{todo\ x} p_1(x)\ln(p_1(x)\text{'}\ p_2(x)). \tag{6.4}$$

Temos que a divergência de *Kullback-Leibler* dada em (6.4) não é simétrica pois

$$\text{g u}(p_1(x)\text{c}p_2(x)) \neq \text{g u}(p_2(x)\text{c}p_1(x))\text{c}$$

desta maneira, para termos uma divergência simétrica fazemos

$$\text{g u}_s(p_1(x)\text{c}p_2(x)) = \frac{1}{2}(\text{g u}(p_1(x)\text{c}p_2(x)) + \text{g u}(p_2(x)\text{c}p_1(x)))\text{c} \tag{6.5}$$

sendo desta forma g $\text{u}_s(p_1(x)\text{c}p_2(x))$ a divergência de *Kullback-Leibler* média entre

$$\text{g u}(p_1(x)\text{c}p_2(x)) \text{ e g u}(p_2(x)\text{c}p_1(x)).$$

6.4 Aplicações da análise de performance

Nesta seção, analisamos a performance em dados simulados para testar a eficiência do método empregado. Nesta análise, verificamos para qual comprimento de seqüências de moléculas o modelo identifica e separa corretamente as proteínas e, utilizando a divergência de *Kullback Leibler*, verificamos qual deve ser a medida de divergência entre as distribuições, que simulam proteínas, para que haja uma separação correta das observações em relação às componentes..

Duas simulações são feitas para testarmos a performance do método considerando, nestas simulações duas componentes na mistura.

Verificamos a eficiência do modelo, usando o índice de referência proposto na seção

6.2.

Para executarmos esta análise, procedemos da seguinte forma, geramos 3 observações sendo que f $_1$ e f $_2 \sim$ ' $u,tino3\ i{=},(N\mathrm{c}\mathrm{T}_1)$ e f $_3 \sim$ ' $u,tino3\ i{=},(N\mathrm{c}\mathrm{T}_2)$, onde

$$\begin{cases} \mathrm{T}_1 = \mathrm{k}(\mathrm{w}_{11}\mathrm{cw}_{12}\mathrm{c...cw}_{1k}) + (1-\mathrm{k})(\mathrm{w}_{21}\mathrm{cw}_{22}\mathrm{c...cw}_{2k}) \\ \mathrm{T}_2 = (1-\mathrm{k})(\mathrm{w}_{11}\mathrm{cw}_{12}\mathrm{c...cw}_{1k}) + \mathrm{k}(\mathrm{w}_{21}\mathrm{cw}_{22}\mathrm{c...cw}_{2k}) \end{cases} \mathrm{c} \tag{6.6}$$

para k $\in [0.5 : 1.0]$.

Iniciamos a análise com k $= 0\mathrm{c}5$ e desta forma, $\mathrm{T}_1 \neq \mathrm{T}_2$ e as distribuições em (6.6) e (6.7) são idênticas. Assim, as observações f $_1\mathrm{cf}_2$ e f $_3$ podem ser vistas como sendo observações geradas de uma mesma componente. Gradativamente, incrementamos o valor de k até chegarmos em k $= 1$. Quando k $= 1\mathrm{c}$as observações f $_1$ e f $_2$ são geradas de uma componente e f $_3$ é gerada de outra. Para cada incremento de k, geramos as observações utilizando comprimentos diferentes de seqüências e medimos as divergências de *Kullback Leibler* entre as distribuições das observações geradas.

6.4.1 Caso 1: Bases nitrogenadas

Para o caso 1, as três observações foram geradas como é citada em 6.4 sendo que

$$\begin{cases} \mathrm{T}_1 = \mathrm{k}(\mathrm{w}_{11}\mathrm{cw}_{12}\mathrm{cw}_{13}\mathrm{cw}_{14}) + (1-\mathrm{k})(\mathrm{w}_{21}\mathrm{cw}_{22}\mathrm{cw}_{23}\mathrm{cw}_{24}) \\ \mathrm{T}_2 = (1-\mathrm{k})(\mathrm{w}_{11}\mathrm{cw}_{12}\mathrm{cw}_{13}\mathrm{cw}_{14}) + \mathrm{k}(\mathrm{w}_{21}\mathrm{cw}_{22}\mathrm{cw}_{23}\mathrm{cw}_{24}) \end{cases} \mathrm{c} \tag{6.7}$$

para k $\in [0\mathrm{c}5 : 1\mathrm{c}0]\mathrm{c}$onde os parâmetros $(\mathrm{w}_{11}\mathrm{cw}_{12}\mathrm{cw}_{13}\mathrm{cw}_{14})$ e $(\mathrm{w}_{21}\mathrm{cw}_{22}\mathrm{cw}_{23}\mathrm{cw}_{24})$ são os mesmos utilizados no exemplo 5.3.2. Estas observações simulam o comportamento protéico visto através das bases nitrogenadas.

Foram usados 6 diferentes comprimentos de seqüências (50,100,200,500,1000,2000) e 9 divergências de *Kullback-Leibler* (0;0,001;0,004;0,009;0,016;0,025;0,037;0,060;0,066).

Combinando cada comprimento de seqüência e cada divergência de *Kullback-Leibler* Simulamos 2000 observações multinomiais calculando o índice de referência.

Resultados para o caso 1

Executando a análise de performance obtemos os resultados exibidos nas Figuras 28, 29 e 30.

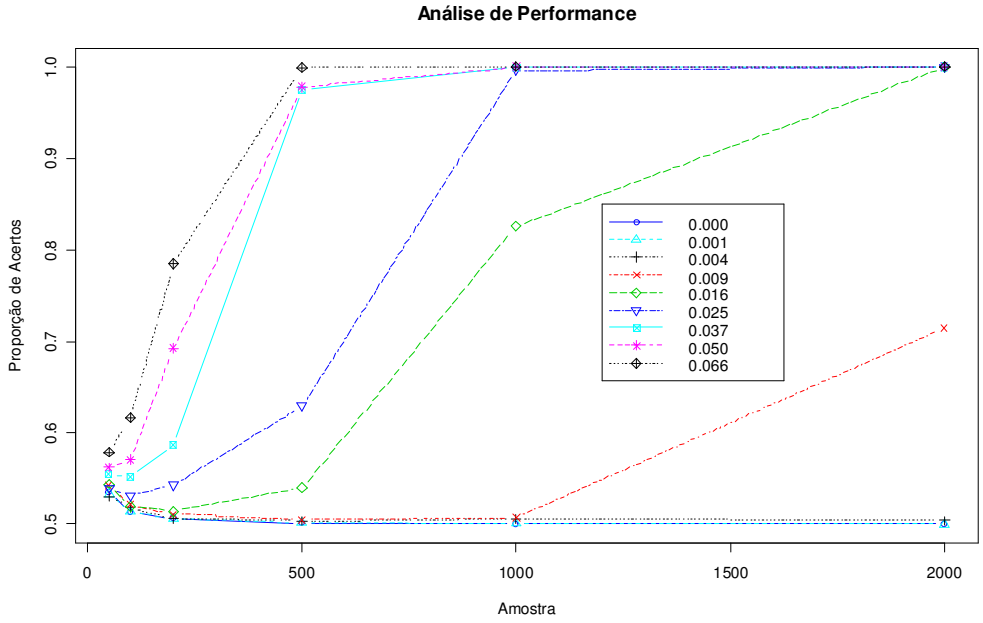


Figura 28: Comprimento das seqüências *versus* índice de referência

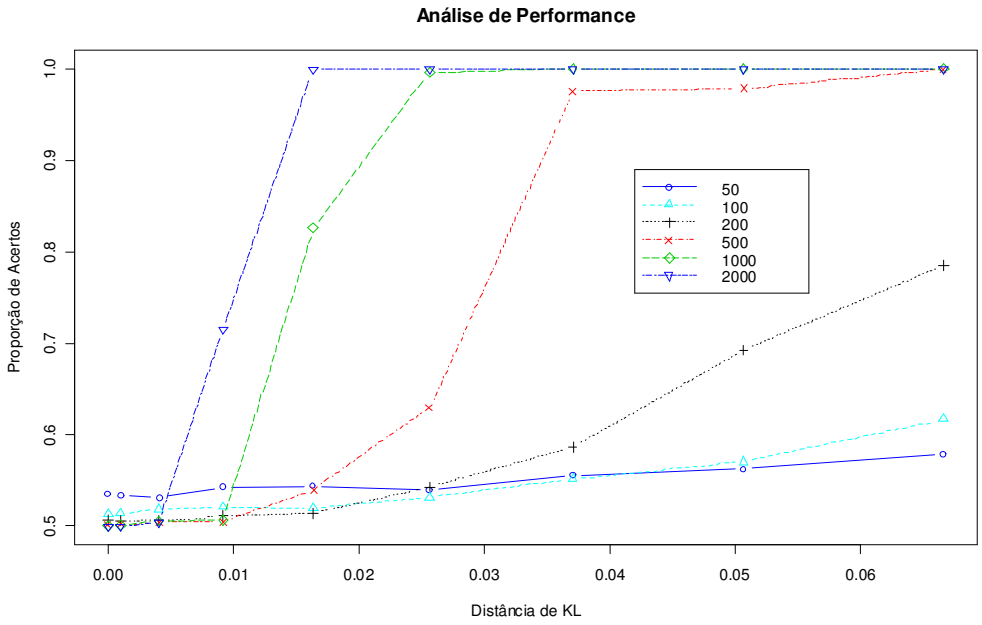


Figura 29: Divergência de *Kullback-Leibler versus* índice de referência

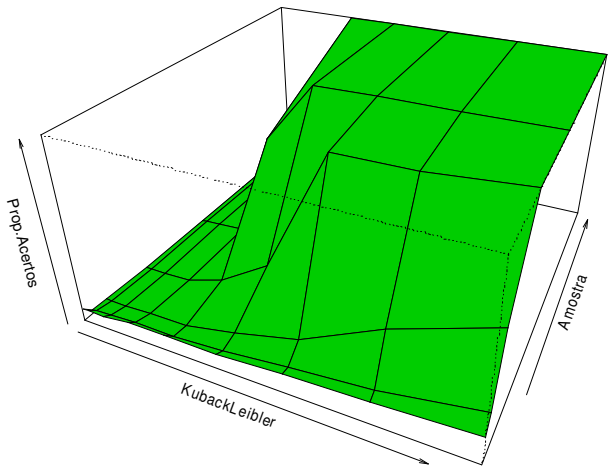


Figura 30: Gráfico em Perspectiva

Temos que a proporção de acertos, que é citado nas Figuras 28, 29 e 30, varia entre 0,5 e 1.

Quando o índice é próximo de 0,5, isto indica que o modelo não separou as proteínas corretamente em relação às componentes da mistura.

O índice de referência próximo de 1, mostra que o método conseguiu identificar as proteínas similares e dissimilares agrupando em uma mesma componente as proteínas similares e, separando em componentes distintas as proteínas dissimilares.

Na Figura 28 temos que o modelo não conseguiu separar de forma correta as proteínas que possuam seqüências de bases com comprimentos menores que 100. Para seqüências de tamanho 500, verificamos uma separação satisfatória apenas nas proteínas que possuam uma divergência de *Kullback-Leibler* acima de 0,037, sendo que para uma divergência de 0,066 a separação ocorre perfeitamente. A partir de um comprimento de tamanho 1000, o método separa com 100% de eficiência as proteínas que possuam uma divergência igual ou maior que 0,025 e, para comprimento igual a 2000 o método é eficiente em proteínas que possuam uma divergência igual ou maior que 0,016. Temos a informação, através da Figura 28 que o método é ineficiente para comprimento de seqüências de tamanho 2000 para proteínas que possuam uma divergência de *Kullback-Leibler* abaixo de 0,009.

Através da Figura 29, temos que o modelo só se mostra eficiente para proteínas que possuam uma divergência de 0,015 quando o comprimento das seqüências de bases nitrogenadas é igual ou superior a 2000. Quando a divergência de *Kullback-Leibler* é 0,025, é necessário um comprimento igual ou maior a 1000 para que o método seja eficiente. Para observações com uma divergência de 0,035, com um comprimento de tamanho 500 atinge-se resultados satisfatórios sendo que, com este comprimento, o método separa com 100% de acerto quando a divergência é de 0,065.

6.4.2 Caso 2: Aminoácidos

Nesta seção as três observações foram geradas como no caso anterior sendo que

$$\begin{cases} T_1 = k(w_{11}cw_{12}cw_{13}...cw_{119}cw_{120}) + (1-k)(w_{21}cw_{22}cw_{23}c...cw_{219}cw_{220}) \\ T_2 = (1-k)(w_{11}cw_{12}cw_{13}...cw_{119}cw_{120}) + k(w_{21}cw_{22}cw_{23}c...cw_{219}cw_{220}) \end{cases} c \qquad (6.8)$$

para $k \in [0c5 : 1.0]$. Temos que os parâmetro $(w_{11}cw_{12}c...cw_{120})$ e $(w_{21}cw_{22}c...cw_{220})$ são os mesmos utilizados no exemplo 5.3.4. no capítulo 5. As observações geradas simulam o comportamento de proteínas sendo vistas através dos aminoácidos.

Foram usados 6 diferentes comprimentos de seqüências (50,100,200,500,1000,2000) e 9 divergências de *Kullback-Leibler* (0;0,006;0,026;0,060;0,108;0,174;0,259;0,372;0,538). Simulamos 2000 observações multinomiais combinando cada comprimento de seqüência com cada divergência de *Kullback-Leibler* calculando o índice de referência para o modelo.

Resultados para o caso 2:

Nas Figuras 31, 32, e 33 exibimos os resultados gráficos obtidos.

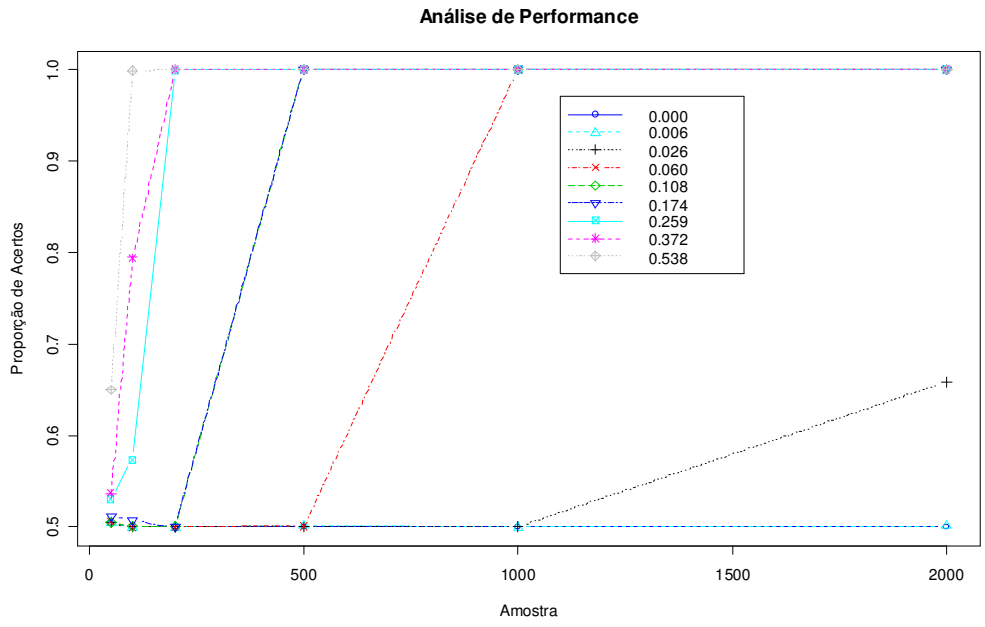


Figura 31: Comprimento das seqüências *versus* índice de referência

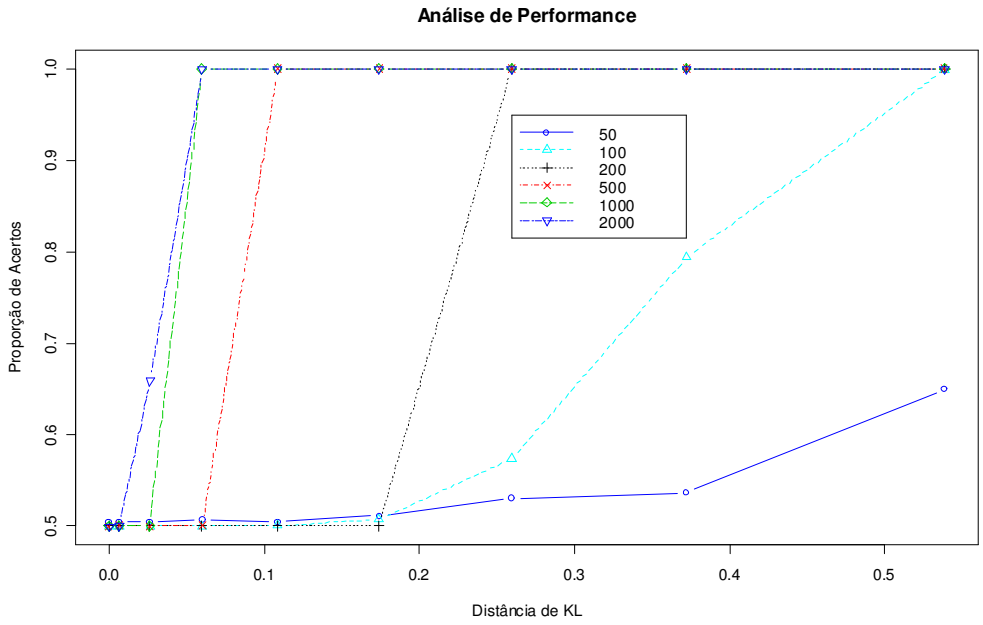


Figura 32: Divergência de *Kullback-Leibler versus* índice de referência

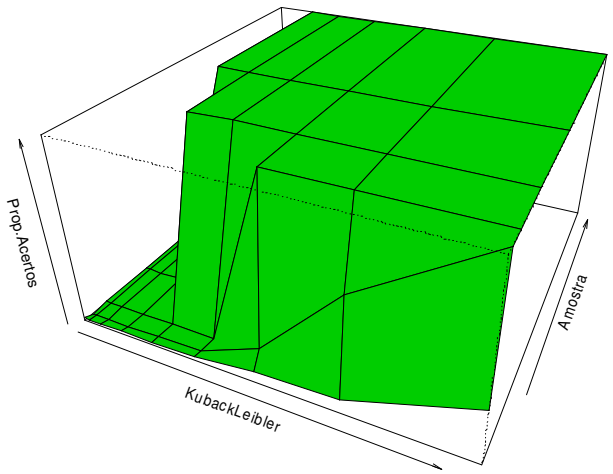


Figura 33: Gráfico em Perspectiva

Na Figura 31, podemos notar que, para um comprimento de seqüências de aminoácidos de tamanho menor que 50 o método se mostra eficiente apenas em proteínas com divergência de 0,538. O modelo é eficiente para comprimento de tamanho 100 para proteínas com uma divergência de *Kullback-Leibler* igual ou superior a 0,259. Para um comprimento de tamanho 500, o método separa com 100% de acertos as proteínas que possuam uma divergência igual ou maior que 0,174. Ao atingir um comprimento de tamanho 1000 para as seqüências de aminoácidos, a Figura 31 nos indica que o modelo é eficiente para proteínas com divergência igual ou superior a 0,060. O método foi incapaz de separar as proteínas com divergência de *Kullbach-Leibler* inferior a 0,026 em seqüências de tamanho 2000.

Temos a informação através da Figura 32 que, para proteínas que possuam uma divergência inferior ou igual a 0,1 o modelo é eficiente para seqüências de tamanhos maiores ou iguais a 500. Quando as proteínas possuem uma divergência de 0,25, a Figura 32 nos indica que o método proposto é eficiente para comprimentos iguais ou maiores a que 200. Com 100% de acertos o método separa proteínas com divergência de 0,5 e com seqüência de aminoácidos de tamanho igual ou superior a 100.

Capítulo 7

Seleção de modelos

7.1 Introdução

Neste trabalho, propomos uma metodologia que através de um modelo de mistura seja capaz de separar proteínas dissimilares e agrupar as similares. Ao propormos o método, consideramos que a população de onde provêm as observações é particionada em duas componentes representando a presença de dois grupos protéicos. Como o nosso interesse é modelar proteínas sem limitar o número de grupos protéicos ou limitar o número de componentes na população, faz-se necessário generalizarmos a metodologia proposta, ou seja, trabalharmos com um valor ' de componentes desconhecido.

Neste capítulo apresentamos dois métodos de seleção de modelos, utilizados por nós a fim de estimar o valor de ' e definir quantos grupos protéicos estão presentes em nossas análises. Os métodos estudados e aplicados são o fator de *Bayes* e o DIC, *deviance information criterion.*

7.2 Fator de *Bayes*

Em 1961, no livro *Theory of Probability*, Jefreys propõe um método estatístico que quantifica a evidência de um modelo em relação à outro através de um número. Este número

é conhecido como fator de *Bayes*.

O fator de *Bayes* é um instrumento utilizado na seleção de modelos e está relacionado ao teste de verossimilhança.

7.2.1 Definição do fator de *Bayes*

Suponha que um conjunto de observações t onde t $= (t_1 ct_2 c...ct_n)$ seja produzido sob um de dois modelos M_0 ou M_1de acordo com a densidade de probabilidade $p(t|M_0)$ ou $p(t|M_1)$.

Suponha também que sejam dadas as distribuições de probabilidade *a priori* $\pi(M_0)$ e $\pi(M_1)$. Assim, utilizando uma abordagem bayesiana, temos a produção das distribuições *a posteriori* $p(M_0|t)$ e $p(M_1|t)$.

Através do teorema de *Bayes* (4.8) temos para $k = 0c1$

$$p(\ _k|t) = \frac{p(t|\ _k)\pi(\ _k)}{p(t|\ _0)\pi(\ _0) + p(t|\ _1)\pi(\ _1)}.$$

Fazendo o quociente entre $p(\ _1|t)$ e $p(\ _0|t)$ temos

$$\frac{p(\ _1|t)}{p(\ _0|t)} = \frac{p(t|\ _1)}{p(t|\ _0)}\frac{\pi(\ _1)}{\pi(\ _0)}c$$

onde, o fator de *Bayes* ou FB_{10} é definido como

$$5\ B_{10} = \frac{p(t|\ _1)}{p(t|\ _0)}c \tag{7.1}$$

e pode ser interpretado como a evidência dos dados em favor de M_1 contra M_0.

Segundo Kass e Raftery (1995), em casos onde $p(t|\ _1)$ e $p(t|\ _0)$ são distribuições sem parâmetros livres $5\ B_{10}$ é denotado simplesmente como razão de verossimilhanças. Em casos onde existem parâmetros desconhecidos sob um ou sob os dois modelos, $5\ B_{10}$ continua sendo uma razão de verossimilhanças mas a densidade $p(t|\ _k)$, para $k = 0c1$,

é obtido através de integração sob o espaço paramétrico. Assim, para $k = 0c1$ temos

$$p(t\ |'\ _k) = \int p(t\ |w_k c'\ _k)\pi(w_k|'\ _k)dw_k c \tag{7.2}$$

onde w_k é o parâmetro ou vetor de parâmetro sob $'\ _k$, $\pi(w_k|'\ _k)$ é a distribuição *a priori* para $'\ _k$ e $p(t\ |w_k c'\ _k)$ é a função de verossimilhança de w_k.

Quando temos mais de dois modelos sendo verificados, podemos reescrever (7.1) como

$$5\ B_{ij} = \frac{p(t\ |'\ _i)}{p(t\ |'\ _j)} c$$

para $i \neq j$.

As densidades contínuas com espaço n-dimensional envolvidas no cálculo do fator de *Bayes*, freqüentemente, são complexas para se calcular analiticamente e desta forma são utilizadas aproximações.

Reescrevendo (7.2) temos

$$p(t\ ) = \int p(t\ |w)\pi(w)dw c \tag{7.3}$$

que, segundo Raftery e Banfield (1990) pode ser aproximada através do método de Monte Carlo fazendo

$$\widehat{p(t\ )} = \frac{1}{3}\sum_{i=1}^{m} p(t\ |w^{(i)}).$$

Segundo Geweke (1989), a precisão do método de Monte Carlo pode ser melhorada pelo método *Importance Sampling* que consiste em gerar $w^{(i)}$,$i = 1c2c...c3$ camostras de uma densidade $\pi^*(w)$ ponderada através de pesos. Desta forma $p(t\ )$ é aproximada por

$$\widehat{p(t\ )} = \frac{\sum_{i=1}^{m} w_i P(t\ |w^{(i)})}{\sum_{i=1}^{m} w_i} c \tag{7.4}$$

onde $w_i = \frac{\pi(\theta^{(i)})}{\pi^*(\theta^{(i)})}$.

Temos que através dos métodos *MCMC* obtemos uma amostra da densidade *a posteriori*

$$P(\mathrm{w}|\mathrm{t}\ ) = \frac{P(\mathrm{t}\ |\mathrm{w})\pi(\mathrm{w})}{P(\mathrm{t}\ )}. \tag{7.5}$$

Substituindo $\pi^*(\mathrm{w})$ por $\frac{P(Y|\theta)\pi(\theta)}{P(Y)}$ em (7.4) obtemos

$$\widehat{p(\mathrm{t}\ )} = \frac{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)})}{\pi^*(\theta^{(i)})} P(\mathrm{t}\ |\mathrm{w}^{(i)})}{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)})}{\pi^*(\theta^{(i)})}} =$$

$$\frac{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)})P(Y)}{P(Y\ \theta^{(i)})\pi(\theta^{(i)})} P(\mathrm{t}\ |\mathrm{w}^{(i)})}{\sum_{i=1}^{m} \frac{\pi(\theta^{(i)})P(Y)}{P(Y|\theta^{(i)})\pi(\theta^{(i)})}} =$$

$$\frac{\sum_{i=1}^{m} P(\mathrm{t}\ )}{\sum_{i=1}^{m} \frac{P(Y)}{P(Y\ \theta^{(i)})}} = \frac{3\ P(\mathrm{t}\ )}{P(\mathrm{t}\ ) \sum_{i=1}^{m} \frac{1}{P(Y|\theta^{(i)})}} =$$

$$\frac{3}{\sum_{i=1}^{m} \frac{1}{P(Y|\theta^{(i)})}} = \frac{3}{\sum_{i=1}^{m} (P(\mathrm{t}\ |\mathrm{w}^{(i)}))^{-1}} =$$

$$\left[\frac{1}{3} \sum_{i=1}^{m} (P(\mathrm{t}\ |\mathrm{w}^{(i)}))^{-1} \right]^{-1} \mathrm{c} \tag{7.6}$$

que é o inverso da média da verossimilhança invertida, ou seja, a média harmônica considerada como uma estimativa para $P(\mathrm{t}\ )$.

7.2.2 Aplicações do fator de *Bayes*

Fazemos duas aplicações do fator de *Bayes.* Para cada aplicação utilizamos 4 modelos com mistura de multinomiais sendo que a diferença entre estes modelos está no número de componentes na mistura. O primeiro modelo, ' $_1$cconsidera a população sem partições, ou seja, ' $_1$ é um modelo sem mistura de distribuições multinomiais. O modelo ' $_2$ considera a população particionada em duas componentes. No modelo ' $_3$ temos mistura de três

componentes e no modelo ’ $_4$ temos mistura de quatro componentes.

A função de verossimilhança do modelo ’ $_1$ é dada por

$$\mathrm{u}\,(\mathrm{w}|x) \propto \pi_{l=1}^{k} \mathrm{w}_l^{\sum_{i=1}^{n} x_{il}} \mathrm{c}$$

e as demais funções de verossimilhança são dadas como (4.10).

Na primeira aplicação, os dados são gerados como no exemplo 5.3.2 ou seja,

$$\mathrm{f}_{\ i} \sim \text{’} \quad u,tino3\ i{=},(300; \mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{cw}_{j3}\mathrm{cw}_{j4})\mathrm{c}$$

para $i = 1\mathrm{c}2\mathrm{c}3$ e $j = 1\mathrm{c}2$ componentes, simulando o comportamento de proteínas observadas através de bases nitrogenadas. Temos que 300 representa o comprimento da seqüência de bases.

Na segunda aplicação geramos os dados simulando o comportamento de proteínas observadas através dos aminoácidos como no exemplo 5.3.4, ou seja,

$$\mathrm{f}_{\ i} \sim \text{’} \quad u,tino3\ i{=},(1000; \mathrm{w}_{j1}\mathrm{cw}_{j2}\mathrm{cw}_{j3}\mathrm{c}...\mathrm{cw}_{j18}\mathrm{cw}_{j19}\mathrm{cw}_{j20})\mathrm{c}$$

onde $i = 1\mathrm{c}2\mathrm{c}3\mathrm{c}\, j = 1\mathrm{c}2$ componentes, e 1000 representa o comprimento da seqüência de aminoácidos.

Tanto na primeira aplicação quanto na segunda, duas observações foram geradas utilizando as proporções das moléculas vindas de um grupo protéico e a outra observação foi gerada utilizando as proporções das moléculas vindas de outro grupo protéico como é dito na seção 5.2. A aplicação do fator de *Bayes* nos permite avaliar a precisão na identificação do número de componentes nos modelos com mistura sendo que as componentes representam os grupos protéicos presentes na população.

Os parâmetros são gerados e estimados através das distribuições condicionais (5.7), (5.8), (4.17) e (4.18) utilizando o algoritmo *Gibbs Sampling* descrito na seção 4.5.2. São feitas 50000 iterações com um descarte inicial de 500 valores e com saltos de 10 em 10 foram selecionados os valores restantes.

7.2.3 Resultados das aplicações do fator de *Bayes*

Como é dito na seção 7.2.2, o fator de *Bayes* é calculado como

$$B_{ij} = \frac{\text{'}_i}{\text{'}_j}$$

onde $\text{'}_i = \int p((\,|\mathrm{w_i c'}_i)\pi(\mathrm{w_i}|\text{'}_i)d\mathrm{w_i}$ e $\text{'}_j = \int p((\,|\mathrm{w_j c'}_j)\pi(\mathrm{w_j}|\text{'}_j)d\mathrm{w_j}$ para $i\mathrm{c}j = 1\mathrm{c}2\mathrm{c}3\mathrm{c}4$ sendo que $i \neq j$.

Neste cálculo, se o resultado for maior que 1, o modelo escolhido é o modelo '_i e se o resultado for menor que 1, o modelo escolhido é o modelo '_j .

Na Tabela 14 temos os resultados do cálculo de $\text{'}_i = \int p((\,|\mathrm{w_i c'}_i)\pi(\mathrm{w_i}|\text{'}_i)d\mathrm{w_i}$, sendo $i = 1\mathrm{c}2\mathrm{c}3\mathrm{c}4\mathrm{c}$ para as duas aplicações feitas.

Tabela 14: Resultados dos cálculos de R_i

Modelo	R_i : Bases	R_i : Aminoácidos
'_1	$8{,}70 \times 10^{-152}$	$2{,}08 \times 10^{-272}$
'_2	$1{,}95 \times 10^{-147}$	$7{,}69 \times 10^{-142}$
'_3	$1{,}59 \times 10^{-148}$	$8{,}88 \times 10^{-270}$
'_4	$2{,}50 \times 10^{-148}$	$1{,}32 \times 10^{-269}$

Os resultados dos cálculos do fator de *Bayes*, são exibidos na Tabela 15 para as duas aplicações.

Tabela 15: Resultados dos cálculos do fator de *Bayes*

Modelo	Bases	Aminoácidos
B_{12}	$4{,}45\times10^{-5}$	$2{,}70\times10^{-131}$
B_{13}	0,00054	0,002
B_{14}	0,00034	0,001
B_{23}	12,28	$8{,}66\times10^{127}$
B_{24}	7,81	$5{,}82\times10^{127}$
B_{34}	0,63	0,672

Segundo a Tabela 15, nas duas aplicações, temos que o modelo com duas componentes, ’ $_2$ é sempre escolhido quando comparado aos demais e é o mais favorecido numericamente. Quando comparamos os modelos ’ $_1$ e ’ $_3$ o modelo ’ $_3$ é o indicado como melhor, quando comparamos os modelos ’ $_1$ e ’ $_4$ temos que o modelo 4 é o melhor, quando comparamos ’ $_3$ e ’ $_4$ temos que o modelo ’ $_3$ é indicado como o melhor. Nota-se que o modelo ’ $_1$ não é apontado como melhor modelo em nenhuma das comparações.

Desta forma podemos afirmar que o fator de Bayes identificou o melhor modelo, ou seja ’ $_2$ com bastante precisão.

7.3 DIC: *deviance information criterion*

Uma forma de se comparar modelos é combinar uma medida de ajuste, que pode ser a deviance, com uma medida de complexidade do modelo que pode ser o número de parâmetros. Esta forma ocorre no AIC, *Akaike informatiom criterion*, Akaike (1973) e no DIC, *deviance information criterion,* Spiegelhalter (2002), entre outros métodos de seleção de modelos.

Apresentamos nesta seção uma proposta para seleção de modelos utilizando uma estrutura *bayesiana* através de um balanceamento entre as medidas de ajuste e de complexidade que é o DIC.

7.3.1 Definição do DIC

Dempster (1974), sugeriu que considerar a distribuição *a posteriori* da log verossimilhança dos dados é equivalente a examinar a distribuição *a posteriori* de

$$D(\mathrm{w}) = -2\log P((\,|\mathrm{w}) + 2\log s((\,)\mathrm{c}$$

onde s (( ) é uma função apenas dos dados e $D(\mathrm{w})$ é chamada de *deviance* bayesiana.

A distribuição *a posteriori* de $D(\mathrm{w})$ é baseada em $P(\mathrm{w}|(\,)$ onde

$$P(\mathrm{w}|(\,) \quad P((\,|\mathrm{w})P(\mathrm{w}).$$

Segundo Spiegelhalter (2002), podemos verificar o ajuste de um modelo através da esperança *a posteriori* da *deviance,*

$$\overline{D} = .\ _{\theta\, y}[D(\mathrm{w})]\mathrm{c} \tag{7.7}$$

e a medida de complexidade de um modelo P_D, que é o número de parâmetros efetivos no modelo, é definido como sendo a esperança *a posteriori* da *deviance* menos a *deviance* calculada nos valores dos parâmetros estimados *a posteriori*,

$$P_D = .\ _{\theta\, y}[D] - D(.\ _{\theta|y}[\mathrm{w}]) =$$

$$\overline{D} - D(\overline{\mathrm{w}}). \tag{7.8}$$

Utilizando estes conceitos, Spiegelhalter (2002), mostra que modelos podem ser comparados utilizado uma *deviance information criterion* definida como

$$DIC = \overline{D} + P_D = D(\overline{\mathrm{w}}) + 2P_D. \tag{7.9}$$

O *DIC* pode ser calculado durante o processo de estimação dos parâmetros utilizando

métodos *MCMC* monitorando os parâmetros estimados w sendo que, a cada iteração do processo calcula-se $D(w)$. No final das iterações, basta tomar a média amostral dos valores de $D(w)$ menos a estimativa da *deviance* calculada usando as médias amostrais dos valores simulados w.

O modelo que possui o melhor ajuste é aquele que apresenta o menor valor de DIC.

7.3.2 Aplicações do DIC

São feitas duas aplicações do *DIC*. Os modelos e os dados foram os mesmos que os modelos e os dados utilizados nas aplicações do fator de *Bayes* na seção 7.2.2.

Como na seção 7.2.2, são feitas 50000 iterações com um descarte inicial de 500 valores e com saltos de 10 em 10 foram selecionados os valores restantes. Os parâmetros são gerados e estimados através das distribuições condicionais (5.7), (5.8), (4.17) e (4.18) utilizando o algoritmo *Gibbs Sampling* descrito na seção 4.5.2.

7.3.3 Resultados das aplicações do *DIC*

Utilizando (7.7), (7.8) e (7.9) obtemos os resultados expostos nas Tabelas 16 e 17.

Tabela 16: Resultados para aplicação do *DIC* em bases nitrogenadas

Modelos	$\overline{D}$	$D(\overline{w})$	P_D	DIC
$'_1$	42,25	39,18	3,06	45,31
$'_2$	31,06	22,45	8,60	39,67
$'_3$	35,72	30,43	5,29	41,01
$'_4$	40,12	36,41	3,71	43,83

Na Tabela 16, temos na primeira coluna respectivamente os modelos sem mistura, com duas componentes, com três componentes e com quatro componentes. Na segunda coluna temos as médias das *deviances,* a terceira coluna apresenta as *deviances* calculadas para os valores médios estimados. Na quarta coluna temos o valor aproximado do número

efetivo de parâmetros para cada modelo, e na quinta coluna temos os valores calculados do *DIC* para cada um dos modelos.

Nesta Tabela temos os resultados do *DIC* utilizando as proporções de bases nitrogenadas como observações. Temos que o melhor modelo é o modelo com duas componentes pois é o que apresenta o menor valor de *DIC,* ou seja, 39,67 . Em seguida temos o modelo com três componentes com *DIC* igual a 41,01 seguido do modelo com quatro componentes com um *DIC* igual a 43,83. O pior modelo é o modelo sem mistura apresentando um *DIC* igual a 45,31.

Tabela 17: Resultados para aplicação do *DIC* em aminoácidos

Modelos	$\overline{D}$	$D(\overline{\mathrm{w}})$	P_D	DIC
’ $_1$	692,64	675,14	17,50	710,14
’ $_2$	142,76	93,98	48,78	191,54
’ $_3$	167,00	115,23	51,76	218,76
’ $_4$	186,89	133,82	53,06	239,95

Na Tabela 17 temos os resultados do *DIC* utilizando as proporções de aminoácidos como observações para os modelos utilizados. Temos que o melhor modelo é o modelo com duas componentes, com *DIC* igual a 191,54. O segundo melhor modelo é o modelo com três componentes com um *DIC* igual a 218,76. Em seguida, com quatro componentes temos um *DIC* igual a 239,95, e o pior modelo, com um *DIC* igual a 710,14, temos o modelo sem mistura de distribuições.

Notamos que tanto na Tabela 16 quanto na Tabela 17, sem contar os modelos sem mistura, obtivemos valores de P_D distante do número efetivo de parâmetros nos modelos. Nos modelos sem mistura, e conseqüentemente sem variáveis latentes, a aproximação de P_D é boa. A razão para isto pode estar na presença das variáveis latentes no modelo empregado.

Nesta dissertação, todos os programas e gráficos foram feitos no software estatístico R que está disponível gratuitamente no site http://www.r-project.org/.

Capítulo 8

Discussões e Conclusões

As proteínas são componentes fundamentais para os seres vivos pois executam as mais variadas funções que vão desde a formação das estruturas celulares até o processamento de materiais através de reações químicas. Dada esta diversidade de funções, um aspecto importante da pesquisa pós-genoma é identificar a função, ou funções, desempenhada pelas proteínas. Esta função está diretamente relacionada à forma estrutural da mesma, que por sua vez está relacionada a sua composição. As proteínas são compostas por unidades básicas que são os aminoácidos sendo que, estes aminoácidos são codificados pelas bases nitrogenadas. Temos que proteínas com composição similar tendem a possuir estruturas similares e, portanto, funções similares.

O foco da pesquisa aqui relatada é adaptar uma metodologia estatística para identificar proteínas com funções similares a partir de sua composição. A metodologia adapta o modelo de mistura com variáveis latentes para usá-lo como identificação de grupos de proteínas similares.

No capítulo 6 desenvolvemos um estudo de simulação para verificar a performance da metodologia proposta. Neste estudo, medimos quão bem o modelo identifica proteínas similares. Para tanto, propomos um índice de referência que mede a concordância entre a situação real utilizada na geração dos dados e a estimativa produzida pelo método.

Quando utilizamos seqüências de bases nitrogenadas e, empregamos as proporções

destas bases no modelo de mistura, os resultados obtidos são bastante satisfatórios quando temos seqüências com comprimento igual ou maior que 1000 e com distribuições com divergência igual ou maior que 0.04. Isto é comprovado no exemplo 5.1 na Tabela 7, onde as estimativas encontradas para os parâmetros das distribuições multinomiais do modelo são próximas das proporções com que as observações foram geradas. As estimativas das ponderações exibidas na Tabela 8 expressam o fato de que uma componente que gere mais observações tenha uma ponderação maior. Com estes tipos de observações também temos bons resultados na análise de performance feitas no capítulo 6 exibida no caso 1.

Quando as proporções de aminoácidos que são empregadas no modelo, os resultados são satisfatórios quando temos seqüências com comprimento igual ou maior que 1000 e com distribuições com divergência igual ou maior que 0.25. Na Tabela 9 do capítulo 5 temos que as estimativas dos parâmetros variam desde estimativas precisas até estimativas não tão boas, mas, de uma forma geral, não apresentaram valores muito distantes dos valores reais. Já os valores das estimativas das ponderações exibidas na Tabela 10, apresentaram resultados não condizentes com o que se esperava, pois o método forneceu uma ponderação maior para a componente que gerou menos observações. Em relação à análise de performance, os resultados obtidos foram semelhantes ao caso das seqüências de bases nitrogenadas sendo que, utilizando aminoácidos a separação feita pelo modelo é mais rápida. Consideramos este resultado preliminar e estudos mais aprofundados devem ser realizados no futuro para confirmar estes resultados e identificar suas causas.

No capítulo 7, aplicamos as técnicas de seleção de modelos fator de *Bayes* e DIC, *deviance information criterion,* tanto para modelos onde as observações são as proporções de bases nitrogenadas quanto para modelos onde as observações são as proporções de aminoácidos com o objetivo de estimar o número de componentes nos modelos. Nos dois casos obtivemos resultados bastante satisfatórios nos modelos utilizados, ou seja, os métodos de seleção de modelos conseguiram detectar o número de grupos funcionais de proteínas na amostra submetida aos métodos.

O trabalho realizado nesta dissertação tem como ponto positivo a aplicabilidade em

situações reais. Em vários casos de pesquisas de curas para doenças ou de novos medicamentos as proteínas estão envolvidas. Existem situações onde não se conhece a função da proteína e se sabe apenas que determinado fragmento de sua estrutura é similar a outra. Neste ambiente o modelo proposto pode ser um apoio relevante para encontrar a função protéica. A idéia básica é submeter um grupo de proteínas contendo proteínas com funções desconhecidas juntamente a proteínas com funções conhecidas e estimar a função das proteínas desconhecidas através do grupo a que são atribuídas.

Bibliografia

[1] Akaike, H.1973. Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B.N. Petrov and F.Csáki), pp. 267-281. Budapest: Akadémiai Kiadó.

[2] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. 2002. *Fundamentos da biologia celular, uma introdução à biologia molecular da célula.* Artmed Editora S.A. Porto Alegre RS

[3] Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal Royal Statistics Society.* **36**, 192-236.

[4] Brown, T.A. 1999. *Genome.* Willey, New York.

[5] Brown, M.P., Hughey, R., Krogh, A., Mian, I.S., Sjöjander, K., Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. Em Hunter, L., Searls, D., Shavlik, J., *ISMB*-93, Menlo Park, CA. AAAI/MIT Press 47-55.

[6] Casella, G., Berger, R.L. 1990. *Statistical Inference.* Duxbury Press, California.

[7] Cowles, M.K., Carlin, B.P. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal American Statistics Association.* ***91****, 883-904.*

[8] Dempster, A.P. 1974. The direct use of likelihood for significance testing. In *Proccedings of Conference on Foundational Questions in Statistical Inference*. 335-352. Department of Theoretical Statistics: University of Aarhus.

[9] Diebolt, J., Robert, C.P. 1994. Estimation of finite mixture distribution through bayesian sampling. *Journal Royal Statistics Society*, série B. **56**, n. 2, 363-375.

[10] Gelfand, A.E., Smith, A.F.M. 1990. Sampling based approaches to calculating marginal densities. *Journal American Statistics Association.* **87**, 523-531.

[11] Gelman, A.B., Carlim, J.S., Stern, H.S., Rubin, D.B. 1995. *Bayesian Data Analysis.* Chapman & Hall/CRC, New York.

[12] Gelman, A.B., Rubin, D.B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science.* **7**, 457-511.

[13] Gelman, A.B., Rubin, D.B. 1992. A single series from the Gibbs sampler provides a false sense of secuty. *Bayesian Statistics.* **4**, 625-631.

[14] Geman, S., Geman, D. 1984. Stochastic relaxation, Gibbs distribution and the bayesian retoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **6**, 721-741.

[15] Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrika. **57**, 1317-1340.

[16] Geweke, J. 1992. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. *Bayesian Statistics.* **4**, 169-193.

[17] Gilks, W.R., Richardson, S., Spiegelhalter, D.J. 1996. *Markov Chain Monte Carlo in practice.* Chapman & Hall/CRC, New York.

[18] Hastings, W.K.,1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* **57**, 97-109.

[19] Jeffreys, H. 1961. *Theory of probability*. London, Oxford University Press.

[20] Karplus, K. 1995. Regularizers for estimating distributions of amino acids from small samples. Technical Report UCSC-CRL-95-11, University of California, Santa Cruz.

[21] Kass, R.E., Raftery, A.E. 1995. Bayes factor. *American Statistical Association*. **90**, N° 430, 773-795.

[22] Lehninger, A.L., Nelson, D.L., Cox, M.M. 1993. *Principles of Biochemistry*. Worth Publishers.

[23] Levine, R.R.J. 1981. Sex differences in schizophrenia: timing or subtypes?. *Psychological* Bulletin. **90**, 432-444.

[24] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal Chem. Phys.*, **21**, 1087-1092.

[25] Mengersen, K. L. e Robert, C. P. 1996. Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics*, 5 (Alicante, 1994), pages 255–276. Oxford University Press, New York

[26] Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions A*. **185**,

71-110.

[27] Raftery, A.E., Banfield, J.D. 1990. Stopping the Gibbs sampler the use of morphology and other issues in spatial statistics. *Annal of the Institute of Statistical Mathematics*, **43**, 32-43.

[28] Raftery, A.E., Lewis, S.M. 1992. How many iterations in Gibbs sampler? *Bayesian Statistics*. **4**, 763-773.

[29] Ritter, C., Tanner, M.A. 1992. Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. Journal American Statistics Association. **87**, 861-868.

[30] Robert, C. P. e Soubiran, C. 1993. Estimation of a normal mixture model through Gibbs sampling and prior feedback. *Test*, **2**(1-2):125–146.

[31] Robert, C. P. 1996. Mixtures of distributions: inference and estimation. In Markov chain *Monte Carlo in practice*, pages 441–464. Chapman and Hall, London.

[32] Roeder, K. e Wasserman, L. 1995. Bayesian density estimation using mixtures of normals. *Relatório Técnico* 09, Department of Statistics, Carnegie Mellon University.

[33] Sjölander, K., Karplus, K., Brown, M., Hughey, R, Krogh, A., Mian, I.S., Haussler, D. 1996. Dirichlet mixture: a method for improved detection of weak but significant protein sequence homology. *CABIOS*. **12**, 327-345.

[34] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A. 2002. Bayesian measures of model complexity and fit. *J.R. Statist. Soc. B* 64, *part* 4, 583-639.

[35] Stryer, L. 1995. *Biochemistry*. W.H. Freeman and Company, New York.

[36] Titterington, D.M., Smith, A.F.M., Makov, U.E. 1987. *Statistical analysis of finite mixture distributions*. Wiley, New York.

[37] Winkler, R.L., Hays, W.L. 1975. *Statistics: Probability, Inference and Decision*. Holt, Rinehart and Winston, New York.