FUNDAMENTOS E APLICAÇÕES DOS CRITÉRIOS DE INFORMAÇÃO: AKAIKE E BAYESIANO

PAULO CÉSAR EMILIANO

2009

PAULO CÉSAR EMILIANO

FUNDAMENTOS E APLICAÇÕES DOS CRITÉRIOS DE INFORMAÇÃO: AKAIKE E BAYESIANO

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

Orientador

Prof. Dr. Mário Javier Ferrua Vivanco

Co-orientador

Prof. Dr. Fortunato Silva de Menezes

LAVRAS

MINAS GERAIS-BRASIL

2009

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Emiliano, Paulo César.
Fundamentos e aplicações dos critérios de informação: Akaike e Bayesiano / Paulo César Emiliano. – Lavras : UFLA, 2009.
92 p. : il.

Dissertação (Mestrado) – Universidade Federal de Lavras, 2009.
Orientador: Mário Javier Ferrua Vivanco.
Bibliografia.

1. Critério de Informação de Akaike. 2. Entropia . 3. Critério de Informação de Schwarz. 4. Informação de Kullback-Leibler 5. Seleção de Modelos. I. Universidade Federal de Lavras. II. Título.

CDD – 536.73

PAULO CÉSAR EMILIANO

FUNDAMENTOS E APLICAÇÕES DOS CRITÉRIOS DE INFORMAÇÃO: AKAIKE E BAYESIANO

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 19 de fevereiro de 2009.

Prof. Dr. Fortunato Silva de Menezes	UFLA
Prof. Dr. Marcelo Angelo Cirillo	UFLA
Prof. Dr. Telde Natel Custódio	UFSJ

Prof. Dr. Mário Javier Ferrua Vivanco
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

Aos meus pais, Francisco e Alzira ,
que souberam conduzir com
muita sabedoria a minha
formação.

" If you have an apple and I have an apple and we exchange apples
then you and I still have one apple.
But if you have an idea and I have an idea and we exchange these ideas,
then each of us will have two ideas."
George Bernard Shaw

AGRADECIMENTOS

Primeiramente a Deus, que deu-me forças em todos os momentos de minha vida, e a Nossa Senhora Aparecida, que sempre intercede por mim e da qual sou devoto.

Meus sinceros agradecimentos ao professor Mário Javier Ferrua Vivanco, pela paciência com que me orientou, disponibilidade em auxiliar-me a qualquer momento, pelas críticas e sugestões.

Aos meus pais, Francisco e Alzira, pela confiança, compreensão, carinho, apoio e tudo que sou devo a eles.

Aos meus irmãos Rosemeire e Washington, pelo carinho, compreensão e torcida em todos os momentos.

A todos os colegas de mestrado e doutorado em Estatística, em especial ao Ed Carlos, Altemir, Ricardo, Augusto, Tânia, Patrícia, Denise, Ana Paula, Isabel, Hiron, Stephânia e Richardson.

Aos meus professores Hélia, grande amiga e companheira, que ensinou-me a entender o que aquelas letrinhas significavam quando eu tinha seis anos, e até hoje eu não esqueci; ao professor William por introduzir-me ao mundo maravilhoso da matemática, de uma forma que apaixonei-me por ela; à professora Cássia, pelos freqüentes incentivos que dava à nossa turma acreditando em nós e incentivando-nos.

A todos da Escola Estadual Santa Tereza, professores, "tias" da cantina, amigos, que foram fundamentais em minha formação.

A todos da Universidade Federal de Viçosa, que de uma forma ou de outra contribuíram para a realização deste trabalho. Em especial aos professores Olímpio, Margareth e Paulo Tadeu, a quem muito admiro e que foi muito importante na

consolidação do meu conhecimento em matemática.

Aos funcionários do Departamento de Ciências Exatas: Edila, Josi, Joyce, Maria, Selminha e Vânia, pela simpatia e boa vontade no atendimento.

Aos professores do Departamento de Ciências Exatas, pelos ensinamentos prestados.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, pela oportunidade da realização deste curso.

À FAPEMIG, pela bolsa de estudos, essencial para a realização deste trabalho.

Aos demais que, direta ou indiretamente, contribuíram para a elaboração deste trabalho.

Paulo César Emiliano

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Emiliano, Paulo César. **Fundamentos e Aplicações dos Critérios de Informação:** Akaike e Bayesiano. 2009. 92p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras. *

Objetivou-se com este estudo apresentar os fundamentos do critério de informação de Akaike (AIC) e do critério de informação Bayesiano (BIC), amplamente utilizados na seleção de modelos, e geralmente pouco entendidos. A seleção de modelos é de vital importância em estudos científicos, devendo portanto estar embasada em princípios científicos concretos, como a parcimônia. O AIC e o BIC são critérios que penalizam a verossimilhança, para que um modelo mais parcimonioso seja selecionado. Estes critérios baseiam-se nos conceitos de informação e entropia, que são fundamentais para o completo entendimento dos mesmos. Procurou-se explicar tais conceitos para que o entendimento desses critérios fosse completo. Também foram dadas duas aplicações do AIC e BIC, em regressão e na seleção de modelos normais. Os resultados obtidos ao utilizar-se os dois critérios foram os mesmos para as duas aplicações feitas, e embora os mesmos modelos tenham sido selecionados, o AIC e o BIC não necessariamente proporcionam os mesmos resultados.

Palavras-chave: Critério de Informação de Akaike, Entropia, Critério de Informação de Schwarz, Informação de Kullback-Leibler, Seleção de Modelos.

* **Comitê Orientador:** Mário Javier Ferrua Vivanco - UFLA (Orientador), Fortunato Silva de Menezes (Co-orientador)

ABSTRACT

Emiliano, Paulo César. **Fundamentals and Applications Criteria for Information:** Akaike and Bayesian. 2009. 92p. Dissertation (Master in Statistics and Agricultural Experimentation) Federal University of Lavras, Lavras.*

This study presented the foundations of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion. (BIC), largely used in the selection of models, and usually little understood. The selection of models is essential in scientific studies, consequently, it should be based on solid scientific foundations, as the parsimony. The AIC and BIC are criteria that punish the likelihood, so that a more parsimonious model is selected. These criteria are based on concepts of information and entropy, that are fundamental for their complete understanding. It was tried to explain such concepts in order to make the understanding of these criteria complete and clear. Two applications of AIC and BIC were Also given, both in regression and in the selection of normal models. The results obtained when using the two methods were the same for the two done applications. But although the same models have been selected -AIC and BIC- they do not necessarily provide the same results.

Key-words: Akaike Information Criterion, Bayesian Information Criterion, Entropy, Kullback-Leibler Information, Model Selection.

*Guindance Committee: Mário Javier Ferrua Vivanco - UFLA. (Adviser), Fortunato Silva de Menezes - UFLA. (Co-Adviser)

1 INTRODUÇÃO

Muitas pessoas têm o dom da ciência, são cientistas e tentam entender os fenômenos que há muito intrigam os homens. Porém, a maioria da população não estuda estes fenômenos, seja porque os acha complicados demais ou porque não têm acesso à informação para entendê-los. Cabe, pois, aos cientistas levar a informação e explicar os fenômenos a estas pessoas da forma mais simples possível.

Em geral um fenômeno em estudo pode ser explicado através de um modelo. Os modelos são os principais instrumentos utilizados na estatística. Eles são uma versão simplificada de algum problema ou situação da vida real e destinam-se a ilustrar certos aspectos do problema, sem contudo, se ater a todos os detalhes.

Geralmente os fenômenos observados são muito complexos e é impraticável descrever tudo aquilo que é observado com total exatidão. Dificilmente consegue-se traduzir em simbologias e fórmulas matemáticas aquilo que é visto com perfeita exatidão. Se isto for possível, deve-se ao fato do fenômeno ser perfeitamente conhecido e um modelo determinístico o explica. Um modelo determinístico é estabelecido quando tudo relacionado ao fenômeno em estudo é conhecido, e por isso ele é, exatamente o mecanismo de geração dos dados obtidos no estudo.

Mas em situações práticas o total conhecimento do fenômeno não acontece, o que torna impossível descrever o mesmo através de um modelo determinístico. Faz-se uso então dos modelos estatísticos, aqueles em que há uma parte sistemática e outra parte aleatória, como por exemplo, os modelos lineares generalizados. Neste tipo de modelo, não se pode determinar quais dados serão obtidos antecipadamente, mas o conjunto do qual os resultados são obtidos é usualmente conhecido. Ao se aproximar um fenômeno por um modelo probabilístico, haverá perda de informação ao fazer-se tal modelagem, sendo que esta perda deve ser mínima

para não comprometer o entendimento do fenômeno em estudo.

Não raro, tem-se mais de um modelo para descrever o mesmo fenômeno, haja vista que não há uma receita a ser seguida, tendo cada pesquisador a liberdade de modelar o fenômeno seguindo a metodologia que julgar mais adequada. Desse modo, ao se deparar com dois (ou mais modelos) é natural questionar: "Dentre estes modelos qual deles é o mais adequado?". O conceito de melhor modelo é controverso, mas um bom modelo deve conseguir equilibrar a qualidade do ajuste e a complexidade, sendo esta, em geral, medida pelo número de parâmetros presentes no modelo; quanto mais parâmetros, mais complexo o modelo, sendo pois mais difícil interpretar o modelo. A seleção do "melhor" modelo torna-se então evidente.

Burnham & Anderson (2004), enfatizam a importância de selecionar modelos baseados em princípios científicos. Diversas são as metodologias utilizadas para selecionar modelos tais como C_p de Mallows, Regressão Stepwise, Critério de Informação de Akaike (AIC), Critério de Informação Bayesiano (BIC), Critério de Informação Generalizado (GIC), dentre outros.

As metodologias acima citadas, baseiam-se nos conceitos de Informação e Entropia. Estes conceitos são de fundamental importância para que se possa ter completo entendimento dos critérios AIC e BIC, que serão objetos de estudo neste trabalho.

Nos critérios AIC e BIC cada modelo dá um valor e o modelo que apresentar o menor valor AIC (ou BIC) é considerado como o "melhor" modelo. Um questionamento natural que se faz é: "Por que o Critério com menor AIC (ou BIC) é selecionado?".

Objetivou-se com este trabalho explicar, ilustrar e comparar os critérios AIC e BIC, amplamente utilizados para a seleção de modelos e por vezes pouco entendi-

dos. Através de algumas aplicações, espera-se que a metodologia destes critérios seja entendida para que, ao se utilizar tais critérios, tenha-se perfeita consciência do resultado obtido e se saiba interpretá-lo com total segurança.

2 REFERENCIAL TEÓRICO

Nesta seção, serão apresentados alguns conceitos que serão úteis para atingir o objetivo proposto neste trabalho.

2.1 Modelos

Em estudos nas mais diversas áreas, tais como ciências sociais, epidemiologia, zootecnia, etc, há vários aspectos que são não determinísticos. Assim sendo, modelos puramente matemáticos não são adequados para modelar esse tipo de estudo. Um caminho para a modelagem de fenômenos não determinísticos são os modelos probabilísticos.

De acordo com Stevenson (2001), um modelo é uma versão simplificada de algum problema ou situação da vida real destinado a ilustrar certos aspectos do mesmo sem levar em conta todos os detalhes. Além disso, o modelo permite checar se sua forma funcional está representando bem o fenômeno em estudo, sem porém deixar de levar em conta o conhecimento do pesquisador acerca do assunto.

Para fenômenos complexos*, é bastante raro ter só um modelo plausível, mas vários para escolher um dentre eles. Em tais situações, a seleção do modelo se torna um problema fundamental. Porém Ghosh & Samanta (2001), afirmam que para muitos cientistas, modelos são sinônimos de paradigmas. Assim, o problema de escolher um modelo só aparece quando aquela ciência estiver nas encruzilhadas. Por exemplo, quando físicos tinham que escolher entre a gravitação na Teoria Clássica de Newton e a gravitação na Teoria da relatividade de Einstein.

Na estatística clássica, normalmente a seleção de modelos é feita na fase de análise exploratória dos dados. Uma análise cuidadosa de dados deve sempre con-

*Aqueles em que há muitas variáveis interferindo no modelo, sendo estas muitas das vezes desconhecidas

siderar o problema de determinação do modelo, isto é, o problema da avaliação e escolha do modelo que melhor represente a situação em estudo (Miranda, 2006). Todo subseqüente estatístico depende da análise do modelo selecionado.

Ocasionalmente, há estudos de sensibilidade da análise subseqüente com respeito ao modelo selecionado. Porém, a estatística, em geral, não enfatiza a seleção de modelos, nem dá uma devida certeza acerca do modelo que é assumido através de convenção ou seleção por análise exploratória. Entretanto, há certas áreas da estatística clássica em que a seleção do modelo desempenha um papel importante, como por exemplo, regressão linear e séries temporais. Assim, o problema torna-se de seleção de modelos (Ghosh & Samanta, 2001).

De acordo com Mazerolle (2004), seleção de modelo é a tarefa de escolher um modelo estatístico de um conjunto de modelos plausíveis. Em sua forma mais básica, esta é uma das tarefas fundamentais das pesquisas científicas. Dos tantos modelos plausíveis que poderiam ser ajustados aos dados, como pode-se escolher um bom modelo?. A modelagem estatística geralmente decide entre um conjunto de possíveis modelos, conjunto este que deve ser selecionado pelo pesquisador. Freqüentemente, modelos simples, como polinômios, são usados como ponto de partida. Burnham & Anderson (2004) enfatizam a importância de selecionar modelos com bace em princípios científicos.

Ao se estudar um fenômeno, o conhecimento prévio que o pesquisador tem acerca deste é de fundamental importância e deve ser levada em conta. Porém, este deve embasar-se também em outros princípios científicos para fazer sustentar suas conclusões acerca do fenômeno. De acordo com Mazerolle (2004), três princípios regulam nossa capacidade de fazer inferência nas ciências:

1- Simplicidade e parcimônia

Sugerem que a explicação mais simples é passível de ser a mais provável.

2- Trabalhando Hipóteses

A seleção de modelos traduz-se em testar para os dados em mãos uma série de modelos plausíveis.

3- O poder da evidência

Dá uma indicação de qual modelo é o melhor entre os modelos testados, e o poder do teste para cada modelo.

Conforme Mazerolle (2004), seria ingênuo esperar que os melhores resultados incluam todas as variáveis no modelo. Isto viola o princípio científico fundamentado na parcimônia, que requer que dentre todos os modelos que expliquem bem os dados, deve-se escolher o mais simples. Assim, deve-se conciliar um modelo mais simples, mas que explique bem o fenômeno em estudo.

Segundo Konishi & Kitagawa (2008), uma vez que o conjunto de possíveis modelos foi selecionado, a análise matemática permite determinar o melhor destes modelos. O significado de "melhor" é controverso. Uma boa técnica de seleção de modelos equilibrará qualidade do ajuste e complexidade. Modelos mais complexos poderão melhor adaptar sua forma para ajustar-se aos dados (por exemplo, um polinômio de quinta-ordem pode ajustar exatamente seis pontos), mas muitos parâmetros podem não representar nada útil ou explicável.

De acordo com Mazerolle (2004), a qualidade do ajuste é geralmente determinada usando-se razão de verossimilhanças ou uma aproximação dela, conduzindo a um teste qui-quadrado. A complexidade é geralmente medida contando o número de parâmetros inclusos no modelo. Entretanto, antes de se construir modelos (por exemplo, um modelo de regressão linear ou qualquer outro modelo generalizado) deve-se ter em mente que não existem modelos verdadeiros. Tem-se apenas modelos aproximados da realidade. O que se faz então é minimizar a perda de

informações. George Box fez uma famosa afirmativa acerca disso: “Todos os modelos são errados, mas alguns são úteis”[1].

2.2 Informação

A palavra informação vem do latim “informare”, dar forma, pôr em forma ou aparência, criar, representar, apresentar, criar uma idéia ou noção, algo que é colocado em forma, em ordem. Como se pode ver, informação é um termo altamente polissêmico (que tem vários significados) (Ribeiro, 2008).

Segundo Ribeiro (2008), a teoria da informação é um ramo do conhecimento humano cujos objetivos envolvem a conceituação matemática do termo informação e a construção de modelos capazes de descrever os processos de comunicação. O artigo “A Mathematical Theory of Communications”, publicado por Claude Shannon em 1948, lançou as bases para a moderna teoria das comunicações Shannon (1948), apud Ribeiro, (2008). Qualquer processo de comunicação envolve transferência de informação entre dois ou mais pontos. Segundo Fernandes & Azevedo (2006), o problema fundamental das comunicações é o de reproduzir em um ponto, exatamente ou aproximadamente, uma mensagem selecionada em um outro ponto.

De acordo com Shannon (1948) apud Ribeiro (2008), um sistema de comunicação consiste de 5 partes:

1- Uma fonte de informação que produz uma mensagem ou seqüência de mensagens a serem comunicadas ao terminal receptor;

2- Um transmissor (codificador) que opera na mensagem de modo que esta possa ser transmitida sobre o canal;

3- Um canal que é o meio pelo qual a informação será transmitida. Este meio

[1]Tradução nossa. “All models are wrong but some are useful”(Draper & Smith, 1998)

contém ruído (em casos ideais o ruído é desconsiderado) e irá alterar de alguma forma a mensagem original;

4- O receptor (decodificador), que apenas faz a função inversa do transmissor de modo a obter a mensagem original;

5- O destino, para quem a mensagem é encaminhada.

Esquematicamente, tem-se a Figura 1 abaixo (Ash, 1965):

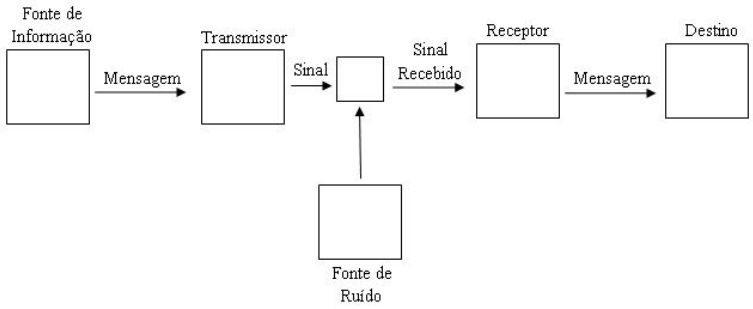

FIGURA 1: Modelo esquemático de um sistema geral de comunicação.

Segundo Shannon (1948) apud Ribeiro (2008), uma fonte de informação é um elemento participante do processo de comunicação que produz informação, enquanto que o destinatário é o elemento que recebe a informação produzida por essa fonte. Em uma conversação os participantes costumeiramente se revezam nos papéis de fonte e destinatário, e a informação circula na forma de palavras, possivelmente selecionadas de um vocabulário conhecido por todo o grupo.

Se um português disser a um polaco “Bom dia”, provavelmente não haverá transmissão de informação entre os dois. No entanto, se o português disser “Dzien dobry”, provavelmente o polaco irá retribuir com um sorriso, pois entendeu a saudação. Logo, para que haja transmissão de informação, o código usado na comunicação tem de ser perceptível por ambas as partes.

Segundo Ash (1965), um conjunto de palavras-código capaz de representar todas as saídas possíveis de uma fonte constitui um código para a fonte de informação. Codificadores são elementos (seres humanos, circuitos, programas, etc), que representam as mensagens geradas pela fonte empregando um código específico. Um decodificador é responsável por desfazer o mapeamento realizado por um codificador.

De acordo com Ash (1965), Shannon desenvolveu a teoria da informação e transmissão de sinais digitais baseados em seqüências de zeros e uns. É aí que define o problema fundamental da comunicação como o de "reproduzir num local, de forma aproximada ou exata, uma mensagem selecionada noutro local". Assim estabeleceu-se então o esquema de transmissão de informação, hoje clássico, com uma mensagem que parte de uma fonte, é codificada e emitida por um transmissor, passa por um canal de comunicação, sofre perturbações designadas por ruídos, e chega depois ao receptor, passando por um sistema de decodificação. Ao falar de "uma mensagem selecionada", Shannon refere-se a uma seqüência informativa que pode ser escolhida dentre muitas outras que aparecerão com iguais ou diferentes probabilidades. Define então a quantidade de informação com base na sua incerteza ou dificuldade de previsão.

Supondo, por exemplo, que um emissor transmita a mensagem "bom dia", letra por letra, ao emitir as primeiras letras, há uma expectativa da parte do receptor, que vê surgir as letras "b", "o", "m", um espaço, e depois o "d" e o "i". O "a" final é quase inútil, pois sua probabilidade de ocorrência é tão grande, para dar sentido à seqüência anterior, que a quantidade de informação transmitida por essa letra é muito menor que a transmitida pelas primeiras. Assim, quanto menor é a incerteza ou dificuldade de previsão, menor é a quantidade de informação, e vice-versa (Ash, 1965).

Se, por exemplo, houver o evento X="O sol nasce", a resposta à pergunta "O sol nascerá hoje?" não traz nenhuma informação; entretanto, se fez a pergunta "O Cruzeiro será o campeão mundial de 2009?" Como isso é pouco provável, uma resposta positiva a essa pergunta oferece uma quantidade de informação muito maior que divulgar uma resposta negativa. Assim, eventos improváveis contém mais informações do que os eventos mais prováveis (Ribeiro, 2008).

De acordo com Fernandes & Azevedo (2006), a teoria da informação de Shannon é apropriada para medir incerteza sobre um espaço desordenado, isto é, ela é útil para analisar variáveis qualitativas nominais, tais como sexo, raça, etc., pois não é possível uma ordenação dos seus resultados. Neste sentido não é possível definir uma distância entre os elementos do espaço, tais como a distância entre o sexo masculino e o sexo feminino.

A noção de distância, acima referida, pode ser entendida a partir da seguinte definição (Domingues, 1982):

Definição 2.1 *Dado um conjunto $M \neq \emptyset$ seja $d : M \times M \longrightarrow \mathbb{R}_+$ e indique-se por $d(x, y)$ a imagem de um par genérico $(x, y) \in M \times M$, através da função d. Diz-se que d é uma **distância** sobre M se as seguintes condições se verificam:*

$$d(x, y) = 0 \Longleftrightarrow x = y, \forall x, y \in M \tag{2.1}$$

$$d(x, y) = d(y, x), \forall x, y \in M \tag{2.2}$$

$$d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in M \tag{2.3}$$

Por exemplo, a informação de Shannon é usada sobre um espaço de letras do alfabeto, já que letras não têm "distâncias" entre elas, não sendo possível quantificar o quanto a letra "m" se distancia da letra "e".

De acordo com Bolfarine & Sandoval (2000), uma medida alternativa de in-

formação foi criada por Fisher, para medir incerteza sobre um espaço ordenado, isto é, a informação de Fisher pode ser usada para variáveis qualitativas ordinais que permitem uma ordenação dos seus resultados (tais como conceitos finais em uma disciplina, peso de pessoas, etc.). Para informação sobre valores de parâmetros contínuos, como as alturas de pessoas, a informação de Fisher é usada, já que tamanhos estimados têm uma distância bem definida.

Conforme Bolfarine & Sandoval (2000), a informação de Fisher é assim definida:

Definição 2.2 *A quantidade*

$$I_F(\theta) = E\left[\left(\frac{\partial \log\left(f\left(X|\theta\right)\right)}{\partial\theta}\right)^2\right]$$

é denominada informação de Fisher de θ.

Se há uma amostra aleatória $X_1, X_2, \ldots X_n$, da variável aleatória X com função de densidade de probabilidade $f(x|\theta)$ e informação de Fisher $I_F(\theta)$, a informação total de Fisher de θ correspondente à amostra observada é a soma da informação de Fisher das n observações da amostra, isto é,

$$E\left[\left(\frac{\partial \log L\left(\theta|\boldsymbol{X}\right)}{\partial\theta}\right)^2\right] = nI_F\left(\theta\right),$$

em que $\log L\left(X|\theta\right)$ é a função de log verossimilhança, que será definida em 2.21.

Sabendo como a informação é gerada, como se pode medir quanta informação é produzida? Como quantificar uma determinada mensagem recebida? Com propósito de responder estas perguntas considere-se a situação abaixo descrita em Silva (2008):

Exemplo

Um sistema deve transmitir o estado do tempo. Suponha que se classifica o tempo da seguinte forma: limpo, nublado, chuvoso e nevoeiro. Define-se informação como a quantidade de incerteza que o receptor tem acerca da mensagem que está recebendo. Por exemplo, suponha que o receptor conhece as seguintes probabilidades para o estado do tempo:

Estado do tempo	*Probabilidade*
Limpo	*0.65*
Nublado	*0.20*
Chuvoso	*0.10*
Nevoeiro	*0.05*

Como a probabilidade do tempo estar limpo é grande, na maioria das vezes, o tempo está limpo, e ao se dizer que ele está limpo transmite-se pouca informação. Por outro lado, ao se dizer que ele está com nevoeiro, trata-se de uma situação pouco freqüente, e portanto, transmite-se muita informação.

De acordo com as probabilidades conhecidas, uma seqüencia típica de transmissão diária poderia ser: "limpo limpo limpo limpo limpo nublado nublado chuvoso limpo". Se for usado o seguinte código binário para codificar as mensagens:

Estado do tempo	*Código*
Limpo	*00*
Nublado	*01*
Chuvoso	*10*
Nevoeiro	*11*

a mensagem acima referida é codificada da seguinte forma: "00 00 00 00 00 01 01 10 00", ou seja, o número de "bits" necessários para transmitir é 18.

O número de "bits" necessários para codificar uma determinada informação segue uma relação inversa à probabilidade de ocorrência do evento. Assim quanto

maior for a probabilidade de ocorrência do evento transmitido, (quanto menor a informação transmitida), menos "bits" serão necessários para codificá-la, e quanto menor a probabilidade de ocorrência do evento (maior informação), mais "bits" serão necessários para codificá-la.

Nesta forma de transmissão haverá uma compressão dos dados que acarreta perda de uma pequena parte da informação que foi originalmente transmitida.

Segundo Kawada (1987) apud Konishi & Kitagawa (2008), para quantificar a informação perdida ao ajustarmos um modelo, existem diversas medidas propostas na literatura. Como exemplo tem-se:

1- A Estatística de χ^2, dada por:

$$\chi^2 = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i}.$$

2- A distância de Hellinger, dada por:

$$I_K(g; f) = \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx.$$

3- A informação generalizada, dada por:

$$I_\lambda(g; f) = \frac{1}{\lambda} \int \left\{ \left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right\} g(x)\, dx. \tag{2.4}$$

4- O critério Deviance, dado por:

$$D(\psi) = -2 \left[\log L(\psi; x) - \log L\left(\widehat{\psi}; x \right) \right],$$

em que ψ é o espaço paramétrico e $\widehat{\psi}$ é o espaço restrito.

5- A divergência, dada por:

$$D\left(g;f\right)=\int u\left(t(x)\right)g\left(x\right)dx=\int u\left(\frac{g\left(x\right)}{f\left(x\right)}\right)g\left(x\right)dx, \tag{2.5}$$

sendo que $t(x)=\dfrac{g\left(x\right)}{f\left(x\right)}$.

6- A L^1-norm, dada por:

$$L_1\left(g;f\right)=\int\left|g\left(x\right)-f\left(x\right)\right|dx.$$

7- A L^2-norm, dada por:

$$L_2\left(g;f\right)=\int\left\{g\left(x\right)-f\left(x\right)\right\}^2dx.$$

8- A Informação de Kullback-Leibler, dada por:

$$I\left(g;f\right)=E_g\left[\log\left(\frac{g\left(X\right)}{f\left(X\right)}\right)\right]=\int_{-\infty}^{+\infty}g\left(x\right)\log\left(\frac{g\left(x\right)}{f\left(x\right)}\right)dx, \tag{2.6}$$

sendo f, g f_i e g_i são funções de distribuição quaisquer, $\lambda\in\mathbb{R}_+^*$ e $u(x)$ uma função tal que $u:\mathbb{R}\rightarrow\mathbb{R}_+^*$.

Se em (2.6), $g(x)$ é a "verdadeira" distribuição, ou seja, $g(x)$ é o modelo determinístico, do qual verdadeiramente são gerados os dados (raramente conhecido devido à complexidade do fenômeno) e $f(x)$ for o nosso modelo estatístico selecionado para modelar o fenômeno, o valor da informação de Kullback - Leibler é uma quantificação da similaridade entre nosso modelo estatístico e a "verdadeira" distribuição.

Conforme Mazerolle (2004), Kullback e Leibler definiram esta medida, posteriormente chamada Informação de Kullback-Leibler (K-L) para representar a

informação perdida pela aproximação de nosso modelo da realidade.

De acordo com Konishi & Kitagawa(2008), vale a pena observar que se na equação (2.4) se fizer $\lambda \to 0$ e sob certas condições de regularidade, será obtida a informação de Kullback-Leibler; de fato:

$$\begin{aligned}
\lim_{\lambda\to 0} I_\lambda(g;f) &= \lim_{\lambda\to 0}\frac{1}{\lambda}\int\left\{\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}g(x)\,dx \\
&= \int \lim_{\lambda\to 0}\left[\frac{1}{\lambda}\left\{\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}g(x)\right]dx \\
&= \int g(x)\lim_{\lambda\to 0}\left[\frac{1}{\lambda}\left\{\left(\frac{g(x)}{f(x)}\right)^{\lambda}-1\right\}\right]dx \\
&\overset{L'Hospital}{=} \int g(x)\lim_{\lambda\to 0}\left[\left(\frac{g(x)}{f(x)}\right)^{\lambda}\ln\left(\frac{g(x)}{f(x)}\right)\right]dx \\
&= \int g(x)\ln\left(\frac{g(x)}{f(x)}\right)dx = I(g;f).
\end{aligned}$$

Além disso, se em (2.5), tomar-se $u(x) = log(x)$ encontrar-e-á também a informação de Kullback-Leibler, isto é, ela é um caso especial da divergência. De fato:

$$D(g;f) = \int u\left(\frac{g(x)}{f(x)}\right)g(x)\,dx = \int \log\left(\frac{g(x)}{f(x)}\right)g(x)\,dx = I(g;f).$$

2.2.1 A informação de Kullback-Leibler

Seja X uma variável aleatória discreta com distribuição de probabilidades $p(X)$. De acordo com Ribeiro (2008), Shannon definiu a quantidade de informação associada à ocorrência do evento X_i como:

$$I(X_i) = log\left(\frac{1}{p_i}\right) = -log\,(p_i), \tag{2.7}$$

em que p_i é a probabilidade de ocorrência do evento X_i. A função definida em (2.7) indica o total de conhecimento sobre o resultado de um certo evento, assim como intuitivamente esperava-se, um evento menos provável tem mais informação que outro mais provável. Se o logaritmo tiver base 2, o conteúdo da informação será expresso em $bits$. Se a base do logaritmo é e, então o conteúdo da informação é medido em $nuts$ e finalmente se a base for 10 o conteúdo da informação será medido em $hartley$. Nesse trabalho, é utilizada a base e, pois a informação com a qual Kullback e Leibler trabalham é definida nessa base, porém em alguns exemplos a base 2, também será utilizada.

A utilização do log na função definida por Shannon pode ser explicada facilmente no caso de acontecimentos eqüiprováveis. Por exemplo, se o número de símbolos que constituem o alfabeto é M, então o número de $bits$, N, necessários para representar todos os M símbolos é: $M = 2^N$, sendo $N = \log_2 M$. No caso de símbolos eqüiprováveis: $p\left(s_i\right) = \frac{1}{M}$, logo são necessários $N = \log_2 \frac{1}{p(s_i)}$, $bits$ para representar cada símbolo.

Considere-se uma fonte S cujas saídas são seqüências de elementos selecionados de um conjunto $A = \{a_0, a_1, a_2, \ldots, a_n\}$. Esse conjunto é o alfabeto da fonte e os seus elementos $a_i, i = 0, 1, 2, \ldots, n$, são denominados letras ou símbolos (Ribeiro, 2008). Considerando-se que os símbolos emitidos pela fonte são estatisticamente independentes entre si, estamos na presença de uma fonte sem memória. Nesse caso, a fonte fica completamente descrita pelo seu alfabeto A e pelas probabilidades de ocorrência dos símbolos do alfabeto fonte:

$$P = \{p(a_0), p(a_1), p(a_2), ..., p(a_n)\}, \text{ sendo que } \sum_{i=1}^{n} p(a_i) = 1.$$

A ocorrência do símbolo a_i significa a geração de $I(a_i) = log_2 \frac{1}{p(a_i)}$ $bits$ de

informação.

Como exemplo considere o arremesso de uma moeda em que $P(cara) = \frac{1}{4}$ e $P\,(coroa) = \frac{3}{4}$. Assim o conteúdo da informação é:

$$I\,(cara) = -log_2\left(\frac{1}{4}\right) = 2bits \;\in\; I\,(coroa) = -log_2\left(\frac{3}{4}\right) = 0,41bits.$$

Sendo X e Y dois eventos, é desejável que a função de informação tenha algumas propriedades (Shannon, 1948):

1- Se $P(X = x) = 0$ ou $P(X = x) = 1$, então $I(X) = 0$;

2- Se $0 < P(X = x) < 1$, então $I(X) > 0$;

3- Se $P(X = x) < P(Y = y)$, então $I(X) > I(Y)$;

4- Se X e Y são eventos independentes, então $I(X, Y) = I(X) + I(Y)$.

Em seu artigo publicado em (1948), Shannon demonstrou que só existe uma função, satisfazendo as pressuposições acima:

$$I(X) = -K\sum_{i=1}^{n} p_i \log p_i$$

em que $K > 0$ e $I(X)$ é uma medida de incerteza contida na variável aleatória.

A função $H = -\sum_{i=1}^{n} p_i \log p_i$ (a constante K é meramente uma constante que só depende da unidade de medida) desempenha um papel central na Teoria da Informação, sendo uma medida de incerteza contida na variável aleatória. A função I pode ser transformada na função entropia, definida em certas formulações de mecânica-estatística em que p_i é a probabilidade do sistema estar na fase i. A quantidade I é, por exemplo, a constante do famoso teorema de Boltzmann (Young & Freedman, 2003). Aqui, a quantidade $H = -\sum_{i=1}^{n} p_i \log p_i$ será chamada de **entropia** do conjunto de probabilidades $p_1, p_2, ..., p_n$.

A informação de Kullback-Leibler baseia-se na $Entropia$ de variáveis aleatórias.

2.2.2 Entropia

Entropia (do grego entropé) é uma medida da quantidade de desordem de um sistema.

2.2.2.1 Visão física da entropia

Fisicamente, o conceito de entropia está intimamente associado a conceitos da termodinâmica. Nas linhas a seguir falar-se-á um pouco mais acerca deste assunto.

Segundo Halliday et al. (1996), a energia é um dos conceitos da física com aplicação mais visível no dia-a-dia. Para mover um carro, por exemplo, é necessário obter energia através da queima do combustível. Para os eletrodomésticos funcionarem, depende-se da energia elétrica. O primeiro princípio da termodinâmica ocupa-se do estudo da energia e da sua conservação. Contudo, nem toda a energia gerada está disponível para ser transformada em trabalho útil. Existem processos que só acontecem em um sentido. Segundo o Dicionário Aurélio, que reflete o nosso linguajar coloquial, algo é reversível quando se pode reverter, ou se pode retornar ao estado inicial. Silva (2005), afirma que em Física, um processo é reversível quando pode partir do estado final e alcançar o estado inicial usando os mesmos micro-estados que utilizou para alcançar o estado final. Um livro deslizando sobre uma mesa terá sua energia mecânica convertida em calor; porém o processo inverso jamais foi visto por alguém (um livro que repousasse sobre uma mesa começasse a se mover espontâneamente e a temperatura do livro e da mesa diminuíssem); estes são os processos irreversíveis. O Segundo Princípio da Termodinâmica trata desta questão, assim como das possíveis maneiras de

transformar calor em trabalho (Halliday et al., 1996).

O Segundo Princípio da Termodinâmica apresentado por Kelvin-Planck é o seguinte: "É impossível construir uma máquina térmica que, operando em ciclo, não produza nenhum efeito além da absorção de calor de um reservatório e da realização de uma quantidade igual de trabalho" (Young & Freedman, 2003). Em sua essência, diz que é impossível construir uma máquina que trabalhe com rendimento de 100%. Para saber o quanto da energia pode ser considerada disponível para consumo, é necessário conhecer um outro conceito: o de entropia.

Segundo Silva (2008a), o conceito físico de entropia surgiu na época da máquina a vapor, proposto pelo prussiano Rudolf Emmanuel Clausius (1822-1888), para explicar o máximo de energia que poderia ser transformada em trabalho útil. Tal conceito é definido como (Halliday et al., 1996):

Definição 2.3 *Entropia S é uma propriedade cuja variação dS, no decurso de uma transformação elementar, internamente reversível, de um sistema fechado, se obtém dividindo a quantidade de calor dQ, que o sistema troca nessa transformação, pela temperatura absoluta T a que o sistema se encontra nesse momento. Isto é:*

$$dS = \left(\frac{dQ}{T}\right)_{rev}.$$

Tudo o que se disse acerca da entropia não é suficiente para compreender o verdadeiro significado físico dessa propriedade. Para tal tem-se que recorrer ao método utilizado na termodinâmica estatística, que faz uso da natureza microscópica da matéria para explicar as suas propriedades macroscópicas (Young & Freedman, 2003). A entropia pode ser considerada como uma medida da desordem molecular ou aleatoriedade molecular.

Tendo como referência um sistema de partículas, o conceito de entropia ganha com Boltzmann uma nova conotação. A entropia passa a ser entendida como uma

medida da distribuição das partículas em termos de posição espacial e quantidade de movimento. Aqui, máxima entropia passa a significar distribuição homogênea ou mínima desordem, quando a probabilidade de uma certa partícula se encontrar em uma determinada posição, com uma certa quantidade de movimento é idêntica à probabilidade de qualquer outra partícula específica se encontrar na mesma situação.

De acordo com Nussenzveig (1981), tem-se a seguinte definição de entropia no sentido estatístico de Boltzmann:

Definição 2.4 *A entropia é dada pela equação*

$$S = k\left[\log W\right]$$

em que k é uma constante (unidade termodinâmica da medida da entropia - Constante de Boltzmann) e W é o número de microestados de entropia S (é o número total de estados microscópicos compatível com o estado macroscópico do sistema).

Assim, a variação da entropia de um estado i para um estado j é

$$H_B = S_i - S_j = k\log\left(\frac{W_i}{W_j}\right), \tag{2.8}$$

em que H_B é a variação da entropia de Boltzmann, S_i e S_j são as entropias no estado i e j, respectivamente e W_i e W_j são números de microestados compatíveis com a ocorrência dos macroestados i e j, respectivamente.

Sendo $p(x)$ e $q(x)$ as funções densidades dos estados i e j respectivamente, pode-se reescrever (2.8) como:

$$H_B = k\log\left(\frac{p(x)}{q(x)}\right). \tag{2.9}$$

Conforme Nussenzveig (1981), como fundador da Mecânica Estatística (Huang, 1987), Boltzmann propôs sucessivas "explicações" para o fenômeno do calor, baseadas em uma abordagem probabilística.

Segundo Halliday et al. (1996), à medida que um sistema torna-se mais desorganizado a nível molecular, as posições das suas moléculas tornam-se menos previsíveis e a sua entropia aumenta . Por isso, a entropia da fase sólida é mais baixa do que a das outras fases pois, nos sólidos, as moléculas oscilam em torno de posições de equilíbrio, não podendo mover umas relativamente às outras e, em qualquer momento, as suas posições são previsíveis com uma certa precisão. Na fase gasosa as moléculas movem-se ao acaso, colidindo umas com as outras, mudando de direção, o que torna extremamente difícil prever, com alguma precisão, o estado microscópico ou configuração molecular de um gás. Associado a este caos molecular está um elevado valor da entropia.

2.2.2.2 Visão estatística da entropia

Segundo Chakrabarti & Chakrabarty (2007), um dos desdobramentos mais ricos e polêmicos do conceito probabilístico de entropia desenvolvido por Boltzmann foi sua extensão ao campo da Teoria da Informação. Quando a informação de ordem j é transmitida, a informação transportada é $I_j = -log_2 P_j \; bits$, conforme a expressão (2.7), mas em geral transmiti-se não somente um símbolo, e sim um conjunto deles (mensagem). Assim, tem-se a informação média associada aos n símbolos transportados.

Para medir a quantidade de informação, Shannon criou o conceito estatístico de entropia, que é diferente do conceito homônimo encontrado em termodinâmica. Porque esta denominação foi escolhida? Segundo Vicki (2007) ao que parece, foi o matemático norte-americano de origem húngara, John Von Neumann, quem

sugeriu este termo. Teria dito, ironicamente, "deve chamá-la de $entropia$ por duas razões: primeiro, porque essa mesma função matemática já é utilizada em termodinâmica, com esse nome; segundo, e mais importante, porque pouca gente sabe realmente o que é entropia e, se usar esse termo numa discussão, sairá sempre ganhando".

De acordo Mackay (2005) a entropia é definida como :

Definição 2.5 *A média ponderada das auto-informações por sua probabilidade de ocorrência é o que chamamos de* ***entropia****, isto é:*

$$H(X) = \sum_{i=1}^{n} p_i I_{p_i} = -\sum_{i=1}^{n} p_i \log p_i \tag{2.10}$$

sendo p_i a probabilidade do evento X_i.

Pode-se também ver a equação (2.10) como

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i = -E\left[\log p_i\right].$$

Este conceito de entropia é útil para medir a quantidade de informação transmitida por uma fonte.

Segundo Wiener (1970) apud Martins (1995), referindo-se a uma sugestão de J. Von Neumann e abstraindo o sinal de negativo, N. Wiener propôs uma extensão do conceito para distribuições contínuas, e definiu:

Definição 2.6 *Seja uma variável aleatória X, contínua, real e centrada (média zero) com uma função de densidade de probabilidade $g(x)$. A* ***entropia*** *é definida por*

$$H_E = \log\left(\frac{g(x)}{f(x)}\right), \tag{2.11}$$

em que H_E é a entropia estatística, $g(x)$ é a "verdadeira" distribuição e $f(x)$ é o nosso modelo estatístico.

Comparando-se as equações (2.9) e (2.11), nota-se que a entropia estatística é a mesma entropia de Boltzmann, a não ser pelo sinal que foi abstraído e pela constante k que é a constante de Boltzmann. Ou seja,

$$H_E = -H_B.$$

Sendo o conceito de entropia conhecido, pode-se perguntar: O que significa a entropia de uma fonte? Significa que, embora não se possa prever qual o símbolo que a fonte irá produzir a seguir, em média espera-se obter $I\ bits$ de informação por símbolo, ou $nI\ bits$ numa mensagem de n símbolos, se n for elevado (Fernandes & Azevedo (2006)).

Assim, dizer que um sinal (uma seqüência) de símbolos tem uma entropia informacional de, por exemplo, $1,75\ bits$ por símbolo significa que pode-se converter a mensagem original em uma seqüência de $0's$ e $1's$ (dígitos binários), de maneira que em média existam $1,75$ dígitos binários por cada símbolo do sinal original. O *em média* aqui quer dizer que alguns símbolos vão precisar de mais dígitos binários para serem codificados (os mais raros) e que outros símbolos vão precisar de menos dígitos binários para serem codificados (os mais comuns).

Exemplo

Suponha que tem-se 4 símbolos (A, C, G, T) com probabilidades de ocorrência iguais a $p_A = \frac{1}{2}$; $p_C = \frac{1}{4}$; $p_G = \frac{1}{8}$; $p_T = \frac{1}{8}$. Estas probabilidades dão as

seguintes quantidades de informação para cada símbolo:

$$I_A = -log_2\left(\frac{1}{2}\right) = 1bit;$$
$$I_C = -log_2\left(\frac{1}{4}\right) = 2bits;$$
$$I_G = -log_2\left(\frac{1}{8}\right) = 3bits;$$
$$I_T = -log_2\left(\frac{1}{8}\right) = 3bits.$$

Portanto, a entropia de uma seqüência desses símbolos é:

$$H = -\sum p_i \log p_i = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = 1,75bit,$$

ou seja, 1,75 símbolos por bits. Pode-se codificar cada um dos quatro símbolos por um número de dígitos binários igual à sua quantidade de informação. Por exemplo:

$$A = 0;$$
$$C = 10;$$
$$G = 110;$$
$$T = 111.$$

Portanto, uma seqüência como:

$$ATCAGAAC,$$

que tem freqüências de ocorrência dos 4 *símbolos iguais às definidas anteriormente pode ser codificada por* 01111001100010*, usando* 14 *dígitos binários para*

codificar 8 símbolos, o que dá uma média de $\frac{14}{8} = 1,75\ bits$ *por símbolo.*

Um código como o exemplificado acima é chamado de código de Shannon-Fano (Cover & Thomas, 1991). Esse código tem a propriedade de que pode ser decodificado sem precisar de espaços entre os símbolos.

Com o conceito de entropia pode-se definir a quantidade de informação transmitida e os limites ótimos de compressão dessa informação (Mackay, 2005). Em 1948, o cabo elétrico de "banda mais larga" então existente podia transmitir 1.800 conversas telefônicas simultâneas. Vinte e cinco anos mais tarde, um cabo telefônico podia transmitir 230.000 conversas simultâneas. Hoje, uma nova fibra ótica com a espessura de um cabelo humano, pode comportar $6,4$ milhões de conversas. No entanto, mesmo com esta largura de banda, os limites teóricos de capacidade de canal determinados por Shannon estão muito aquém dos praticados. Os engenheiros sabem que ainda há muito que melhorar.

Sejam $X_n = \{x_1, x_2, ..., x_n\}$ um conjunto de n observações independentes amostradas aleatoriamente de uma distribuição (modelo) de probabilidades desconhecida $g(x)$ (verdadeiro modelo, do qual retiramos nossos dados), e seja $f(x)$ um modelo arbitrário especificado. O que se quer é avaliar a qualidade do ajuste ao se aproximar o modelo $g(x)$ pelo modelo $f(x)$.

A informação de Kullback-Leibler quantifica essa perda de informações (Konishi & Kitagawa, 2008):

Definição 2.7 *A Informação de Kullback-Leibler é definida por:*

$$I\left(g;f\right) = E_g\left[-H_B\right] = E_g\left[\log\left(\frac{g\left(y\right)}{f\left(y\right)}\right)\right] = \int_{-\infty}^{+\infty} g\left(y\right)\log\left(\frac{g\left(y\right)}{f\left(y\right)}\right)dy \quad (2.12)$$

em que H_B *é a entropia de Boltzmann,* g *é a distribuição da qual são gerados os dados,* f *é a distribuição utilizada para aproximar* g *e* E_g *representa a esperança,*

com respeito a distribuição de probabilidade g.

A equação (2.12) pode também, ser expressa como:

$$I(g;f) = E_g[\log g(x)] - E_g[\log f(x)] \tag{2.13}$$

ou equivalentemente

$$I(g;f) = \int_{-\infty}^{+\infty} g(x) \log[g(x)]\, dx - \int_{-\infty}^{+\infty} g(x) \log[f(x)]\, dx. \tag{2.14}$$

Conforme Konishi & Kitagawa (2008), a Informação de Kullback-Leibler têm as seguintes propriedades:

(P1) Para quaisquer funções de densidade de probabilidade f e g, $I(g;f) \geq 0$;

(P2) Se f e g são funções de densidade de probabilidade e $I(g;f) = 0$, então $f(x) = g(x), \forall x \in \mathbb{R}$;

(P3) Se f e g são duas funções de densidade de probabilidade e $f \to g$, então $I(g;f) \to 0$.

Nota-se que o primeiro termo na equação (2.13) é uma constante, que depende somente do verdadeiro modelo g. Assim, somente o segundo termo de (2.14) é importante na avaliação do modelo estatístico $f(x)$, pois se houver dois modelos candidatos f_1 e f_2, ao compará-los obter-se-á:

$$I(f_1, g) = \int g(x) \ln(g(x))\, dx - \int g(x) \ln(f_1(x))\, dx$$

e

$$I(f_2, g) = \int g(x) \ln(g(x))\, dx - \int g(x) \ln(f_2(x))\, dx.$$

Logo

$$
\begin{aligned}
I(f_1, g) - I(f_2, g) &= \left(\int g(x) \ln(g(x))\, dx - \int g(x) \ln(f_1(x))\, dx \right) \\
&\quad - \left(\int g(x) \ln(g(x))\, dx - \int g(x) \ln(f_2(x))\, dx \right) \\
&= \int g(x) \ln(f_2(x))\, dx - \int g(x) \ln(f_1(x))\, dx. \quad (2.15)
\end{aligned}
$$

Assim vê-se que a primeira parte da equação (2.13) é cancelada, e a equação só depende do segundo termo, chamado de log verossimilhança esperada (Konishi & Kitagawa, 2008). Entretanto a segunda parte ainda depende da função desconhecida g.

$$
E_g[\ln(f(x))] = \int \ln(f(x))\, g(x)\, dx = \int \ln(f(x))\, dG(x). \qquad (2.16)
$$

Em que g é a verdadeira distribuição, f é o odelo que aproxima g e G é a função de distribuição acumulada de g.

Considerar-se-á um exemplo dado por Burnham & Anderson (2002) para ilustrar a K-L informação:

Exemplo

Seja g um distribuição gama com parâmetros $\alpha = 4$ e $\beta = 4$. Consideram-se os modelos g_i, $i = 1, 2, 3, 4$ como sendo aproximações do verdadeiro modelo, em que g_1 é uma Weibull com parâmetros $\alpha = 2$ e $\beta = 20$, g_2 é uma log-normal com parâmetros $\alpha = 2$ e $\sigma^2 = 2$, g_3 é uma inversa Gaussiana com parâmetros $\alpha = 16$ e $\beta = 64$, g_4 é uma distribuição F com parâmetros $\alpha = 4$ e $\beta = 10$.

De acordo com Johnson et al. (1994) tem-se:

$$g(x) = \frac{1}{4^4\Gamma(4)} x^{4-1} e^{-\frac{x}{4}} = \frac{1}{1536} x^3 e^{-\frac{x}{4}}$$

$$g_1(x) = \frac{20}{2^{20}} x^{20-1} e^{-\left(\frac{x}{2}\right)^{20}} = \frac{5}{2^{18}} x^{19} e^{-\frac{x^{20}}{2^{20}}}$$

$$g_2(x) = \frac{1}{x\sqrt{2\pi}\sqrt{2}} e^{-(\ln x - 2)/2\times 2} = \frac{1}{2\sqrt{\pi}x} e^{-(\ln x - 2)/4}$$

$$g_3(x) = \left(\frac{64}{2\pi x^3}\right)^{1/2} e^{\left\{-\frac{64}{2\times 16}\left(\frac{x}{16} - 2 + \frac{16}{x}\right)\right\}} = \frac{4\sqrt{2\pi}}{\pi} x^{-3/2} e^{\left\{-2\left(\frac{x}{16} - 2 + \frac{16}{x}\right)\right\}}$$

$$\begin{aligned} g_4(x) &= \frac{\Gamma[(4+10)/2]}{\Gamma(4/2)\Gamma(10/2)} \left(\frac{4}{10}\right)^{4/2} x^{(4-2)/2} (1 + (4/10)x)^{-(4+10)/2} \\ &= \frac{\Gamma(7)}{\Gamma(2)\Gamma(5)} \left(\frac{2}{5}\right)^2 x (1 + (2/5)x)^{-14/2} \\ &= \frac{24}{5} x \left(1 + \frac{2}{5}x\right)^{-7} \end{aligned}$$

Nas figuras abaixo tem-se o gráfico destas distribuições.

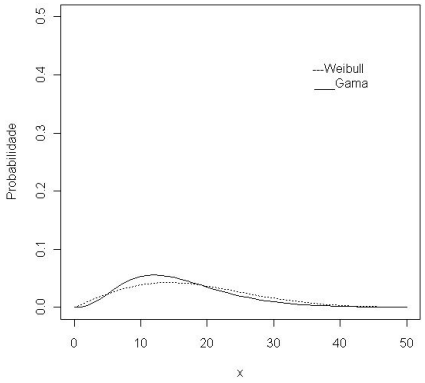

FIGURA 2: Representação gráfica das distribuições Gama(4,4) - linha contínua - e Weibull(2,20) - linha pontilhada

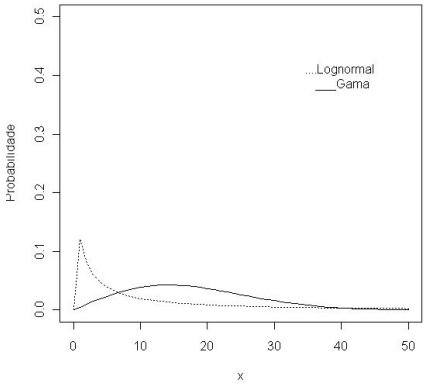

FIGURA 3: Representação das distribuições Gama(4,4) - linha contínua - e Lognormal(2,2) - linha pontilhada

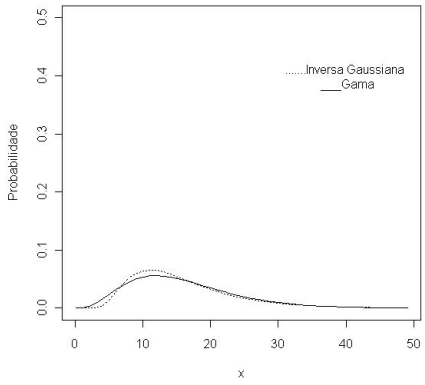

FIGURA 4: Representação gráfica das distribuições Gama(4,4) - linha contínua - e Inversa Gaussiana(16,64) - linha pontilhada

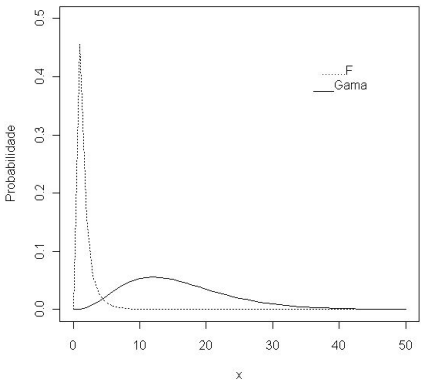

FIGURA 5: Representação gráfica da distribuição Gama(4,4) - linha contínua - e da distribuição F(4,10) - linha pontilhada

Em uma primeira análise, puramente visual, pode-se dizer que as distribuições Weibull e Inversa Gaussiana estão muito mais "próximas" da distribuição Gama que as distribuições Lognormal e F. Vejamos isto através da informação de

Kullback-Leibler, calculando a K-L informação para cada distribuição.

$$
\begin{aligned}
I(g,g_1) &= \int g(x)\ln\left(\frac{g(x)}{g_1(x)}\right)dx = \int g(x)\ln(g(x))\,dx - \int g(x)\ln(g_1(x))\,dx \\
&= \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{x^3e^{-\frac{x}{4}}}{1536}\right)dx - \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{5}{2^{18}}x^{19}e^{-\frac{x^{20}}{2^{20}}}\right)dx \\
&= \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(-\ln(1536)+3\ln(x)-\frac{x}{4}\right)dx}_{(I)} - \\
&- \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(\ln\left(\frac{5}{2^{18}}\right)+19\ln(x)-\frac{x^{20}}{2^{20}}\right)dx}_{(II)} \qquad (2.17)
\end{aligned}
$$

Efetuando as integrações e os cálculos necessários em (2.17) tem-se (I)=$3,40970$ *e* $(II) = 3,3635$ *e assim* $I(g,g_1) = 3,40970 - 3,3635 = 0,04620$.

Para $g_2(x)$ *tem-se*

$$
\begin{aligned}
I(g,g_2) &= \int g(x)\ln\left(\frac{g(x)}{g_2(x)}\right)dx = \int g(x)\ln(g(x))\,dx - \int g(x)\ln(g_2(x))\,dx \\
&= \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{x^3e^{-\frac{x}{4}}}{1536}\right)dx - \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{1}{2\sqrt{\pi}x}e^{-(\ln x-2)/4}\right)dx \\
&= \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(-\ln(1536)+3\ln(x)-\frac{x}{4}\right)dx}_{(III)} - \\
&- \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(-\ln\left(2\sqrt{\pi}\right)-\frac{\ln x}{4}+\frac{1}{2}\right)dx}_{(IV)}. \qquad (2.18)
\end{aligned}
$$

Novamente, efetuando as integrações e os cálculos necessários em (2.18) e notando que $(I) = (III)$ *obtém-se* $(III) = 3,40970$ *e* $(IV) = 2,73735$*, assim* $I(g,g_2) = 3,40970 - 2,73735 = 0,67235$.

Para $g_3(x)$ tem-se

$$\begin{aligned} I(g,g_3) &= \int g(x)\ln\left(\frac{g(x)}{g_3(x)}\right)dx = \int g(x)\ln(g(x))\,dx - \int g(x)\ln(g_3(x))\,dx \\ &= \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{x^3e^{-\frac{x}{4}}}{1536}\right)dx - \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{4\sqrt{2\pi}}{\pi}xe^{\left\{-2\left(\frac{x}{16}-2+\frac{16}{x}\right)\right\}}\right)dx \\ &= \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(-\ln(1536)+3\ln(x)-\frac{x}{4}\right)dx}_{(V)} - \\ &- \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(\ln\left(\frac{4\sqrt{2\pi}}{\pi}\right)+\ln(x)-\frac{x}{8}+4-\frac{32}{x}\right)dx}_{(VI)} \end{aligned} \quad (2.19)$$

Novamente, efetuando as integrações e os cálculos necessários em (2.19) e notando que $(I) = (V)$ obtém-se $(V) = 3,40970$ e $(VI) = 3,34962$ e assim $I(g,g_3) = 3,40970 - 3,34962 = 0,06008$.

Para $g_4(x)$ tem-se

$$\begin{aligned} I(g,g_4) &= \int g(x)\ln\left(\frac{g(x)}{g_4(x)}\right)dx = \int g(x)\ln(g(x))\,dx - \int g(x)\ln(g_4(x))\,dx \\ &= \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{x^3e^{-\frac{x}{4}}}{1536}\right)dx - \int \frac{x^3e^{-\frac{x}{4}}}{1536}\ln\left(\frac{24}{5}x\left(1+\frac{2}{5}x\right)^{-7}\right)dx \\ &= \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(-\ln(1536)+3\ln(x)-\frac{x}{4}\right)dx}_{(VII)} - \\ &- \underbrace{\frac{1}{1536}\int x^3e^{-\frac{x}{4}}\left(\ln\left(\frac{24}{5}\right)+\ln(x)-7\ln\left(1+\frac{2}{5}x\right)\right)dx}_{(VIII)} \end{aligned} \quad (2.20)$$

Novamente, efetuando as integrações e os cálculos necessários em (2.20) e notando que $(I) = (VII)$ obtém-se $(VII) = 3,40970$ e $(VIII) = -2,33585$ e assim $I(g,g_4) = 3,40970 - (-2,33585) = 5,74555$.

Resumidamente, tem-se a seguinte tabela:

Modelo	*K-L informação*	*Posição*
Weibull(2,20)	*0,0462*	*1*
Lognormal(2,2)	*0,67235*	*3*
Inversa Gausssiana(16,64)	*0,06008*	*2*
F(4,10)	*5,74555*	*4*

De acordo com os resultados da K-L Informação, a distribuição que melhor "aproxima" a distribuição gama(4,4) é a distribuição Weibull, seguida pela inversa Gaussiana, a lognormal e a F, respectivamente. Isso condiz com a análise gráfica feita anteriormente e também está de acordo com a propriedade (P3), pois à medida que a distribuição torna-se mais "próxima" da gama, vê-se que $I(g, g_i)$ *diminui.*

Conforme Akaike (1974), a K-L informação é apropriada para testar se um dado modelo é adequado, entretanto o seu uso é limitado, pois ela depende da distribuição g, que é desconhecida. Se uma boa estimativa para a log verossimilhança esperada puder ser obtida através dos dados, esta estimativa poderá ser utilizada como um critério para comparar modelos.

Para analisar a estrutura de um dado fenômeno assumem-se modelos paramétricos $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$ tendo p parâmetros, e em seguida maximiza-se a função de verossimilhança (descrita na seção seguinte) para se estimar o parâmetro $\boldsymbol{\theta}$.

2.2.3 A função de verossimilhança

O método mais importante de achar estimativas é o método de máxima verossimilhança, introduzido por R. A. Fisher. Conforme Bolfarine & Sandoval (2000) a função de verossimilhança é definida como:

Definição 2.8 *Seja* $\{X_1, X_2, ..., X_n\}$ *uma amostra aleatória independente e identicamente distribuída, de tamanho* n *da variável aleatória* X *com função de densidade* $g(x|\theta)$*, com* $\theta \in \Theta$*, em que* Θ *é o espaço paramétrico. A* ***função de verossimilhança*** *de* θ *correspondente à amostra aleatória observada é dada por:*

$$L(\theta; X_1, X_2, ..., X_n) = \prod_{i=1}^{n} g(X_i|\,\theta) = g(X_1|\theta)\, g(X_2|\theta) \ldots g(X_n|\theta)\,. \quad (2.21)$$

Se a amostra tiver sido obtida, os valores de $\{x_1, x_2, ..., x_n\}$ serão conhecidos. Como θ é desconhecido, pode-se propor o seguinte: Para qual valor de θ a função $L(x_1, x_2, \ldots, x_n; \theta)$ será máxima? (Meyer, 1983).

Definição 2.9 *O estimador de máxima verossimilhança de* θ*, isto é,* $\widehat{\theta}$*, é aquele valor de* θ *que maximiza* $L(\theta; X_1, X_2, ..., X_n)$.

Segundo Ferreira (2005), o método de máxima verossimilhança estima os valores dos parâmetros da distribuição em estudo, maximizando a função de verossimilhança. O estimador de máxima verossimilhança, é aquele valor de θ, que maximiza (2.21). Para obter o estimador de máxima verossimilhança, toma-se a derivada primeira de $L(\theta; x_1, x_2, ..., x_n)$ com respeito a θ, iguala-se a zero e resolve-se para θ, obtendo-se os pontos críticos; aquele ponto (se existir) que maximiza $L(\theta; x_1, x_2, ..., x_n)$ é a estimativa de máxima verossimilhança para θ. Havendo mais de um parâmetro, para encontrar os estimadores de máxima verossimilhança dos parâmetros, deve-se primeiro tomar as derivadas parciais da função de verossimilhança com respeito a cada um deles, a seguir igualar a derivada a zero e resolver o sistema obtido. Isto é,

$$\frac{\partial L(\boldsymbol{\theta}; x_1, x_2, ..., x_n)}{\partial \boldsymbol{\theta}} = 0. \quad (2.22)$$

Como a função de verossimilhança $L(\boldsymbol{\theta}; x_1, x_2, ..., x_n)$ e a função log verossimilhança $logL(\boldsymbol{\theta}; x_1, x_2, ..., x_n)$ assumem máximo para o mesmo valor, muitas das vezes é preferível trabalhar com a função log verossimilhança, por esta ser bem menos complicada de trabalhar e encontrar os pontos críticos. A função $S = logL(\boldsymbol{\theta}; x_1, x_2, ..., x_n)$ é chamada função suporte (Cramér, 1973).

Segundo Konishi & Kitagawa (2008), os estimadores de máxima verossimilhança têm muitas propriedades da teoria das grandes amostras que torna o seu resultado mais atrativo. São elas:

- Os estimadores são assintoticamente consistentes, o que significa que quanto maior o tamanho da amostra, mais próximos os valores das estimativas estarão dos verdadeiros valores. Formalmente tem-se:

  Definição 2.10 *Um estimador $\widehat{\boldsymbol{\theta}}$ do parâmetro $\boldsymbol{\theta}$ é um estimador consistente se:* $\lim_{n\to\infty} P\left(\left|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right| \geq \boldsymbol{\epsilon}\right) = 0$*, para qualquer* $\boldsymbol{\epsilon} > 0$.

- Os parâmetros estimados são assintoticamente, normalmente distribuídos. Formalmente tem-se:

  Teorema 2.1 *Seja $\widehat{\boldsymbol{\theta}}$ um estimador de máxima verossimilhança do parâmetro $\boldsymbol{\theta}$, então a distribuição de*

  $$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{p} N\left(0, -\left[E\left[\frac{\partial^2 \ln(X, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right]\right]^{-1}\right).$$

  Vale a pena observar que a variância é justamente a inversa da informação de Fisher.

- Eles também são assintoticamente eficientes, e quanto maior a amostra, maior precisão das estimativas.

- Os estimadores de máxima verossimilhança são também estatísticas suficientes, isto é, são estatísticas que condensam os Ω de tal forma que não são perdidas informações acerca de θ. Tal conceito pode assim ser formalizado:

  Definição 2.11 *Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória de densidade $f(\cdot;\theta)$. Uma estatística $S = s(X_1, X_2, \ldots, X_n)$ é dita ser uma estatística suficiente se e só se a distribuição condicional de $X_1, X_2, \ldots, X_n)$ dado $S = s$ não depender de θ para qualquer valor de $s \in S$.*

- Ele também tem a propriedade da invariância, que pode ser formalizada como:

  Definição 2.12 *Seja $\widehat{\Theta} = \widehat{\vartheta}(X_1, X_2, \ldots, X_n)$ um estimador de máxima verossimilhança de θ com função de densidade $f(\cdot;\theta)$, sendo θ unidimensional. Se $\tau(\cdot)$ é uma função inversível, então o estimador de máxima verossimilhança de $\tau(\theta)$ é $\tau\left(\widehat{\theta}\right)$.*

Estas são excelentes propriedades da teoria das grandes amostras.

Uma outra propriedade, que não necessariamente estes estimadores têm, é o não-enviesamento. Um estimador é não-viesado se sua esperança é igual ao valor estimado. Formalmente tem-se:

Definição 2.13 *Um estimador $\widehat{\boldsymbol{\theta}}$ do parâmetro $\boldsymbol{\theta}$ é um estimador não viesado quando a sua distribuição amostral está centrada no próprio parâmetro, isto é, $E\left[\widehat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$.*

2.2.4 O estimador da função suporte

Depois que o vetor de parâmetros $\boldsymbol{\theta}$ foi estimado, ele é substituído no modelo $f(x|\boldsymbol{\theta})$ e passa-se a trabalhar com o modelo $f(x|\widehat{\boldsymbol{\theta}})$. Assim, ao invés de (2.16)

tem-se

$$E_g\left[\ln f\Big(x|\widehat{\boldsymbol{\theta}}\Big)\right] = \int \ln f\Big(x|\widehat{\boldsymbol{\theta}}\Big) g\left(x\right) dx = \int \ln f\Big(x|\widehat{\boldsymbol{\theta}}\Big) dG\left(x\right). \qquad (2.23)$$

Tendo como base estimadores de máxima verossimilhança, deseja-se encontrar um bom estimador para (2.23). Segundo Konishi & Kitagawa (2008), uma estimativa da função suporte esperada, pode ser obtida substituindo a distribuição de probabilidade desconhecida G na equação (2.23) por uma função de distribuição empírica $\widehat{G}$ baseada nos dados X. Isto pode ser entendido nas definições feitas a seguir.

Definição 2.14 *Sejam $X = \{x_1, x_2, ..., x_n\}$ os dados observados de uma distribuição $G(x)$. A função de distribuição empírica $\widehat{G}$ é a função de densidade acumulada que dá $\dfrac{1}{n}$ de probabilidade para cada X_i. Formalmente,*

$$\widehat{G}_n\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right)$$

em que

$$I\left(X_i \leq x\right) = \begin{cases} 1, \ se\ X_i \leq x \\ 0, \ se\ X_i \geq x. \end{cases}$$

Wasserman(2005), mostra o seguinte teorema:

Teorema 2.2 *Sejam $X_1, X_2, ..., X_n \sim G$ e seja $\widehat{G}_n$ a função densidade acumulada empírica. Então:*

- *Para qualquer valor de x fixo,*

$$E\left(\widehat{G}_n\left(x\right)\right) = G\left(x\right) \qquad (2.24)$$

$$Var\left(\widehat{G}_n(x)\right) = \frac{G(x)(1-G(x))}{n}$$

- $\sup\left[\left|\widehat{G}_n(x) - G(x)\right| \to 0\right]$.

Definição 2.15 *Um **funcional estatístico** $T(G)$ é qualquer função de G, em que G é uma distribuição e T uma função qualquer.*

São exemplos de funcionais:

- A média $\mu = \int x dG(x)$,
- A variância $\sigma^2 = \int (x-\mu)^2 dG(x)$,
- A mediana $m = G^{-1}\left(\frac{1}{2}\right)$.

Um funcional da forma $\int u(x)\, dG(x)$ é dito ser um funcional linear. No caso contínuo, $\int u(x)\, dG(x)$ é definido como sendo $\int u(x)\, g(x)\, dx$ e no caso discreto é definido como sendo $\sum_i u(x_i)\, g(x_i)$.

Definição 2.16 *O estimador para $\theta = T(G)$ é definido por $\widehat{\theta}_n = \widehat{G}_n$.*

Se um funcional pode ser escrito na forma $T(G) = \int u(x) dG(x)$, Konishi & Kitagawa (2008) mostram que o estimador correspondente é dado por

$$T\left(\widehat{G}\right) = \int u(x) d\widehat{G}(x) = \sum_{i=1}^{n} \widehat{g}(x_i)\, u(x_i) = \frac{1}{n}\sum_{i=1}^{n} u(x_i) \qquad (2.25)$$

ou seja, substitui-se a função densidade de probabilidade acumulada G pela função de distribuição acumulada empírica $\widehat{G}$, e a função densidade $\widehat{g}_n = \frac{1}{n}$ para cada observação X_i.

Assim, se por exemplo, a função de densidade acumulada G for substituída por $\widehat{G}$, será obtido o seguinte estimador para a média μ:

$$T\left(\widehat{G}\right) = \int x d\widehat{G}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$

que é exatamente a média amostral.

De (2.25) vê-se que pode-se estimar a função suporte esperada por:

$$\begin{aligned} E_{\widehat{G}}\left[\log f\left(x|\widehat{\boldsymbol{\theta}}\right)\right] &= \int \log f\left(x|\widehat{\boldsymbol{\theta}}\right) d\widehat{G}\left(x\right) \\ &= \sum_{i=1}^{n} \widehat{g}\left(x_i|\widehat{\boldsymbol{\theta}}\right) \log f\left(x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log f\left(x_i|\widehat{\boldsymbol{\theta}}\right). \end{aligned} \tag{2.26}$$

Nota-se que o estimador da função suporte esperada $E_G\Big[logf\Big(x|\widehat{\boldsymbol{\theta}}\Big)\Big]$ é $n^{-1}L\left(\widehat{\boldsymbol{\theta}}\right)$ e a função suporte $L\left(\widehat{\boldsymbol{\theta}}\right)$ é um estimador de $nE_G\Big[logf\Big(x|\widehat{\boldsymbol{\theta}}\Big)\Big]$.

3 OS CRITÉRIOS DE INFORMAÇÃO AIC E BIC

Com o intuito de comparar n modelos, $g_1(x|\boldsymbol{\theta_1})$, $g_2(x|\boldsymbol{\theta_2})$, ..., $g_n(x|\boldsymbol{\theta_n})$, pode-se simplesmente comparar as magnitudes da função suporte maximizada, isto é, $L\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)$, mas tal método não dá uma verdadeira comparação, haja vista que, em não conhecendo o verdadeiro modelo $g(x)$, primeiramente utiliza-se o método da máxima verossimilhança para estimar-se os parâmetros $\boldsymbol{\theta_i}$ de cada modelo $g_i(x), i = 1, 2, ...n$, posteriormente utilizar-se-á os mesmos dados para estimar-se $E_G\left[logf(x|\widehat{\boldsymbol{\theta}})\right]$, isto introduz um viés em $L\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)$, sendo que, a magnitude deste viés varia de acordo com a dimensão do vetor de parâmetros.

De acordo com a Definição (2.13) o viés é dado por

$$b(G) = E_{G(\boldsymbol{x_n})}\left[\log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}(X_n)\right) - nE_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}(X_n)\right)\right]\right], \quad (3.1)$$

em que a esperança é tomada com respeito à distribuição conjunta.

Vê-se assim que os critérios de informação são construídos para avaliar e corrigir o viés da função suporte. Segundo Konishi & Kitagawa (2008), um critério de informação tem a forma que se segue:

$$\begin{aligned} CI\left(\boldsymbol{X_n}, \widehat{G}\right) &= -2\left(\log(\text{verossimilhança}) - \text{viés}\right) \\ &= -2\sum_{i=1}^{n}\log f\left(X_n|\widehat{\boldsymbol{\theta}}(X_n)\right) + 2\left(b(G)\right). \end{aligned} \quad (3.2)$$

Alguns critérios comuns na literatura também podem ser utilizados para seleção de modelos. Esses critérios levam em consideração a complexidade do modelo no critério de seleção. São critérios que essencialmente, penalizam a verossimilhança, utilizando o número de variáveis do modelo e, eventualmente o tamanho da amostra. Esta penalização é feita subtraindo-se do valor da verossimilhança

uma determinada quantidade, que depende do quão complexo é o modelo (quanto mais parâmetros, mais complexo).

Akaike (1974), propôs utilizar a informação de Kullback-Leibler para a seleção de modelos. Ele estabeleceu uma relação entre a máxima verossimilhança e a informação de Kullback-Leibler desenvolvendo então um critério para estimar a informação de Kullback-Leibler, o posteriormente chamado, Critério de Informação de Akaike(AIC).

Critérios de seleção de modelos como o Critério de Informação de Akaike (AIC) e Critério de Informação Bayesiano (BIC), são freqüentemente utilizados para selecionar modelos em diversas áreas. Segundo esses critérios, o melhor modelo será aquele que apresentar menor valor de AIC ou BIC.

Por serem resultados assintóticos, os resultados deste trabalho são válidos para "grandes" amostras, sendo o conceito de "grande" amostra difícil de se definir, pois tal conceito depende da área de estudo, da disponibilidade de recursos para uma amostra maior, dentre outros fatores. Se houver convicção de que a amostra em mãos não é "'grande", pode-se utilizar as correções destes critérios, já existentes, para pequenas amostras. Tais correções não serão alvo desse estudo, mas podem ser encontradas em (Burnham & Anderson, 2002).

3.1 Critério de informação de Akaike

O Critério de informação de Akaike (AIC) desenvolvido por Hirotugu Akaike sob o nome de "um critério de informação" em 1971 e proposto, em Akaike (1974), é uma medida relativa da qualidade de ajuste de um modelo estatístico estimado. Fundamenta-se no conceito de entropia, oferecendo uma medida relativa das informações perdidas, quando um determinado modelo é usado para descrever a realidade. Akaike encontrou uma relação entre a esperança relativa da K-L informação

e a função suporte maximizada, permitindo uma maior interação entre a prática e a teoria, em seleção de modelos e análises de conjuntos de dados complexos (Burnham & Anderson, 2002).

Akaike (1974), mostrou que o viés é dado assintoticamente por:

$$b\left(G\right) = tr\left\{I\left(\boldsymbol{\theta}_0\right) J\left(\boldsymbol{\theta}_0\right)^{-1}\right\}, \tag{3.3}$$

sendo $J\left(\boldsymbol{\theta}_0\right)$ e $I\left(\boldsymbol{\theta}_0\right)$ dados por (6.6) e (6.10), respectivamente. A derivação desse resultado é carregada de cálculos matemáticos e por isso encontra-se nos anexos.

O AIC é um critério que avalia a qualidade do ajuste do modelo paramétrico, estimado pelo método da máxima verossimilhança. Ele baseia-se no fato de que o viés (3.3) tende ao número de parâmetros a serem estimados no modelo, pois sob a suposição de que existe um $\boldsymbol{\theta}_0 \in \Theta$ tal que $g(x) = f(x|\boldsymbol{\theta}_0)$, tem-se a igualdade das expressões (6.6) e (6.10), isto é, $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ e assim obter-se-à em (3.3) que:

$$\begin{aligned} b\left(G\right) &= E_{G(\boldsymbol{x_n})}\left[\log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right) - nE_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right)\right]\right] \\ &= tr\left\{I\left(\boldsymbol{\theta_0}\right) J\left(\boldsymbol{\theta_0}^{-1}\right)\right\} = tr\left(I_p\right) = p, \end{aligned} \tag{3.4}$$

em que p é o número de parâmetros a serem estimados no modelo.

Com esse resultado, Akaike (1974) definiu seu critério de informação como:

$$AIC = -2\left(\text{Função suporte maximizada}\right) + 2\left(\text{número de parâmetros}\right),$$

$$AIC = -2\log L\left(\widehat{\boldsymbol{\theta}}\right) + 2\left(k\right) \tag{3.5}$$

O AIC não é uma prova sobre o modelo, no sentido de testar hipóteses, mas

uma ferramenta para a seleção de modelos; não é um teste de hipóteses, não há significância e nem valor-p. Dado um conjunto de dados e vários modelos concorrentes, pode-se classificá-los de acordo com o seu AIC, com aqueles tendo os menores valores de AIC sendo os melhores (Burnham & Anderson, 2002). A partir do valor do AIC pode-se inferir que, por exemplo, os três principais modelos estão em um empate e os restantes são muito piores, mas não se deve atribuir um valor cima do qual um determinado modelo é "rejeitado".

Esse critério está implementado em grande parte dos softwares estatísticos, tais como SAS, R, Statistica, etc. Por si só, o valor do AIC para um determinado conjunto de dados não tem qualquer significado. O AIC torna-se útil quando são comparados diversos modelos. O modelo com o menor AIC é o "melhor" modelo, dentre os modelos comparados. Se apenas modelos ruins forem considerados, o AIC selecionará o melhor dentre estes modelos.

3.2 Critério de informação bayesiano

O Critério de informação Bayesiano (BIC), também chamado de Critério de Schwarz, foi proposto por Schwarz (1978), e é um critério de avaliação de modelos definido em termos da probabilidade a posteriori, sendo assim chamado porque Schwarz deu um argumento Bayesiano para prová-lo. A seguir serão descritos alguns conceitos que levarão à construção deste critério ao final desta subseção.

- **O teorema de Bayes**

De acordo com Bolfarine & Sandoval (2000), quando dois ou mais eventos de um espaço amostral são levados em consideração conjuntamente, passa a haver sentido conjecturar se a ocorrência ou não de um afeta a ocorrência ou não do outro, isto é, se são independentes ou não. Intuitivamente, somos levados à definição de que dois eventos são independentes se, $P[A \cap B] = P[A]\,P[B]$. Entretanto,

se há dependência entre os eventos, passa a haver sentido falar na probabilidade de que um evento ocorra dado que outro ocorreu ou não. Esta dependência motiva a definição de probabilidade condicional. Finalmente, os conceitos de independência e probabilidade condicional levarão ao teorema de Bayes.

Mood et al. (1974), definem probabilidade condicional, independência e subconjuntos mutuamente exclusivos como se segue:

Definição 3.1 *A **probabilidade condicional** de um evento A dado um evento B, denotada por $P[A|B]$ é definida por:*

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

se $P[B] > 0$ e é indefinida se $P[B] = 0$.

Definição 3.2 *Dois eventos A e B são ditos **independentes** se, e só se, qualquer uma das três condições é verdadeira*

- $P[A \cap B] = P[A]\,P[B]$,
- $P[A|B] = P[A]$, *se* $P[B] > 0$,
- $P[B|A] = P[B]$, *se* $P[B] > 0$.

Definição 3.3 *Dois conjuntos A e B, subconjuntos de Ω, são definidos como sendo **mutuamente exclusivos** (disjuntos) se $A \cap B = \emptyset$. Subconjuntos $A_1, A_2, \ldots$ são ditos mutuamente exclusivos se $A_i \cup A_j = \emptyset$ para todo $i \neq j$, $i, j \in \mathbb{N}$.*

Teorema 3.1 *Se $(\Omega, \mathcal{A}, P[\cdot])$ é um espaço de probabilidades e $B_1, B_2, \ldots, B_n$ é uma coleção de eventos mutuamente exclusivos em $\mathcal{A}$, satisfazendo $\Omega = \bigcup_{j=1}^{n} B_j$*

e $P[B_j] > 0$*, para* $j = 1, 2, \ldots, n$*, então para todo* $A \in \mathcal{A}$*, tal que* $P[A] > 0$*, tem-se:*

$$P[B_k|A] = \frac{P[A|B_k]\,P[B_k]}{\sum_{j=1}^{n} P[A|B_k]\,P[B_k]}, \tag{3.6}$$

sendo Ω *o espaço amostral e* $\mathcal{A}$ *o espaço paramétrico.*

Conforme Konishi & Kitagawa (2008), sejam $M_1, M_2, \ldots, M_k$, k modelos candidatos, cada um dos modelos M_i com uma distribuição de probabilidades $f_i(x|\theta_i)$ e uma priori, $\pi_i(\boldsymbol{\theta_i})$ para o k_i-ésimo vetor θ_i. Se são dadas n observações $\boldsymbol{x_n} = \{x_1, x_2, \ldots, x_n\}$, então para o $i-$ésimo modelo M_i, a distribuição marginal de $\boldsymbol{x_n}$ é dada por:

$$p_i(x_n) = \int f_i(\boldsymbol{x_n}|\boldsymbol{\theta_i})\,\pi_i(\boldsymbol{\theta_i})\,d\boldsymbol{\theta_i}. \tag{3.7}$$

Essa quantidade pode considerada como a verossimilhança para o i-ésimo modelo e será referida como verossimilhança marginal dos dados.

Sendo $P(M_i)$ a distribuição a priori do i-ésimo modelo, por (3.6) a distribuição a posteriori será (Burnham & Anderson, 2002):

$$P(M_i|\boldsymbol{x_n}) = \frac{p_i(\boldsymbol{x_n})\,P(M_j)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})\,P(M_j)} \tag{3.8}$$

Segundo Paulino et al. (2003), a probabilidade a posteriori indica a probabilidade dos dados serem gerados do i-ésimo modelo quando os dados $\boldsymbol{x_n}$ são observados. Se um modelo está sendo selecionado de r modelos, seria natural adotar o modelo que tem a maior probabilidade a posteriori. Esse princípio mostra que o modelo que maximiza o numerador $p_j(\boldsymbol{x_n})\,P(M_j)$ deve ser selecionado, pois todos os modelos compartilham do mesmo denominador em (3.8). Se as dis-

tribuições a priori $P(M_i)$ são iguais em todos os modelos, então o modelo que maximiza a probabilidade marginal dos dados $p_i(\boldsymbol{x_n})$, deve ser selecionado. Assim, se uma aproximação para o probabilidade marginal expressa em termos da integral em (3.8) puder ser obtida, a necessidade básica de encontrar a integral problema-por-problema desaparece, isto faz do BIC um critério satisfatório para seleção de modelos.

De acordo com Konishi & Kitagawa (2008), o BIC é definido como:

$$\begin{aligned} -2logp_i\left(\boldsymbol{x_n}\right) &= -2log\int f_i\left(\boldsymbol{x_n}|\boldsymbol{\theta_i}\right)\pi_i\left(\boldsymbol{\theta_i}\right)d\boldsymbol{\theta_i} \\ &\approx -2logf_i\left(\boldsymbol{x_n}|\widehat{\boldsymbol{\theta_i}}\right)+k_ilogn \end{aligned} \tag{3.9}$$

em que $\widehat{\boldsymbol{\theta_i}}$ é o estimador de máxima verossimilhança para o k_i-ésimo vetor paramétrico θ_i do modelo $f_i\left(\boldsymbol{x_n}|\boldsymbol{\theta_i}\right)$.

Conseqüentemente, dos r modelos avaliados usando o método de máxima verossimilhança, o modelo que minimizar o valor do BIC é o melhor modelo para os dados.

Assim, sob a suposição de que todos os modelos têm distribuição de probabilidades a priori iguais, a probabilidade posteriori, obtida usando a informação do dados, serve para contrastar os modelos e ajuda na identificação do modelo que gerou os dados.

Sejam M_1 e M_2 dois modelos que quer-se comparar. Para cada modelo tem-se as verossimilhanças marginais $p_i(\boldsymbol{x_n})$, as prioris $P(M_i)$ e as posterioris $P(M_i|\boldsymbol{x_n})$ com $i=\{1,2\}$, assim, a razão à posteriori em favor do modelo M_1 versus o mo-

delo M_2 é:

$$\frac{P\left(M_1|\boldsymbol{x_n}\right)}{P\left(M_1|\boldsymbol{x_n}\right)} = \frac{\frac{p_1(\boldsymbol{x_n})P(M_1)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})P(M_j)}}{\frac{p_2(\boldsymbol{x_n})P(M_2)}{\sum_{j=1}^{n} p_j(\boldsymbol{x_n})P(M_j)}} = \frac{p_1\left(\boldsymbol{x_n}\right)P\left(M_1\right)}{p_2\left(\boldsymbol{x_n}\right)P\left(M_2\right)}.$$

A razão

$$\frac{p_1\left(\boldsymbol{x_n}\right)}{p_2\left(\boldsymbol{x_n}\right)} \tag{3.10}$$

é chamada de *Fator de Bayes*.

Segundo Konishi & Kitagawa (2008), Akaike mostrou que a comparação baseada no fator de Akaike é assintoticamente equivalente à comparação através do fator de Bayes.

O problema em encontrar o valor do BIC reside no fato de ter-se que calcular o valor da integral em (3.7). Isso é feito utilizando-se a aproximação de Laplace para integrais.

- **A aproximação de Laplace para integrais**

Considere a aproximação de Laplace para a integral

$$\int \exp\left\{nq\left(\boldsymbol{\theta}\right)\right\}d\boldsymbol{\theta}, \tag{3.11}$$

em que $\boldsymbol{\theta}$ é um vetor de parâmetros p-dimensional e $q(\boldsymbol{\theta})$ é uma função real p-dimensional.

A grande vantagem da aproximação de Laplace é o fato de que quando o número n de observações é grande, o integrando concentra-se em um vizinhança $\widehat{\boldsymbol{\theta}}$ de $q(\boldsymbol{\theta})$, e conseqüentemente, o valor da integral depende somente do comportamento do integrando na vizinhança de $\widehat{\boldsymbol{\theta}}$.

Assim, $\left.\frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0$ e a expansão de $q(\boldsymbol{\theta})$ em torno de $\widehat{\boldsymbol{\theta}}$ é:

$$q(\boldsymbol{\theta}) = q\left(\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J_q\left(\widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + \ldots, \tag{3.12}$$

em que

$$J_q\left(\widehat{\boldsymbol{\theta}}\right) = -\left.\frac{\partial^2 q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{3.13}$$

Definição 3.4 *Sejam $q(\boldsymbol{\theta})$ uma função de valores reais avaliada em torno de $\widehat{\boldsymbol{\theta}}$, sendo $\boldsymbol{\theta}$ um vetor de parâmetros. Então a* ***aproximação de Laplace*** *para a integral é dada por:*

$$\int \exp\{nq(\boldsymbol{\theta})\, d\boldsymbol{\theta}\} \approx \frac{(2\pi)^{p/2}}{(n)^{p/2}\left|J_q\left(\widehat{\boldsymbol{\theta}}\right)\right|^{p/2}} \exp\left(nq\left(\widehat{\boldsymbol{\theta}}\right)\right) \tag{3.14}$$

em que $J_q\left(\widehat{\boldsymbol{\theta}}\right)$ é definido em (3.13).

Utilizando-se a aproximação de Laplace para aproximar (3.7), que pode ser reescrita como

$$\begin{aligned} p(x_n) &= \int f_i(\boldsymbol{x_n}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \exp\{\log f(\boldsymbol{x_n}|\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \exp\{\ell(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \end{aligned} \tag{3.15}$$

em que $\ell(\boldsymbol{\theta})$ é a função suporte $\ell(\boldsymbol{\theta}) = logf(x_n|\boldsymbol{\theta})$.

Assim sendo, fazendo-se a expansão em séries de Taylor de $\ell(\boldsymbol{\theta})$ e $\pi(\boldsymbol{\theta})$ em torno de $\widehat{\boldsymbol{\theta}}$ obter-se-á respectivamente:

$$\ell(\boldsymbol{\theta}) = \ell\left(\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J\left(\widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + \ldots, \tag{3.16}$$

$$\pi(\boldsymbol{\theta}) = \pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots, \tag{3.17}$$

substituindo (3.16) e (3.17) em (3.15) obtém-se:

$$\begin{aligned} p(x_n) = & \int \exp\left\{\pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots\right\} d\boldsymbol{\theta} \\ & \times \left\{\pi\left(\widehat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \left.\frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \ldots\right\} d\boldsymbol{\theta} \\ & \approx \exp\left\{\ell\left(\widehat{\boldsymbol{\theta}}\right)\right\} \pi\left(\widehat{\boldsymbol{\theta}}\right) \int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta}) \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\} d\boldsymbol{\theta} \end{aligned} \tag{3.18}$$

A integral em (3.18) satisfaz a equação (3.14), conseqüentemente pode ser aproximada utilizando Laplace, e obtém-se:

$$\int \exp\left\{-\frac{n}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta}) \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right\} d\boldsymbol{\theta} = (2\pi)^{p/2} n^{-p/2} \left|J\left(\widehat{\boldsymbol{\theta}}\right)\right|^{-1/2}, \tag{3.19}$$

em que o integrando é uma função de densidade normal p-dimensional com vetor de médias $\widehat{\boldsymbol{\theta}}$ e matriz de covariância $J^{-1}\left(\widehat{\theta}\right)\Big/n$.

Para n grande,

$$p(x_n) \approx \exp\left\{\ell\left(\widehat{\boldsymbol{\theta}}\right)\right\} \pi\left(\widehat{\boldsymbol{\theta}}\right) (2\pi)^{p/2} n^{-p/2} \left|J\left(\widehat{\boldsymbol{\theta}}\right)\right|^{-1/2} \tag{3.20}$$

Tomando o logarítmo em (3.20) e multiplicando a expressão por -2 obtém-se

$$\begin{aligned} -2\log p(x_n) &= -2\log\left\{\int f(x_n|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right\} \\ &= -2\ell\left(\widehat{\boldsymbol{\theta}}\right) + p\log n + \log\left|J\left(\widehat{\boldsymbol{\theta}}\right)\right| - p\log(2\pi) - 2\log\pi\left(\widehat{\boldsymbol{\theta}}\right) \end{aligned} \tag{3.21}$$

Assim, o Critério de Informação Bayesiano pode ser obtido da seguinte forma (ignorando-se os termos constantes no equação):

Definição 3.5 *Seja* $F\left(\boldsymbol{x_n}|\widehat{\boldsymbol{\theta}}\right)$ *um modelo estatístico estimado através do método de máxima verossimilhança. Então o Critério de Informação Bayesiano(BIC) é dado por:*

$$BIC = -2\log f\left(x_n|\boldsymbol{\theta}\right) + p\log n, \tag{3.22}$$

em que $f\left(x_n|\boldsymbol{\theta}\right)$ *é o modelo escolhido,* p *é o número de parâmetros a serem estimados e* n *é o número de observações da amostra.*

3.3 Algumas considerações acerca do AIC e do BIC

Vale a pena salientar algumas características dos critérios AIC e BIC. A maioria dessas considerações são feitas por Burnahm & Anderson(2002), e também estão no website desses autores, onde estão disponíveis outras considerações acerca destes métodos.

- Tanto o AIC quanto o BIC fundamentam-se na verossimilhança, impondo entretanto diferentes penalizações;
- O AIC e o BIC servem para comparar modelos encaixados, mas podem ser aplicados também em modelos não encaixados;
- Para $n > 8$, o valor do AIC para um determinado modelo será sempre menor que o valor do BIC, mas os resultados não necessariamente o serão;
- O AIC e o BIC servem para comparar quaisquer quantidade de modelos, e não somente dois, como muitos pensam;
- O AIC e o BIC são critérios assintóticos e já existem correções para estes;
- O AIC e o BIC servem para estudar estruturas de covariâncias;

- A seleção dos modelos é feita pelo pesquisador e, se somente modelos ruins forem selecionados, o AIC fará a seleção do melhor dentre eles.

4 APLICAÇÕES DO AIC E BIC

4.1 Os dados

Para a realização desse trabalho foram avaliados dois conjuntos de dados distintos.

O primeiro conjunto de dados é disponibilizado em Triola (1999), e encontra-se no anexo A. Foram extraídas duas amostras de confeitos M&M, pesados os de cores vermelha e amarela. A variável resposta foi o peso em gramas de cada elemento amostral. Utilizando o AIC e o BIC desejou-se testar se os pesos dos confeitos amarelos e vermelhos seguem a mesma distribuição.

O segundo conjunto de dados foi obtido de Rawlings et al. (1998). Trata-se de um estudo das características que influenciam a produção aérea de biomassa na grama de pântano. Foram amostrados três tipos de vegetação Spartina, em três localidades (Oak Island, Smith Island, and Snows Marsh). Em cada localidade, cinco amostras aleatórias do substrato de terra de cada tipo de vegetação foram coletadas, totalizando 45 amostras.

Foram analisadas 14 características físico-químicas da terra durante vários meses, porém os dados usados nesse estudo envolvem só a amostragem de setembro, em que foram analisadas as variáveis: salinidade (Sal), pH (pH), potássio (K) em ppm, sódio (Na) em ppm , zinco (Zn) em ppm e a variável resposta foi a biomassa aérea em gm^{-2}. O propósito do estudo foi utilizar regressão linear múltipla para relacionar a produção de biomassa com as cinco variáveis estudadas.

4.2 Igualdade de médias e / ou de variâncias de distribuições normais

Uma utilidade dos critérios de Akaike e de Schwarz é testar se os dados oriundos de uma distribuição normal tem mesma média e variância; ou mesma média

e variâncias diferentes, ou diferentes médias e mesma variância ou se provém de uma normal com médias e variâncias diferentes.

Sejam dois conjuntos de dados $\{y_1, y_{2,}\ldots, y_n\}$ e $\{y_{n+1}, y_{n+2,}\ldots, y_{n+m}\}$, sendo que $y_1, y_{2,}\ldots, y_n \sim N\left(\mu_1, \sigma_1^2\right)$ e $y_{n+1}, y_{n+2,}\ldots, y_{n+m} \sim N\left(\mu_2, \sigma_2^2\right)$. Deseja-se verificar se:

$$\mu_1 = \mu_2 = \mu \quad \text{e} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \ \text{ ou} \tag{4.1}$$

$$\mu_1 \neq \mu_2 \quad \text{e} \quad \sigma_1^2 \neq \sigma_2^2 \ \text{ ou} \tag{4.2}$$

$$\mu_1 \neq \mu_2 \quad \text{e} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \ \text{ ou} \tag{4.3}$$

$$\mu_1 = \mu_2 = \mu \quad \text{e} \quad \sigma_1^2 \neq \sigma_2^2 \tag{4.4}$$

Tem-se que

$$f\left(y_1|\mu_1, \sigma_1^2\right) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_1}{\sigma_1}\right)^2\right\}, i = 1, 2, \ldots, n,$$

e

$$f\left(y_2|\mu_2, \sigma_2^2\right) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_2}{\sigma_2}\right)^2\right\}, i = n+1, n+2, \ldots, n+m,$$

E a função de densidade conjunta é dada por:

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{\theta}) = & \quad f\left(y_1, \ldots, y_n, y_{n+1}, \ldots, y_{n+m}|\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) \\ = & \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\left(\frac{y_i - \mu_1}{\sqrt{2}\sigma_1}\right)^2\right]\right\} \prod_{i=n+1}^{n+m}\left\{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\left(\frac{y_i - \mu_1}{\sqrt{2}\sigma_2}\right)^2\right]\right\} \end{aligned}$$

Assim, a função suporte é:

$$L(\boldsymbol{\theta}) = \log\left\{\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2}\left(\frac{y_i-\mu_1}{\sigma_1}\right)^2}\right]\prod_{i=n+1}^{n+m}\left[\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{1}{2}\left(\frac{y_i-\mu_2}{\sigma_2}\right)^2}\right]\right\}$$

$$= -\frac{n}{2}\log\left(2\pi\sigma_1^2\right) - \frac{\sum_{i=1}^{n}\left(y_i-\mu_1\right)^2}{2\sigma_1^2} - \frac{m}{2}\log\left(2\pi\sigma_2^2\right) - \frac{\sum_{i=n+1}^{n+m}\left(y_i-\mu_2\right)^2}{2\sigma_2^2} \quad (4.5)$$

em que $\boldsymbol{\theta} = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$.

Serão obtidas as situações descritas em (4.1), (4.2) (4.3), e (4.4). Será feita agora a derivação dos critérios de Akaike e Schwarz para cada uma delas.

Caso 1: $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Para o caso descrito em (4.1), ou seja, $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$ existem dois parâmetros μ e σ^2 desconhecidos. Esta suposição é equivalente a termos $n+m$ observações $y_1, y_2, \ldots, y_{n+m}$ de uma distribuição normal , isto é,

$$y_1, y_2, \ldots, y_{n+m}, \sim N\left(\mu, \sigma^2\right).$$

Sob a suposição (4.1) tem-se de (4.5) que

$$L\left(\boldsymbol{\theta}\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\sigma^2} - \frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{\sum_{i=n+1}^{n+m}\left(y_i-\mu\right)^2}{2\sigma^2}$$

$$L\left(\boldsymbol{\theta}\right) = -\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}\left(y_i-\mu\right)^2, \quad (4.6)$$

sendo $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$.

Maximizando (4.6) tem-se:

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{n+m}{2}\left[\log\left(2\pi\widehat{\sigma}^2\right) + 1\right], \quad (4.7)$$

em que

$$\widehat{\mu} = \frac{1}{n+m} \sum_{i=1}^{n+m} y_i \tag{4.8}$$

e

$$\widehat{\sigma}_2^2 = \frac{1}{n+m} \sum_{i=1}^{n+m} (y_i - \widehat{\mu})^2. \tag{4.9}$$

Os cálculos inerentes a esses resultados encontram-se no Anexo C.

O valor do AIC é dado por:

$$AIC = -2\,(\text{Função suporte maximizada}) + 2\,(\text{número de parâmetros})\,,$$

$$AIC = -2\left(\log L\left(\widehat{\boldsymbol{\theta}}\right)\right) + 2\,(k) \tag{4.10}$$

em que $L(\widehat{\theta})$ é a verossimilhança maximizada e k o número de parâmetros desconhecidos e estimados.

Substituindo (4.7) em (4.10), tem-se:

$$AIC_1 = -2\left\{\frac{n+m}{2}\left[\log\left(2\pi\widehat{\sigma}^2\right) + 1\right]\right\} + 2(2) = (n+m)\left[\log\left(2\pi\widehat{\sigma}^2\right) + 1\right] + 4$$

$$AIC_1 = (n+m)\left(\log\widehat{\sigma}^2 + \log 2\pi + 1\right) + 4 \tag{4.11}$$

O valor do BIC é dado por:

$$BIC = -2\,(\text{Função suporte maximizada}) + (\text{número de parâmetros})\log n,$$

$$BIC = -2\left(\log L\left(\widehat{\boldsymbol{\theta}}\right)\right) + (k)\log n \tag{4.12}$$

em que $L(\widehat{\theta})$ é a função de verossimilhança maximizada e k o número de parâmetros desconhecidos e estimados.

Substituindo (4.7) em (4.12), tem-se:

$$\begin{aligned} BIC_1 &= -2\left\{\frac{n+m}{2}\left[\log\left(2\pi\widehat{\sigma}^2\right)+1\right]\right\}+2\log(n) \\ &= (n+m)\left[\log\left(2\pi\widehat{\sigma}^2\right)+1\right]+2\log(n+m) \\ BIC_1 &= (n+m)\left(\log\widehat{\sigma}^2+\log 2\pi+1\right)+2\log(n+m) \end{aligned} \tag{4.13}$$

Caso 2: $\mu_1 \neq \mu_2$ e $\sigma_1^2 \neq \sigma_2^2$

Se todos os parâmetros são desconhecidos tem-se então $\boldsymbol{\theta} = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)$ e assim a função em (4.5) é expressa como:

$$\begin{aligned} L(\boldsymbol{\theta}) &= L\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) = -\frac{n}{2}\log\left(2\pi\sigma_1^2\right) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n}(y_i-\mu_1)^2 \\ &- \frac{m}{2}\log\left(2\pi\sigma_2^2\right) - \frac{1}{2\sigma_2^2}\sum_{i=n+1}^{n+m}(y_i-\mu_2)^2 \end{aligned} \tag{4.14}$$

Logo,

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma_1^2}\right) - \frac{\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2}{2\widehat{\sigma_1^2}} - \frac{m}{2}\log\left(2\pi\sigma_2^2\right) - \frac{\sum_{i=n+1}^{m}(y_i-\widehat{\mu_2})^2}{2\widehat{\sigma_2^2}}, \tag{4.15}$$

e $\widehat{\mu_1}$, $\widehat{\mu_2}$, $\widehat{\sigma_1^2}$ e $\widehat{\sigma_2^2}$ são dados por respectivamente por (4.16), (4.17), (4.18) e (4.19).

$$\widehat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.16}$$

$$\widehat{\mu_2} = \frac{1}{m}\sum_{i=n+1}^{n+m} y_i \tag{4.17}$$

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2 \tag{4.18}$$

$$\widehat{\sigma_2^2} = \frac{1}{m}\sum_{i=1}^{n}(y_i-\widehat{\mu_2})^2. \tag{4.19}$$

Todos os cálculos necessários para a obtenção desses resultados encontram-se no Anexo C.

Substituindo (4.15) em (4.10), já multiplicando pelo fator -2, tem-se:

$$AIC_2 = n\log\Big(2\pi\widehat{\sigma_1^2}\Big) + \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu_1})^2}{\widehat{\sigma_1^2}} + m\log\Big(2\pi\widehat{\sigma_2^2}\Big) + \frac{\sum_{i=n+1}^{m}(y_i - \widehat{\mu_2})^2}{\widehat{\sigma_2^2}} + 2(4)$$

$$AIC_2 = n\log\Big(2\pi\widehat{\sigma_1^2}\Big) + \frac{n\widehat{\sigma_1^2}}{\widehat{\sigma_1^2}} + m\log\Big(2\pi\widehat{\sigma_2^2}\Big) + \frac{m\widehat{\sigma_2^2}}{\widehat{\sigma_2^2}} + 8$$

$$AIC_2 = (n+m)\log(2\pi) + n\log\widehat{\sigma_1^2} + m\log\widehat{\sigma_2^2} + (n+m) + 8$$

$$AIC_2 = (n+m)(\log(2\pi) + 1) + n\log\sigma_1^2 + m\log\sigma_2^2 + 8 \tag{4.20}$$

O valor do BIC é dado por:

$$BIC = -2(\log L(\boldsymbol{\theta})) + (k)\log n, \tag{4.21}$$

Substituindo (4.15) em (4.21), tem-se:

$$\begin{aligned} BIC_2 &= -2\left(-\frac{n}{2}\log\Big(2\pi\widehat{\sigma_1^2}\Big) - \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu_1})^2}{2\widehat{\sigma_1^2}}\right. \\ &- \left.\frac{m}{2}\log\big(2\pi\sigma_2^2\big) - \frac{\sum_{i=n+1}^{m}(y_i - \widehat{\mu_2})^2}{2\widehat{\sigma_2^2}}\right) + 4\log n \end{aligned}$$

$$BIC_2 = n\log\Big(2\pi\widehat{\sigma_1^2}\Big) + \frac{n\widehat{\sigma_1^2}}{\widehat{\sigma_1^2}} + m\log\Big(2\pi\widehat{\sigma_2^2}\Big) + \frac{m\widehat{\sigma_2^2}}{\widehat{\sigma_2^2}} + 4\log n$$

$$BIC_2 = (n+m)\log(2\pi) + n\log\widehat{\sigma_1^2} + m\log\widehat{\sigma_2^2} + (n+m) + 4\log n$$

$$BIC_2 = (n+m)(\log(2\pi)+1) + n\log\sigma_1^2 + m\log\sigma_2^2 + 4\log n \qquad (4.22)$$

Caso 3: $\mu_1 \neq \mu_2$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$

No caso em que $\mu_1 \neq \mu_2$ $\sigma_1^2 = \sigma_2^2 = \sigma^2$, tem-se três parâmetros desconhecidos μ_1, μ_2 e σ^2, que devem ser estimados a fim de obter a estimativa da função suporte. De (4.5) tem-se:

$$L(\boldsymbol{\theta}) = -\frac{n+m}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\mu_1)^2 + \sum_{i=n+1}^{n+m}(y_i-\mu_2)^2\right] \qquad (4.23)$$

em que $\boldsymbol{\theta} = \big(\mu_1, \mu_2, \sigma^2\big)$.

A função suporte estimada é dada por

$$L\Big(\widehat{\boldsymbol{\theta}}\Big) = -\frac{m+n}{2}\left(\log\Big(2\pi\widehat{\sigma^2}\Big) + 1\right) \qquad (4.24)$$

Sendo os estimadores de μ_1, μ_2, e σ^2 dados respectivamente por:

$$\widehat{\mu_1} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (4.25)$$

$$\widehat{\mu_2} = \frac{\sum_{i=n+1}^{n+m} y_i}{m} \qquad (4.26)$$

$$\widehat{\sigma^2} = \frac{1}{(n+m)}\left[\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2 + \sum_{i=n+1}^{n+m}(y_i-\widehat{\mu_2})^2\right] \qquad (4.27)$$

Substituindo (4.24) em (4.10) tem-se:

$$AIC_3 = -2\left(-\frac{m+n}{2}\left(\log\left(2\pi\widehat{\sigma^2}\right)+1\right)\right)+2\times 3$$

$$AIC_3 = (m+n)\left(\log\left(2\pi\widehat{\sigma^2}\right)+1\right)+6$$

$$AIC_3 = (n+m)\log\widehat{\sigma^2}+(n+m)\left(\log 2\pi+1\right)+6 \tag{4.28}$$

Sendo valor do BIC dado por

$$BIC = -2\left(\log L\left(\widehat{\boldsymbol{\theta}}\right)\right)+(k)\log n, \tag{4.29}$$

substitui-se (4.24) em (4.29), e tem-se:

$$BIC_3 = -2\left(-\frac{m+n}{2}\left(\log\left(2\pi\widehat{\sigma^2}\right)+1\right)\right)+3\log n$$

E assim

$$BIC_3 = (n+m)\log\widehat{\sigma^2}+(n+m)\left(\log 2\pi+1\right)+3\log n \tag{4.30}$$

Caso 4: $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 \neq \sigma_2^2$

Neste caso tem-se 3 parâmetros desconhecidos μ, σ_1^2, e σ_2^2, e $\boldsymbol{\theta} = \left(\mu, \sigma_1^2, \sigma_2^2\right)$.

Assim sendo, tem-se em (4.5):

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma_1^2\right)-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma_1^2}-\frac{m}{2}\log\left(2\pi\sigma_2^2\right)-\frac{\sum_{i=n+1}^{n+m}(y_i-\mu)^2}{2\sigma_2^2}. \tag{4.31}$$

E assim

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{(n+m)}{2}\left(\log 2\pi + 1\right) - \frac{n}{2}\log\widehat{\sigma_1^2} - \frac{m}{2}\log\widehat{\sigma_2^2} \tag{4.32}$$

Sendo que

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2 \tag{4.33}$$

$$\widehat{\sigma_2^2} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i - \widehat{\mu})^2 \tag{4.34}$$

e o estimador de μ é encontrado resolvendo-se a equação

$$\widehat{\mu^3} + A\widehat{\mu^2} + B\widehat{\mu} + C = 0 \tag{4.35}$$

em que A, B e C, são dados respectivamente por (6.34), (6.35) e (6.36).

O passo seguinte é obter o valor de AIC. Substituindo (4.32) em (4.10) tem-se:

$$AIC_4 = -2\left[-\frac{(n+m)}{2}\left(\log 2\pi + 1\right) - \frac{n}{2}\log\widehat{\sigma_1^2} - \frac{m}{2}\log\widehat{\sigma_2^2}\right] + 2\times 3$$

$$AIC_4 = (n+m)\left(\log 2\pi + 1\right) + n\log\widehat{\sigma_1^2} + m\log\widehat{\sigma_2^2} + 6 \tag{4.36}$$

E finalmente para obter-se o BIC

$$BIC = -2\left(\log L\left(\widehat{\boldsymbol{\theta}}\right)\right) + (k)\log n, \tag{4.37}$$

será substituído (4.32) em (4.37) e daí

$$BIC_4 = -2\left(-\frac{(n+m)}{2}\left(\log 2\pi + 1\right) - \frac{n}{2}\log\widehat{\sigma_1^2} - \frac{m}{2}\log\widehat{\sigma_2^2}\right) + 3\log n,$$

e o valor do BIC é dado por:

$$BIC_4 = (n+m)\left(\log 2\pi + 1\right) + n\log\widehat{\sigma_1^2} + m\log\widehat{\sigma_2^2} + 3\log n \qquad (4.38)$$

4.3 Seleção de variáveis em modelos de regressão

Supondo que se tenha uma variável resposta Y e m variáveis explicativas $X_1, X_2, \ldots, X_m$. O modelo de regressão linear múltipla é dado por

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \varepsilon,$$

em que o erro $\varepsilon \sim N(0, \sigma^2)$.

A distribuição condicional da variável resposta Y dado as variáveis explicativas é

$$f\left(Y|X_1, \ldots, X_m\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y - \beta_0 - \sum_{j=1}^{m}\beta_j X_j\right)^2\right].$$

Assim, se houver um conjunto com n observações, sendo estas independentes $\left\{\left(Y_i, X_{i1}, \ldots, X_{im}\right); i = 1, \ldots, n\right\}$, a verossimilhança para o modelo será dada por

$$L\left(\beta_0, \beta_1, \ldots, \beta_m, \sigma^2\right) = \prod_{i=1}^{n} p\left(Y_i|X_{i1}, \ldots X_{im}\right).$$

Assim, a função suporte será:

$$L\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{m}\beta_j X_{ij}\right)^2, \qquad (4.39)$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)$, sendo que seu estimador de máxima verossimilhança $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_m}\right)$, é obtido como solução do sistema de equações lineares

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{Y},$$

em que

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1m} \\ 1 & X_{21} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{nm} \end{bmatrix}, \text{ e } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.$$

O estimador de máxima verossimilhança de $\widehat{\sigma}^2$ é:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \ldots + \widehat{\beta}_m X_{im}\right)\right\}^2. \qquad (4.40)$$

Substituindo (4.40) em (4.39) tem-se a função suporte maximizada

$$L\left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_m, \widehat{\sigma}^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log d\left(X_1, \ldots, X_m\right) - \frac{n}{2}, \quad (4.41)$$

em que $d\left(X_1, \ldots, X_m\right)$ é a estimativa da variância residual σ^2 do modelo, dada em (4.40).

Como o número de parâmetros a serem estimados no modelo de regressão múltipla é $m + 2$, o AIC deste modelo de acordo com a equação (3.5) será dado

por:

$$AIC = n\left(\log 2\pi + 1\right) + n \log d\left(X_1, \ldots, X_m\right) + 2\left(m + 2\right). \tag{4.42}$$

Na regressão múltipla, nem todas as variáveis explicativas necessariamente influenciarão significativamente a variável resposta. Um modelo estimado com um grande número de variáveis explicativas desnecessárias pode ser instável. Selecionando o modelo com o menor AIC para todas as diferentes possíveis combinações da variável explicativa, espera-se obter um modelo razoável, que equilibre a qualidade do ajuste e a complexidade.

O BIC para este modelo, conforme (3.22), será dado por

$$BIC = n\left(\log 2\pi + 1\right) + n \log d\left(X_1, \ldots, X_m\right) + 2\left(m + 2\right) \log n. \tag{4.43}$$

4.4 Seleção de modelos para os dados M&M e produção de biomassa

Todos os cálculos foram feitos utilizando-se o software R.

4.4.1 Análise dos dados dos pesos de M&M

Para o caso em que $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$ foi obtido:

$$\begin{gathered}
\widehat{\mu} = 0.9138936 \\
\widehat{\sigma^2} = 0.0009435844 \\
L\left(\widehat{\boldsymbol{\theta}}\right) = 97.00677, \\
AIC_1 = -190.0135 \\
BIC_1 = -186.3132,
\end{gathered}$$

Para o segundo caso, em que $\mu_1 \neq \mu_2$ e $\sigma_1^2 \neq \sigma_2^2$ tem-se:

$$\widehat{\mu_1} = 0.9172692$$

$$\widehat{\mu_2} = 0.9097143$$
$$\widehat{\sigma_1^2} = 0.001099581$$
$$\widehat{\sigma_2^2} = 0.0007188707$$
$$L\left(\widehat{\boldsymbol{\theta}}\right) = 97.87383$$
$$AIC_2 = -187.7477$$
$$BIC_2 = -180.3471.$$

Para o terceiro caso, em que $\mu_1 \neq \mu_2$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$ tem-se:

$$\widehat{\mu_1} = 0.9172692$$
$$\widehat{\mu_2} = 0.9097143$$
$$\widehat{\sigma^2} = 0.0009294766$$
$$L\left(\widehat{\boldsymbol{\theta}}\right) = 97.36078$$
$$AIC_3 = -188.7216$$
$$BIC_3 = -183.1711.$$

Para o quarto caso, em que $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 \neq \sigma_2^2$ tem-se:

$$\widehat{\mu} = 0.9128487670$$
$$\widehat{\sigma_1^2} = 0.001119122$$
$$\widehat{\sigma_2^2} = 0.0007188707$$
$$L\left(\widehat{\boldsymbol{\theta}}\right) = 97.64484$$
$$AIC_4 = -189.2897$$
$$BIC_4 = -183.7392.$$

Comparando-se os valores do AIC , obtidos $(AIC_1, AIC_2, AIC_3, AIC_4)$, vê-se que deve-se selecionar o modelo 1, em que $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$, ou seja, pelo critério de Akaike, é mais provável que os pesos dos M&M tenham distribuição normal, com mesma média e mesma variância.

Ao se comparar os modelos utilizando o BIC, os resultados obtidos são os mesmos que aqueles obtidos pelo AIC, ou seja, os dados seguem a distribuição normal, com mesma média e mesma variância, haja vista que o valor do BIC_1 foi o menor deles.

4.4.2 Análise dos dados da produção de biomassa na grama de pântano.

Na Tabela 1 abaixo, tem-se o resultado do AIC e BIC para os 32 modelos possíveis de se obter com os dados.

A partir desta tabela, seleciona-se pelo AIC o modelo que tem pH e Na como sendo o mais provável. O modelo final selecionado foi

$$Y = -475.72892 + 404.94836 \times pH - 0.02333 \times Na.$$

A seleção pelo critério BIC não difere em seus resultados do critério AIC, selecionando o mesmo modelo como sendo o mais provável.

A dificuldade aqui encontrada é ao fazer-se os cálculos para todos os modelos possíveis, pois se houver N variáveis, tem-se 2^N modelos possíveis. Nesse exemplo, o número de variáveis é relativamente pequeno, mas se houvesse, por exemplo, dez variáveis, tería-se $2^{10} = 1024$ modelos possíveis.

Seria impraticável trabalhar com tantos modelos, o que se faz então é uma pré seleção das variáveis utilizando stepwise, ou outro método, e somente depois calcula-se o AIC e o BIC para tais modelos pré selecionados.

TABELA 1: Resultados do estudo da produção aérea de biomassa na grama de pântano.

Modelo	$\widehat{\sigma^2}$	$\widehat{log(L(\widehat{\theta}))}$	AIC	BIC
Y=μ+ε	426021.44	−355.50	715.01	718.62
Y=SAL+ε	421487.01	−355.26	716.52	721.94
Y=pH+ε	170679.44	−334.92	675.84	681.26
Y=K+ε	408179.80	−354.54	715.08	720.50
Y=Na+ε	394486.72	−353.77	713.54	718.96
Y=Zn+ε	259921.99	−344.39	694.77	700.19
Y=SAL+pH+ε	168961.07	−334.69	677.39	684.62
Y=SAL+K+ε	403264.55	−354.27	716.54	723.76
Y=SAL+Na+ε	392962.59	−353.69	715.37	722.60
Y=SAL+Zn+ε	190594.81	−337.41	682.81	690.04
Y=pH+K+ε	150140.21	−332.04	672.07	679.30
Y=pH+Na+ε	145514.93	−331.33	670.67	677.89
Y=pH+Zn+ε	166880.94	−334.42	676.83	684.06
Y=K+Na+ε	394351.87	−353.76	715.53	722.76
Y=K+Zn+ε	249136.22	−343.43	694.86	702.09
Y=Na+Zn+ε	242819.41	−342.85	693.71	700.93
Y=SAL+pH+K+ε	148179.33	−331.74	673.48	682.52
Y=SAL+pH+Na+ε	145253.20	−331.29	672.58	681.62
Y=SAL+pH+Zn+ε	154797.34	−332.72	675.45	684.48
Y=SAL+K+Na+ε	392958.57	−353.69	717.37	726.40
Y=SAL+K+Zn+ε	180423.99	−336.17	682.34	691.38
Y=SAL+Na+Zn+ε	185562.41	−336.80	683.61	692.64
Y=pH+K+Na+ε	144694.09	−331.21	672.41	681.44
Y=pH+K+Zn+ε	148217.11	−331.75	673.49	682.53
Y=pH+Na+Zn+ε	143803.24	−331.07	672.13	681.17
Y=K+Na+Zn+ε	242818.98	−342.85	695.71	704.74
Y=SAL+pH+K+Na+ε	144121.58	−331.12	674.23	685.07
Y=SAL+pH+K+Zn+ε	138517.20	−330.22	672.45	683.29
Y=SAL+pH+Na+Zn+ε	139832.73	−330.44	672.87	683.71
Y=SAL+K+Na+Zn+ε	180079.53	−336.13	684.26	695.10
Y=pH+K+Na+Zn+ε	143070.72	−330.95	673.90	684.74
Y=SAL+pH+K+Na+Zn+ε	797841.82	−369.62	753.24	765.89

5 CONCLUSÕES

Diante do problema da seleção de modelos, pode-se utilizar os critérios de informação Bayesiano e de Akaike para se selecionar modelos satisfatoriamente. Esses critérios baseiam-se em conceitos de fundamental importância, a verossimilhança, a Informação e a Entropia.

O AIC e o BIC podem ser utilizados nas mais diversas áreas; em estatística são amplamente utilizados principalmente em séries temporais e regressão; entretanto a regressão, a geoestatística e outras áreas também utilizam estes critérios.

Nesse trabalho, utilizou-se satisfatoriamente, os critérios para seleção de modelos normais e modelos de regressão; os resultados obtidos foram os mesmos nas aplicações feitas, mas nem sempre isto ocorre, conforme será demonstrado em trabalhos posteriores.

6 ESTUDOS FUTUROS

- Avaliar via simulação via Monte Carlo os desempenhos dos critérios AIC e BIC;
- Comparar o AIC e o BIC com um terceiro e recente método, a Medida L;
- Aplicação e comparação do AIC e BIC em séries temporais, onde estes são amplamente utilizados;
- Avaliar a utilização desses critérios em dados censurados, em que a verossimilhança não pode ser calculada (somente a verossimilhança parcial).

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control,** Boston, v.19, n.6, p.716–723, Dec. 1974.

ASH, R.B. **Information theory**. Illinois: Academic, 1965. 339p.

BOLFARINE, H.; SANDOVAL, M.C. **Introdução á inferência estatística**. São Paulo: Sociedade Brasileira de Matemática, 2000. 125p.

BURNHAM, K.P.; ANDERSON, D.R. **Model selection and multimodel inference:** a pratical information-theoretic approach. New York: Springer, 2002. 488p.

BURNHAM, K.P.; ANDERSON, D.R. Multimodel inference: understanding aic and bic in model selection. **Sociological Methods and Research,** Beverly Hills, v.33, n.2, p.261–304, May 2004.

CHAKRABARTI, C.G.; CHAKRABARTY, I. Boltzmann entropy : probability and information. **Romanian Journal of Physics,** Bucharest, v.52, n.5-6, p.525–528, Jan. 2007.

COVER, T.M.; THOMAS, J.A. **Elements of information theory**. New York: J. Wiley, 1991. 542p.

CRAMÉR, H. **Elementos da teoria de probabilidade e algumas de suas aplicações**. São Paulo: Mestre Jou, 1973. 330p.

DOMINGUES, H.H. **Espaços métricos e introdução à topologia**. São Paulo: Atual, 1982. 183p.

DRAPER, N.R.; SMITH, H. **Applied regression analysis**. 3. ed. New York: J. Wiley, 1998. 706p.

FERNANDES, R. de M.S.; AZEVEDO, T. de S. **Teoria da informação e suas aplicações em compressão e aleatoriedade**. Rio de Janeiro: PESC - COPPE, 2006. Notas de aula. Disponível em: <http://www.ravel.ufrj.br/arquivosPublicacoes/cos702_Rafael_Tiago.pdf>. Acesso em: 20 jul. 2008.

FERREIRA, D.F. **Estatística básica**. Lavras: UFLA, 2005. 664p.

GARBI, G.G. **O romance das equações algébricas:** a história da álgebra. São Paulo: Makron Books, 1997. 253p.

GHOSH, J.K.; SAMANTA, T. Model selection - an overview. **Current Science**, Bangalore, v.80, n.9, p. 1135–1144, May 2001.

HALLIDAY; RESNICK; WALKER. **Fundamentos de física 2:** gravitação, ondas e termodinâmica. 4. ed. Rio de Janeiro: LTC, 1996.

HUANG, K. **Statistical mechanics**. 2. ed. Singapore: J. Wiley, 1987. 493p.

JOHNSON, N.L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions**. 2. ed. New York: J. Wiley, 1994. 756p.

KONISHI, S.; KITAGAWA, G. **Information criteria and statistical modeling**. New York: Springer, 2008. 321p.

MACKAY, D.J. **Information theory, inference, and learning algorithms**. 4. ed. London: Cambridge, 2005. 628p.

MARTINS, R.C. Sobre a atualidade de proposições de Ludwig Boltzmann. **Revista da SBHC,** São Paulo, n.13, p.81–94, 1995.

MAZEROLLE, M.J. **Mouvements et reproduction des amphibiens en tourbières perturbées.** 2004. 78p. Tese (Doutorado em Ciências Florestais) - Université Laval, Québec.

MEYER, P.L. **Probabilidade:** aplicações à estatística. 2. ed. Rio de Janeiro: LTC, 1983. 421p.

MIRANDA, C.G. **O método lasso para o modelo de Cox e sua comparação com propostas tradicionais de seleção de variáveis.** 2006. 97p. Tese (Doutorado em Estatística) - Universidade Federal de Minas Gerais. Belo Horizonte.

MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. **Introduction to the theory of statistics**. 3. ed. New York: J. Wiley, 1974. 564p.

NUSSENZVEIG, H.M. **Curso de física básica 2:** fluidos; oscilações e calor; ondas. 3. ed. São Paulo: E. Blücher, 1981. 315p.

PAULINO, C.D.; TURKMAN, A.A.; MURTEIRA, B.J. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 280p.

RAWLINGS, J.O.; PANTULA, S.G.; DICKEY, D.A. **Applied regression analysis:** a research tool. 2. ed. New York: Springer, 1998. 657p.

RIBEIRO, J.C. **Teoria da informação** - módulo I. Rio de Janeiro, 2007. Notas de aula. Disponível em: <http://pasta.ebah.com.br/download/apostila-teoria-da-informacao-pdf-3985>. Acesso em: 16 jul. 2008.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics,** Hayward, v.6, n.2, p.461–464, Mar. 1978.

SHANNON, C.E. A mathematical theory of communication. **The Bell System Technical Journal,** New York, v.27, p.623–656, Oct. 1948.

SILVA, R.T. da. **Conservação da energia.** Recife, 2005. (Notas de aula). Disponível em: <http//www.fisica.ufpb.br/˜romero>. Acesso em: 18 jul. 2008.

SILVA, V. M.M. da. **Teoria da informação e codificação**. Coimbra: DEEC-FCTUC, 2008. Notas de apoio. Disponível em: <https://woc.uc.pt/deec/class/getmaterial.do?idclass=334&idyear=4>. Acesso em: 20 nov. 2008.

STEVENSON, W.J. **Estatística aplicada à administração**. São Paulo: Harbra, 2001. 495p.

TRIOLA, M.F. **Introdução à estatística**. 7. ed. Rio de Janeiro: LTC, 1999. 410p.

VICKI, V. **A história da criptologia.** Disponível em: <http://www.numaboa.com/criptografia/historia/553-Shannon>. Acesso em: 20 nov. 2007.

WASSERMAN, L. **All of statistics:** a concise course in statistical inference. New York: Springer, 2005. 322p.

WIENER, N. **Cibernética:** ou, controle e comunicação no animal e na máquina. São Paulo: Polígono / Universidade de São Paulo, 1970. 256p.

YOUNG, H.; FISHER, R. **Física II:** termodinâmica e ondas. 10. ed. São Paulo: Pearson Education do Brasil, 2003.

ANEXOS

ANEXO **Páginas**

ANEXO A

TABELA 2: Dados utilizados no estudo de pesos (em gramas) de uma amostra de confeitos M&M.

Observação	Amarelo	Vermelho
1	0.906	0.870
2	0.978	0.933
3	0.926	0.952
4	0.868	0.908
5	0.876	0.911
6	0.968	0.908
7	0.921	0.913
8	0.893	0.983
9	0.939	0.920
10	0.886	0.936
11	0.924	0.891
12	0.910	0.924
13	0.877	0.874
14	0.879	0.908
15	0.941	0.924
16	0.879	0.897
17	0.940	0.912
18	0.960	0.888
19	0.989	0.872
20	0.900	0.898
21	0.917	0.882
22	0.911	
23	0.892	
24	0.886	
25	0.949	
26	0.934	

ANEXO B

TABELA 3: Dados utilizados no estudo das características que influenciam a produção aérea de biomassa na grama de pântano.

Y	SAL	pH	K	Na	Zn
676	33	5.00	1441.67	35185.5	16.4524
516	35	4.75	1299.19	28170.4	13.9852
1052	32	4.20	1154.27	26455.0	15.3276
868	30	4.40	1045.15	25072.9	17.3128
1008	33	5.55	521.62	31664.2	22.3312
436	33	5.05	1273.02	25491.7	12.2778
544	36	4.25	1346.35	20877.3	17.8225
680	30	4.45	1253.88	25621.3	14.3516
640	38	4.75	1242.65	27587.3	13.6826
492	30	4.60	1281.95	26511.7	11.7566
984	30	4.10	553.69	7886.5	9.8820
1400	37	3.45	494.74	14596.0	16.6752
1276	33	3.45	525.97	9826.8	12.3730
1736	36	4.10	571.14	11978.4	9.4058
1004	30	3.50	408.64	10368.6	14.9302
396	30	3.25	646.65	17307.4	31.2865
352	27	3.35	514.03	12822.0	30.1652
328	29	3.20	350.73	8582.6	28.5901
392	34	3.35	496.29	12369.5	19.8795
236	36	3.30	580.92	14731.9	18.5056
392	30	3.25	535.82	15060.6	22.1344
268	28	3.25	490.34	11056.3	28.6101
252	31	3.20	552.39	8118.9	23.1908
236	31	3.20	661.32	13009.5	24.6917
340	35	3.35	672.15	15003.7	22.6758
2436	29	7.10	528.65	10225.0	0.3729
2216	35	7.35	563.13	8024.2	0.2703
2096	35	7.45	497.96	10393.0	0.3205
1660	30	7.45	458.38	8711.6	0.2648
2272	30	7.40	498.25	10239.6	0.2105
824	26	4.85	936.26	20436.0	18.9875
1196	29	4.60	894.79	12519.9	20.9687

...continua...

Continuação da TABELA 3.

Y	SAL	pH	K	Na	Zn
1960	25	5.20	941.36	18979.0	23.9841
2080	26	4.75	1038.79	22986.1	19.9727
1764	26	5.20	898.05	11704.5	21.3864
412	25	4.55	989.87	17721.0	23.7063
416	26	3.95	951.28	16485.2	30.5589
504	26	3.70	939.83	17101.3	26.8415
492	27	3.75	925.42	17849.0	27.7292
636	27	4.15	954.11	16949.6	21.5699
1756	24	5.60	720.72	11344.6	19.6531
1232	27	5.35	782.09	14752.4	20.3295
1400	26	5.50	773.30	13649.8	19.5880
1620	28	5.50	829.26	14533.0	20.1328
1560	28	5.40	856.96	16892.2	19.2420

ANEXO C

Derivação do viés da Função suporte

O estimador de $\boldsymbol{\theta}$ é o vetor de parâmetros p-dimensional $\widehat{\boldsymbol{\theta}}$ que maximiza a função (2.21). Tal estimador é obtido somo solução de (2.22). Isto é, deve-se achar a solução de

$$\frac{\partial L\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(X_i|\boldsymbol{\theta}\right) = 0$$

Tomando a esperança, tem-se:

$$E_{G(X_n)}\left[\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(X_i|\boldsymbol{\theta}\right)\right] = nE_{G(z)}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(Z|\boldsymbol{\theta}\right)\right.$$

Assim, para um modelo contínuo, se $\boldsymbol{\theta_0}$ é solução de

$$E_{G(z)}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(Z|\boldsymbol{\theta}\right)\right] = \int g\left(z\right) \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(z|\boldsymbol{\theta}\right) dz = \mathbf{0}, \tag{6.1}$$

pode ser mostrado que o estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ converge em probabilidade para $\boldsymbol{\theta_0}$ quando $n \to \infty$.

Usando o resultado acima, pode-se avaliar o viés dado por (3.1), quando a função suporte esperada é estimada usando a log verossimilhança do modelo.

O viés

$$b\left(G\right) = E_{G(\boldsymbol{x_n})}\left[\log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right) - nE_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}\left(X_n\right)\right)\right]\right], \tag{6.2}$$

pode ser decomposto como

$$\begin{aligned} b\left(G\right) =& E_{G(\mathbf{x_n})}\left[\log f\left(\mathbf{X_n}|\widehat{\theta}\left(\mathbf{X_n}\right)\right) - nE_{G(Z)}\left[\log f\left(Z|\hat{\theta}\left(\mathbf{X_n}\right)\right)\right]\right] \\ =& E_{G(\mathbf{x_n})}\left[\log f\left(\mathbf{X_n}|\widehat{\theta}\left(\mathbf{X_n}\right)\right) - \log f\left(\mathbf{X_n}|\theta_0\right)\right] \\ &+ E_{G(\mathbf{x_n})}\left[\log f\left(\mathbf{X_n}|\theta_0\right) - nE_{G(Z)}\left[\log f\left(Z|\theta_0\right)\right]\right] \\ &+ E_{G(\mathbf{x_n})}\left[nE_{G(Z)}\left[\log f\left(Z|\theta_0\right)\right] - nE_{G(Z)}\left[\log f\left(Z|\hat{\theta}\left(\mathbf{X_n}\right)\right)\right]\right] \\ =& D_1 + D_2 + D_3. \end{aligned} \tag{6.3}$$

Esquematicamente tem-se a Figura 6 abaixo:

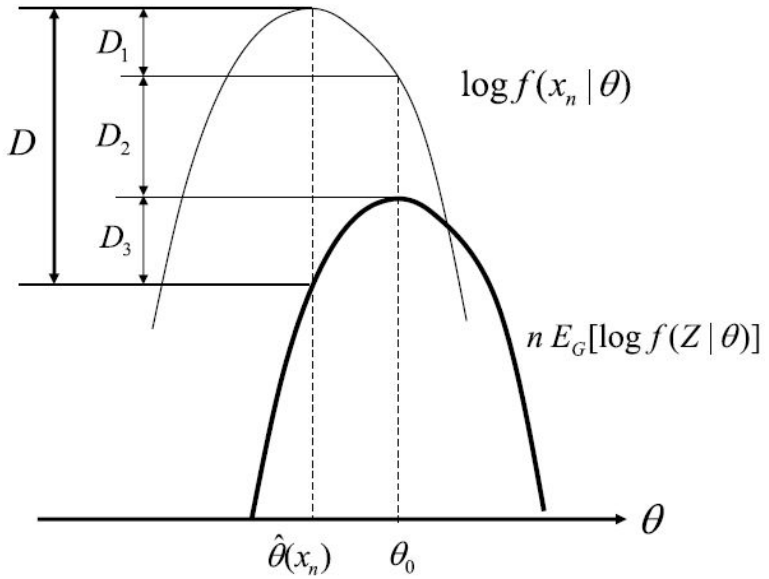

FIGURA 6: Decomposição dos termos do viés.

1 - Cálculo de D_2. Primeiramente será feito este caso, por se tratar do mais simples, pois não contém nenhum estimador. Assim:

$$\begin{aligned} D_2 &= E_{G(\mathbf{x_n})}\left[\log f\left(\mathbf{X_n}|\boldsymbol{\theta_0}\right) - nE_{G(Z)}\left[\log f\left(Z|\boldsymbol{\theta_0}\right)\right]\right] \\ &= E_{G(\mathbf{x_n})}\left[\log f\left(\mathbf{X_n}|\boldsymbol{\theta_0}\right)\right] - nE_{G(Z)}\left[\log f\left(Z|\boldsymbol{\theta_0}\right)\right] \\ &= E_{G(\mathbf{x_n})}\left[\sum_{i=1}^{n}\log f\left(X_i|\boldsymbol{\theta_0}\right)\right] - nE_{G(Z)}\left[\log f\left(Z|\boldsymbol{\theta_0}\right)\right] \\ &= 0 \end{aligned} \tag{6.4}$$

Isto mostra que na Figura 6, embora D_2 varie aleatoriamente dependendo dos dados, sa esperança é zero.

2 - Cálculo de D_3. Para simplicidade das fórmulas, escreva-se primeiramente

$$\eta\left(\boldsymbol{\theta}\right) := E_{G(Z)}\left[\log f\left(Z|\widehat{\boldsymbol{\theta}}\left(\boldsymbol{X}_n\right)\right)\right].$$

Pela expansão em série de Taylor de $\eta\left(\widehat{\boldsymbol{\theta}}\right)$ em torno de $\boldsymbol{\theta_0}$, sendo este solução de (6.1), obtém-se:

$$\begin{aligned}\eta\left(\widehat{\boldsymbol{\theta}}\right) &= \eta\left(\boldsymbol{\theta_0}\right) + \sum_{i=1}^{p}\left(\theta_i - \theta_i^{(0)}\right)\frac{\partial\eta\left(\boldsymbol{\theta_0}\right)}{\partial\theta_i} \\ &+ \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\theta_i - \theta_i^{(0)}\right)\left(\theta_j - \theta_j^{(0)}\right)\frac{\partial^2\eta\left(\boldsymbol{\theta_0}\right)}{\partial\theta_i\partial\theta_j} + \ldots, \qquad (6.5)\end{aligned}$$

em que $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots \widehat{\theta}_p\right)^T$ e $\boldsymbol{\theta_0} = \left(\theta_1^{(0)}, \theta_2^{(0)} \ldots \theta_p^{(0)}\right)^T$. Como $\boldsymbol{\theta_0}$ é solução de (6.1) tem-se

$$\frac{\partial\eta\left(\boldsymbol{\theta_0}\right)}{\partial\theta_i} = E_{G(Z)}\left[\frac{\partial}{\partial\theta_i}\log f\left(Z|\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta_0}}\right] = 0, i = 1, 2, \ldots, p.$$

Assim, (6.5) pode ser aproximado por:

$$\eta\left(\widehat{\boldsymbol{\theta}}\right) = \eta\left(\boldsymbol{\theta_0}\right) - \frac{1}{2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right)^T J\left(\boldsymbol{\theta_0}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}\right),$$

sendo $J\left(\boldsymbol{\theta_0}\right)$ uma $p \times p$ matriz dada por

$$J\left(\boldsymbol{\theta_0}\right) = -E_{G(Z)}\left[\frac{\partial^2\log f\left(Z|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta_0}}\right] = -\int g\left(z\right)\frac{\partial^2\log f\left(Z|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta_0}} dz, \qquad (6.6)$$

e o (a, b)-ésimo elemento é dado por

$$j_{ab} = -E_{G(Z)}\left[\frac{\partial^2\log f\left(Z|\boldsymbol{\theta}\right)}{\partial\theta_a\partial\theta_b}\Big|_{\boldsymbol{\theta_0}}\right] = -\int g\left(z\right)\frac{\partial^2\log f\left(Z|\boldsymbol{\theta}\right)}{\partial\theta_a\partial\theta_b}\Big|_{\boldsymbol{\theta_0}} dz$$

Como D_3 é justamente a esperança de $\eta(\boldsymbol{\theta}_0) - \eta\left(\widehat{\boldsymbol{\theta}}\right)$, com respeito a $G(\boldsymbol{X_n})$, obtém-se a aproximação:

$$\begin{aligned} D_3 &= E_{G(\boldsymbol{X_n})}\left[nE_{G(Z)}\left[\log f\left(Z|\boldsymbol{\theta}_0\right)\right] - n\log f\left(Z|\widehat{\boldsymbol{\theta}}\right)\right] \\ &= \frac{n}{2}E_{G(X_n)}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T J\left(\boldsymbol{\theta}_0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\right] \\ &= \frac{n}{2}E_{G(X_n)}\left[tr\left\{J\left(\widehat{\boldsymbol{\theta}}_0\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0\right)^T\right\}\right] \\ &= \frac{n}{2}tr\left\{J\left(\widehat{\boldsymbol{\theta}}_0\right)E_{G(X_n)}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T\right]\right\}. \end{aligned} \tag{6.7}$$

Pelas propriedades assintóticas dos estimadores de máxima verossimilhança dadas no Teorema 2.1, tem-se que:

$$E_{G(X_n)}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T\right] = \frac{1}{n}J\left(\boldsymbol{\theta}_0\right)^{-1} I\left(\boldsymbol{\theta}_0\right) J\left(\boldsymbol{\theta}_0\right)^{-1}, \tag{6.8}$$

deste modo pela substituição de (6.7) em (6.8), tem-se:

$$D_3 = \frac{1}{2}tr\left\{I\left(\boldsymbol{\theta}_0\right) J\left(\boldsymbol{\theta}_0\right)^{-1}\right\}, \tag{6.9}$$

sendo que $J(\boldsymbol{\theta})$ é dada por (6.6) e $I(\boldsymbol{\theta})$ é a $p \times p$ matriz dada por

$$\begin{aligned} I\left(\boldsymbol{\theta}_0\right) &= E_{G(Z)}\left[\left.\frac{\partial \log f\left(z|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\frac{\partial \log f\left(z|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0}\right] \\ &= \int g\left(z\right)\left.\frac{\partial \log f\left(z|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\frac{\partial \log f\left(z|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0} dz. \end{aligned} \tag{6.10}$$

Resta agora o cálculo de D_3.

3 - Cálculo de D_1. Reescrevendo $L(\boldsymbol{\theta}) = \log f\left(\boldsymbol{X}_n|\boldsymbol{\theta}\right)$, em termos da sua expansão em séries de Taylor, na vizinhança do estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$, obtém-se:

$$L\left(\widehat{\boldsymbol{\theta}}\right) = L(\boldsymbol{\theta}) + \left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \frac{\partial L\left(\widehat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}} + \frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)^T \frac{\partial^2 L\left(\widehat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right) + \ldots. \tag{6.11}$$

Em (6.11), $\widehat{\boldsymbol{\theta}}$ satisfaz a equação $\dfrac{\partial L\left(\widehat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}} = \mathbf{0}$, pelo fato de que o estimador de máxima verossimilhança é dado como solução de $\dfrac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$.

Tem-se que $\dfrac{1}{n}\dfrac{\partial^2 L\left(\widehat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \dfrac{1}{n}\dfrac{\partial^2 \log f\left(\boldsymbol{X_n}|\widehat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}$ converge em probabilidade para $J(\boldsymbol{\theta}_0)$ quando $n \to \infty$, isto vem do fato de que $\widehat{\boldsymbol{\theta}}$ converge para $\widehat{\boldsymbol{\theta}_0}$ e pode ser provado utilizando-se da lei dos grandes números.

Assim, tem-se de (6.11) que

$$L(\boldsymbol{\theta}_0) - L\left(\widehat{\boldsymbol{\theta}}\right) \approx \frac{n}{2}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right).$$

A partir deste resultado, juntamente com (6.8) pode-se calcular D_1.

$$\begin{aligned}
D_1 &= E_{G(\boldsymbol{X}_n)}\left[\log f\left(\boldsymbol{X}_{\mathbf{n}}|\widehat{\boldsymbol{\theta}}(\mathbf{X_n})\right) - \log f(\mathbf{X_n}|\theta_0)\right] \\
&= \frac{n}{2}E_{G(\boldsymbol{X_n})}\left[\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)^T J(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)\right] \\
&= \frac{n}{2}E_{G(\boldsymbol{X_n})}\left[tr\left\{J(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)^T\right\}\right] \\
&= \frac{n}{2}tr\left\{J(\boldsymbol{\theta}_0)\,E_{G(\boldsymbol{X_n})}\left[\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}\right)^T\right]\right\} \\
&= \frac{1}{2}tr\left\{I(\boldsymbol{\theta}_0)\,J(\boldsymbol{\theta}_0)^{-1}\right\} \qquad (6.12)
\end{aligned}$$

Assim, de (6.4), (6.9) e (6.12) tem-se que

$$\begin{aligned}
b(G) &= D_1 + D_2 + D_3 \\
&= \frac{1}{2}tr\left\{I(\boldsymbol{\theta}_0)\,J(\boldsymbol{\theta}_0)^{-1}\right\} + 0 + \frac{1}{2}tr\left\{I(\boldsymbol{\theta}_0)\,J(\boldsymbol{\theta}_0)^{-1}\right\}, \quad (6.13)
\end{aligned}$$

sendo $I(\boldsymbol{\theta}_0)$ e $J(\boldsymbol{\theta}_0)$ dados por (6.6) e (6.10), respectivamente.

ANEXO D

Função suporte para modelos normais.

Tem-se de (4.5) que de forma geral

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma_1^2\right) - \frac{\sum_{i=1}^{n}\left(y_i-\mu_1\right)^2}{2\sigma_1^2} - \frac{m}{2}\log\left(2\pi\sigma_2^2\right) - \frac{\sum_{i=n+1}^{n+m}\left(y_i-\mu_2\right)^2}{2\sigma_2^2}.$$

Desse modo, serão feitas aqui as derivações para os estimadores de máxima verossimilhança para os quatro casos descritos em (4.1), (4.2), (4.3), e (4.4).
Caso 1: $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$
Para este caso, tem-se por (4.6)

$$L\left(\boldsymbol{\theta}\right) = -\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}\left(y_i-\mu\right)^2, \tag{6.14}$$

sendo $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$.

Para maximizar (4.6) faça-se $\dfrac{\partial L\left(\mu,\sigma^2\right)}{\partial\sigma^2} = 0$ e $\dfrac{\partial L\left(\mu,\sigma^2\right)}{\partial\boldsymbol{\theta}} = 0$.
Derivando (4.6) em relação a σ^2, tem-se:

$$\frac{\partial L\left(\mu,\sigma^2\right)}{\partial\sigma^2} = \frac{\partial\left(-\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}\left(y_i-\mu\right)^2\right)}{\partial\sigma^2} = 0$$

$$\frac{\partial L\left(\mu,\sigma^2\right)}{\partial\sigma^2} = -\frac{n+m}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n+m}\left(y_i-\mu\right)^2 = 0$$

$$\frac{1}{\widehat{\sigma}^2}\left(-\frac{n+m}{2} + \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n+m}\left(y_i-\widehat{\mu}\right)^2\right) = 0 \Longrightarrow \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n+m}\left(y_i-\mu\right)^2 = \frac{n+m}{2}$$

$$\widehat{\sigma}^2 = \frac{1}{(n+m)}\sum_{i=1}^{n+m}\left(y_i-\widehat{\mu}\right)^2. \tag{6.15}$$

O estimador de σ^2 é dado por (6.15), e essa equação necessita do estimador de μ, que será encontrado abaixo:

$$\frac{\partial L\left(\mu,\sigma^2\right)}{\partial\mu} = \frac{\partial\left(-\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y_i-\mu\right)^2\right)}{\partial\mu} = 0$$

$$0 - \frac{1}{2\widehat{\sigma}^2} 2 \sum_{i=1}^{m} (y_i - \widehat{\mu})(-1) = 0 \Longrightarrow \sum_{i=1}^{n+m} (y_i - \widehat{\mu}) = 0$$

$$\sum_{i=1}^{n+m} (y_i - \widehat{\mu}) = 0 \sum_{i=1}^{n+m} y_i = \sum_{i=1}^{n+m} \widehat{\mu} \Longrightarrow \widehat{\mu} = \frac{1}{n+m} \sum_{i=1}^{n+m} y_i.$$

Desse modo o estimador de μ é dado por

$$\widehat{\mu} = \frac{1}{n+m} \sum_{i=1}^{n+m} y_i. \tag{6.16}$$

Substituindo os valores encontrados em (6.15) e (6.16) em (4.6), tem-se

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{n+m}{2} \log\left(2\pi\widehat{\sigma}^2\right) - \frac{1}{2\widehat{\sigma}^2} \sum_{i=1}^{n+m} (y_i - \widehat{\mu})^2$$

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{n+m}{2} \log\left(2\pi\widehat{\sigma}^2\right) - \frac{1}{2\widehat{\sigma}^2} (n+m)\, \widehat{\sigma}^2$$

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{n+m}{2} \log\left(2\pi\sigma^2\right) - \frac{n+m}{2}$$

Caso 2: $\mu_1 \neq \mu_2$ e $\sigma_1^2 \neq \sigma_2^2$
Nesse caso, tem-se por (4.14)

$$\begin{aligned} L(\boldsymbol{\theta}) &= L\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) = -\frac{n}{2} \log\left(2\pi\sigma_1^2\right) - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (y_i - \mu_1)^2 \\ &- \frac{m}{2} \log\left(2\pi\sigma_2^2\right) - \frac{1}{2\sigma_2^2} \sum_{i=n+1}^{n+m} (y_i - \mu_2)^2 \end{aligned} \tag{6.17}$$

Derivando (6.17) em relação a μ_1 e igualando a zero, tem-se:

$$\frac{\partial \left(-\frac{n}{2} \log\left(2\pi\sigma_1^2\right) - \dfrac{\sum_{i=1}^{n} (y_i - \mu_1)^2}{2\sigma_1^2} - \dfrac{m}{2} \log\left(2\pi\sigma_2^2\right) - \dfrac{\sum_{i=n+1}^{n+m} (y_i - \mu_2)^2}{2\sigma_2^2} \right)}{\partial \mu_1} = 0,$$

Obtendo assim

$$-\frac{2}{2\widehat{\sigma_1^2}}\sum_{i=1}^{n}\left(y_i-\widehat{\mu_1}\right)(-1)=0\Longrightarrow\sum_{i=1}^{n}\left(y_i-\widehat{\mu_1}\right)=0\Longrightarrow\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\widehat{\mu_1}$$

E finalmente encontra-se o estimador de μ_1, dado por

$$\widehat{\mu_1}=\frac{\sum_{i=1}^{n}y_i}{n} \tag{6.18}$$

Derivando (6.32) em relação a μ_2^2 e igualando a zero, tem-se:

$$\frac{\partial\left(-\frac{n}{2}\log\left(2\pi\sigma_1^2\right)-\frac{\sum_{i=1}^{n}\left(y_i-\mu_1\right)^2}{2\sigma_1^2}-\frac{m}{2}\log\left(2\pi\sigma_2^2\right)-\frac{\sum_{i=n+1}^{n+m}\left(y_i-\mu_2\right)^2}{2\sigma_2^2}\right)}{\partial\mu_2}=0$$

E assim

$$-\frac{2}{2\widehat{\sigma_2^2}}\sum_{i=n+1}^{n+m}\left(y_i-\widehat{\mu_2}\right)(-1)=0\Longrightarrow\sum_{i=n+1}^{n+m}\left(y_i-\widehat{\mu_2}\right)=0\Longrightarrow\sum_{i=n+1}^{n+m}y_i=\sum_{i=n+1}^{n+m}\widehat{\mu_2}$$

Assim, o estimador de μ_2, é

$$\widehat{\mu_2}=\frac{\sum_{i=n+1}^{n+m}y_i}{n} \tag{6.19}$$

Derivando (6.32) em relação a σ_1^2 e igualando a zero, tem-se

$$\frac{\partial L\left(\theta\right)}{\partial\sigma_1^2}=0$$

$$\frac{\partial\left(-\frac{n}{2}\log(2\pi\sigma_1^2)-\frac{\sum_{i=1}^{n}(y_i-\mu_1)^2}{2\sigma_1^2}-\frac{m}{2}\log(2\pi\sigma_2^2)-\frac{\sum_{i=n+1}^{n+m}(y_i-\mu_2)^2}{2\sigma_2^2}\right)}{\partial\sigma_1^2}=0,$$

$$-\frac{n}{2\widehat{\sigma_1^2}}+\frac{1}{2\left(\widehat{\sigma_1^2}\right)^2}\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2=0\Longrightarrow n=\frac{1}{\widehat{\sigma_1^2}}\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2$$

Finalmente obtém-se o estimador de σ_1^2, dado por

$$\widehat{\sigma_1^2}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2 \tag{6.20}$$

Nota-se que o estimador de σ_1^2 depende do estimador de μ_1, expresso por (6.18). Derivando (6.32) em relação a σ_2^2 e igualando a zero, tem-se

$$\frac{\partial\left(-\frac{n}{2}\log\left(2\pi\sigma_1^2\right)-\frac{\sum_{i=1}^{n}(y_i-\mu_1)^2}{2\sigma_2^2}-\frac{m}{2}\log(2\pi\sigma_2^2)-\frac{\sum_{i=n+1}^{n+m}(y_i-\mu_2)^2}{2\sigma_2^2}\right)}{\partial\sigma_1^2}=0$$

$$-\frac{m}{2\widehat{\sigma_2^2}}+\frac{1}{2\left(\widehat{\sigma_2^2}\right)^2}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu_2})^2=0\Longrightarrow\frac{m}{2}=\frac{1}{2\widehat{\sigma_2^2}}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu_2})^2$$

E assim, obtém-se o estimador de σ_2^2, dado por

$$\widehat{\sigma_2^2}=\frac{1}{m}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu_2})^2. \tag{6.21}$$

O estimador de σ_2^2 depende do estimador de μ_2 que é dado pela fórmula (6.19). Substituindo (6.18), (6.19), (6.20) e (6.21) em (6.17) tem-se:

$$L\left(\widehat{\boldsymbol{\theta}}\right)=-\frac{n}{2}\log\left(2\pi\widehat{\sigma_1^2}\right)-\frac{\sum_{i=1}^{n}(y_i-\widehat{\mu_1})^2}{2\widehat{\sigma_1^2}}-\frac{m}{2}\log(2\pi\sigma_2^2)-\frac{\sum_{i=n+1}^{m}(y_i-\widehat{\mu_2})^2}{2\widehat{\sigma_2^2}}. \tag{6.22}$$

Caso 3: $\mu_1 \neq \mu_2$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$
Sob a premissa de que $\mu_1 \neq \mu_2$ e $\sigma_1^2 = \sigma_2^2 = \sigma^2$ tem-se de (4.5):

$$L\left(\boldsymbol{\theta}\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{\sum\limits_{i=1}^{n}\left(y_i - \mu_1\right)^2}{2\sigma^2} - \frac{m}{2}\log\left(2\pi\sigma^2\right) - \frac{\sum\limits_{i=n+1}^{n+m}\left(y_i - \mu_2\right)^2}{2\sigma^2},$$

Daí

$$L(\boldsymbol{\theta}) = -\frac{n+m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mu_1)^2 + \sum_{i=n+1}^{n+m}\left(y_i - \mu_2\right)^2\right] \quad (6.23)$$

Afim de maximizar (6.23), faça-se $\dfrac{\partial L\left(\boldsymbol{\theta}\right)}{\partial\sigma^2} = 0$, $\dfrac{\partial L\left(\boldsymbol{\theta}\right)}{\partial\mu_1} = 0$, e $\dfrac{\partial L\left(\boldsymbol{\theta}\right)}{\partial\mu_2} = 0$.
Derivando (6.23) em relação a σ^2, tem-se

$$\frac{\partial L(\boldsymbol{\theta})}{\partial\sigma^2} = \frac{\partial\left\{-\frac{n+m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum\limits_{i=1}^{n}(y_i - \mu_1)^2 + \sum\limits_{i=n+1}^{n+m}\left(y_i - \mu_2\right)^2\right]\right\}}{\partial\sigma^2} = 0,$$

e assim

$$-\frac{n+m}{2\widehat{\sigma^2}} + \frac{1}{2\left(\widehat{\sigma^2}\right)^2}\left[\sum_{i=1}^{n}\left(y_i - \widehat{\mu_1}\right)^2 + \sum_{i=n+1}^{n+m}\left(y_i - \widehat{\mu_2}\right)^2\right] = 0$$

$$\Downarrow$$

$$(n+m) = \frac{1}{\widehat{\sigma^2}}\left[\sum_{i=1}^{n}\left(y_i - \widehat{\mu_1}\right)^2 + \sum_{i=n+1}^{n+m}\left(y_i - \widehat{\mu_2}\right)^2\right]$$

Desse modo,o estimador de σ^2 é dada por

$$\widehat{\sigma^2} = \frac{1}{(n+m)}\left[\sum_{i=1}^{n}\left(y_i - \widehat{\mu_1}\right)^2 + \sum_{i=n+1}^{n+m}\left(y_i - \widehat{\mu_2}\right)^2\right] \quad (6.24)$$

Vê-se assim que a estimador de σ^2 depende da estimador de μ_1 e μ_2. Tais estimadores serão encontradas abaixo:

Derivando (6.23) em relação a μ_1,tem-se:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_1} = \frac{\partial\left\{-\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(y_i-\mu_1\right)^2+\sum_{i=n+1}^{n+m}\left(y_i-\mu_2\right)^2\right]\right\}}{\partial \mu_1} = 0.$$

Assim

$$\frac{2}{2\widehat{\sigma^2}}\sum_{i=1}^{n}\left(y_i-\widehat{\mu_1}\right)(-1)=0 \Longrightarrow \sum_{i=1}^{n}\left(y_i-\widehat{\mu_1}\right)=0 \Longrightarrow \sum_{i=1}^{n} y_i = n\widehat{\mu_1}.$$

Logo, o estimador de μ_1 é dado por:

$$\widehat{\mu_1} = \frac{\sum\limits_{i=1}^{n} y_i}{n} \tag{6.25}$$

Para encontrar o estimador de μ_2, deve-se derivar (6.23) em relação a μ_2 e igualar a zero, assim:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_2} = \frac{\partial\left\{-\frac{n+m}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(y_i-\mu_1\right)^2+\sum_{i=n+1}^{n+m}\left(y_i-\mu_2\right)^2\right]\right\}}{\partial \mu_2} = 0.$$

Assim

$$\frac{2}{2\widehat{\sigma^2}}\sum_{i=n+1}^{n+m}\left(y_i-\widehat{\mu_2}\right)(-1)=0 \Longrightarrow \sum_{i=n+1}^{n+m}\left(y_i-\widehat{\mu_2}\right)=0 \Longrightarrow \sum_{i=n+1}^{n+m} y_i = n\widehat{\mu_2}.$$

Desse modo, o estimador de μ_2 é dado por:

$$\widehat{\mu_2} = \frac{\sum\limits_{i=n+1}^{n+m} y_i}{m} \tag{6.26}$$

Conseqüentemente, tem-se em (6.23)

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{m+n}{2}\log\left(2\pi\widehat{\sigma^2}\right) - \frac{1}{2\widehat{\sigma^2}}\left[\sum_{i=1}^{n}\left(y_i-\widehat{\mu_1}\right)^2 + \sum_{i=n+1}^{m}\left(y_i-\widehat{\mu_2}\right)^2\right]$$

$$L\Big(\widehat{\boldsymbol{\theta}}\Big) = -\frac{m+n}{2}\log\Big(2\pi\widehat{\sigma^2}\Big) - \frac{1}{2\widehat{\sigma^2}}\Big(n\widehat{\sigma^2} + m\widehat{\sigma^2}\Big)$$

$$L\Big(\widehat{\boldsymbol{\theta}}\Big) = -\frac{m+n}{2}\Big(\log\Big(2\pi\widehat{\sigma^2}\Big) + 1\Big)$$

Em que $\widehat{\sigma^2}$, $\widehat{\mu_1}$, e $\widehat{\mu_2}$, são dados por (6.24), (6.25) e (6.26) respectivamente.
Caso 4: $\mu_1 = \mu_2 = \mu$ e $\sigma_1^2 \neq \sigma_2^2$
De (4.5) tem-se:

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log\big(2\pi\sigma_1^2\big) - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma_1^2} - \frac{m}{2}\log\big(2\pi\sigma_2^2\big) - \frac{\sum_{i=n+1}^{n+m}(y_i-\mu)^2}{2\sigma_2^2} \quad (6.27)$$

A verossimilhança maximizada será dada por

$$L\Big(\widehat{\boldsymbol{\theta}}\Big) = -\frac{n}{2}\log\Big(2\pi\widehat{\sigma_1^2}\Big) - \frac{\sum_{i=1}^{n}(y_i-\widehat{\mu})^2}{2\widehat{\sigma_1^2}} - \frac{m}{2}\log\Big(2\pi\widehat{\sigma_2^2}\Big) - \frac{\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2}{2\widehat{\sigma_2^2}},$$

daí vem que

$$\begin{aligned} L\left(\widehat{\boldsymbol{\theta}}\right) &= -\frac{n+m}{2}\log 2\pi - \frac{n}{2}\log\widehat{\sigma_1^2} - \frac{m}{2}\log\widehat{\sigma_2^2} \\ &- \frac{1}{2\widehat{\sigma_1^2}}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 - \frac{1}{2\widehat{\sigma_2^2}}\sum_{i=n+1}^{m+n}(y_i-\widehat{\mu})^2 \end{aligned}$$

e finalmente

$$L\left(\widehat{\boldsymbol{\theta}}\right) = -\frac{(n+m)}{2}(\log 2\pi + 1) - \frac{n}{2}\log\widehat{\sigma_1^2} - \frac{m}{2}\log\widehat{\sigma_2^2} \quad (6.28)$$

Deve-se agora encontrar o valor da função suporte maximizada; para isto, deriva-se (6.27) em relação a cada parâmetro para se encontrar as estimativas dos parâmetros.

Derivando em relação a σ_1^2 e igualando a zero $\dfrac{\partial L\left(\boldsymbol{\theta}\right)}{\partial \sigma_1^2} = 0$, tem-se:

$$\frac{\partial\left(-\dfrac{n\log\left(2\pi\sigma_1^2\right)}{2} - \dfrac{\sum\limits_{i=1}^{n}(y_i-\mu)^2}{2\sigma_1^2} - \dfrac{m}{2}\log\left(2\pi\sigma_2^2\right) - \dfrac{\sum\limits_{i=n+1}^{n+m}(y_i-\mu)^2}{2\sigma_2^2}\right)}{\partial\sigma_1^2} = 0,$$

Desse modo

$$-\frac{n}{2\widehat{\sigma_1^2}} + \frac{1}{2\left(\widehat{\sigma_1^2}\right)^2}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 = 0 \Longrightarrow \frac{1}{2\widehat{\sigma_1^2}}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 = \frac{n}{2}$$

Assim o estimador de σ_1^2 é dado por

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 \tag{6.29}$$

Derivando (6.27) em relação a σ_2^2 e igualando-se a zero tem-se:

$$\frac{\partial\left(-\dfrac{n}{2}\log\left(2\pi\sigma_1^2\right) - \dfrac{\sum\limits_{i=1}^{n}(y_i-\mu)^2}{2\sigma_1^2} - \dfrac{m}{2}\log\left(2\pi\sigma_2^2\right) - \dfrac{\sum\limits_{i=n+1}^{n+m}(y_i-\mu)^2}{2\sigma_2^2}\right)}{\partial\sigma_2^2} = 0$$

Assim

$$-\frac{m}{2\widehat{\sigma_2^2}} + \frac{1}{2\left(\widehat{\sigma_2^2}\right)^2}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2 = 0 \Longrightarrow \frac{1}{2\widehat{\sigma_2^2}}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2 = \frac{m}{2}$$

E assim obtém-se o estimador de σ_2^2 dado por:

$$\widehat{\sigma_2^2} = \frac{1}{m}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2 \tag{6.30}$$

Fazendo-se $\dfrac{\partial L(\boldsymbol{\theta})}{\partial \mu} = 0$ em (6.27) tem-se:

$$\frac{\partial\left(-\frac{n}{2}\log\left(2\pi\sigma_1^2\right) - \dfrac{\sum\limits_{i=1}^{n}(y_i-\mu)^2}{2\sigma_1^2} - \dfrac{m}{2}\log\left(2\pi\sigma_2^2\right) - \dfrac{\sum\limits_{i=n+1}^{n+m}(y_i-\mu)^2}{2\sigma_2^2}\right)}{\partial\mu} = 0$$

Desse modo

$$-\frac{1}{2\widehat{\sigma_1^2}}2\sum_{i=1}^{n}(y_i-\widehat{\mu})(-1) - \frac{1}{2\widehat{\sigma_2^2}}2\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})(-1) = 0$$

$$\Downarrow$$

$$\frac{1}{\widehat{\sigma_1^2}}\sum_{i=1}^{n}(y_i-\widehat{\mu}) = -\frac{1}{\widehat{\sigma_2^2}}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu}) \tag{6.31}$$

Substituindo os estimadores de σ_1^2 e σ_2^2, obtidos em (6.29) e (6.30) em (6.31) tem-se:

$$\frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n}(y_i-\widehat{\mu})^2}\sum_{i=1}^{n}(y_i-\widehat{\mu}) = -\frac{1}{\frac{1}{m}\sum\limits_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2}\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})$$

$$\Downarrow$$

$$n\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2\sum_{i=1}^{n}(y_i-\widehat{\mu}) = -m\sum_{i=1}^{n}(y_i-\widehat{\mu})^2\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})$$

$$\Downarrow$$

$$n\sum_{i=1}^{n}(y_i-\widehat{\mu})\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})^2 + m\sum_{i=n+1}^{n+m}(y_i-\widehat{\mu})\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 = 0$$

$$\Downarrow$$

$$n\left[\sum_{i=1}^{n}y_i - \sum_{i=1}^{n}\widehat{\mu}\right]\sum_{i=n+1}^{n+m}\left(y_i^2 - 2\widehat{\mu}y_i + \widehat{\mu^2}\right) +$$

$$+m\left[\sum_{i=n+1}^{n+m} y_i - \sum_{i=n+1}^{n+m}\widehat{\mu}\right]\sum_{i=1}^{n}\left(y_i^2 - 2\widehat{\mu}y_i + \widehat{\mu^2}\right) = 0$$

$$\Downarrow$$

$$\underbrace{n\left[\sum_{i=1}^{n} y_i - n\widehat{\mu}\right]\left[\sum_{i=n+1}^{n+m} y_i^2 - 2\widehat{\mu}\sum_{i=n+1}^{n+m} y_i + m\widehat{\mu}^2\right]}_{(P)} +$$

$$+\underbrace{m\left[\sum_{i=n+1}^{n+m} y_i - \sum_{i=n+1}^{n+m}\widehat{\mu}\right]\left[\sum_{i=1}^{n} y_i^2 - 2\widehat{\mu}\sum_{i=1}^{n} y_i + n\widehat{\mu^2}\right]}_{(Q)} = 0$$

Desenvolvendo (P) tem-se

$$n\left[\sum_{i=1}^{n} y_i - n\widehat{\mu}\right]\left[\sum_{i=n+1}^{n+m} y_i^2 - 2\widehat{\mu}\sum_{i=n+1}^{n+m} y_i + m\widehat{\mu^2}\right] = n\left(\sum_{i=1}^{n} y_i\right)\sum_{i=n+1}^{n+m} y_i^2 -$$

$$-2\widehat{\mu}n\left(\sum_{i=1}^{n} y_i\right)\sum_{i=n+1}^{n+m} y_i + mn\widehat{\mu^2}\left(\sum_{i=1}^{n} y_i\right) - n^2\widehat{\mu}\sum_{i=n+1}^{n+m} y_i^2 + 2n^2\widehat{\mu^2}\sum_{i=n+1}^{n+m} y_i - n^2m\widehat{\mu^3}$$

Desenvolvendo (Q) tem-se

$$m\left[\sum_{i=n+1}^{n+m} y_i - m\widehat{\mu}\right]\left[\sum_{i=1}^{n} y_i^2 - 2\widehat{\mu}\sum_{i=1}^{n} y_i + n\widehat{\mu}^2\right] = m\left(\sum_{i=n+1}^{n+m} y_i\right)\sum_{i=1}^{n} y_i^2$$

$$-2m\widehat{\mu}\sum_{i=1}^{n} y_i\sum_{i=n+1}^{n+m} y_i + mn\widehat{\mu^2}\sum_{i=n+1}^{n+m} y_i - m^2\widehat{\mu}\sum_{i=1}^{n} y_i^2 + 2m^2\widehat{\mu^2}\sum_{i=1}^{n} y_i - nm^2\widehat{\mu^3}$$

Juntando-se (P) e (Q) tem-se:

$$\begin{aligned}
0=(P)+(Q) \;=\; & n\left(\sum_{i=1}^{n} y_i\right)\sum_{i=n+1}^{n+m} y_i^2 - 2\widehat{\mu}n\left(\sum_{i=1}^{n} y_i\right)\sum_{i=n+1}^{n+m} y_i + mn\widehat{\mu^2}\left(\sum_{i=1}^{n} y_i\right) \\
& - n^2\widehat{\mu}\sum_{i=n+1}^{n+m} y_i^2 + 2n^2\widehat{\mu^2}\sum_{i=n+1}^{n+m} y_i - n^2m\widehat{\mu^3} + m\left(\sum_{i=n+1}^{n+m} y_i\right)\sum_{i=1}^{n} y_i^2 \\
& - \; 2m\widehat{\mu}\sum_{i=1}^{n} y_i\sum_{i=n+1}^{n+m} y_i + mn\widehat{\mu^2}\sum_{i=n+1}^{n+m} y_i - m^2\widehat{\mu}\sum_{i=1}^{n} y_i^2 \\
& + \quad 2m^2\widehat{\mu^2}\sum_{i=1}^{n} y_i - nm^2\widehat{\mu^3}
\end{aligned}$$

Agrupando-se os termos de grau semelhante tem-se:

$$-\left(nm^2+n^2m\right)\widehat{\mu^3}+\left(2m^2\sum_{i=1}^{n}y_i+mn\sum_{i=n+1}^{n+m}y_i+2n^2\sum_{i=n+1}^{n+m}y_i+mn\sum_{i=1}^{n}y_i\right)\widehat{\mu^2}+$$

$$+\widehat{\mu}\left(-m^2\sum_{i=1}^{n}y_i^2-2m\sum_{i=1}^{n}y_i\sum_{i=n+1}^{n+m}y_i-2n\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i-n^2\sum_{i=n+1}^{n+m}y_i^2\right)+$$

$$+n\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i^2+m\left(\sum_{i=n+1}^{n+m}y_i\right)\sum_{i=1}^{n}y_i^2=0$$

Dividindo-se por $\left(-nm^2-n^2m\right)$ tem-se:

$$\widehat{\mu^3}-\widehat{\mu^2}\frac{\left(2m^2\sum_{i=1}^{n}y_i+mn\sum_{i=n+1}^{n+m}y_i+2n^2\sum_{i=n+1}^{n+m}y_i+mn\sum_{i=1}^{n}y_i\right)}{nm\left(m+n\right)}$$

$$+\ \widehat{\mu}\frac{\left(-m^2\sum_{i=1}^{n}y_i^2-2m\sum_{i=1}^{n}y_i\sum_{i=n+1}^{n+m}y_i-2n\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i-n^2\sum_{i=n+1}^{n+m}y_i^2\right)}{nm\left(m+n\right)}$$

$$-\ \frac{\left(n\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i^2+m\left(\sum_{i=n+1}^{n+m}y_i\right)\sum_{i=1}^{n}y_i^2\right)}{nm\left(m+n\right)}=0$$

Daí segue que

$$\widehat{\mu^3}+\widehat{\mu^2}\left(-\frac{2m\sum_{i=1}^{n}y_i}{n\left(m+n\right)}-\frac{\sum_{i=n+1}^{n+m}y_i}{\left(m+n\right)}-\frac{2n\sum_{i=n+1}^{n+m}y_i}{m\left(m+n\right)}-\frac{\sum_{i=1}^{n}y_i}{\left(m+n\right)}\right)$$

$$+\ \widehat{\mu}\left(\frac{m\sum_{i=1}^{n}y_i^2}{n\left(m+n\right)}+\frac{2\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i}{n\left(m+n\right)}+\frac{2\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i}{m\left(m+n\right)}+\frac{n\sum_{i=n+1}^{n+m}y_i^2}{m\left(m+n\right)}\right)$$

$$-\ \frac{1}{nm\left(m+n\right)}\left(n\left(\sum_{i=1}^{n}y_i\right)\sum_{i=n+1}^{n+m}y_i^2+m\left(\sum_{i=n+1}^{n+m}y_i\right)\sum_{i=1}^{n}y_i^2\right)=0 \quad (6.32)$$

Sejam

$$w=\frac{n}{m+n} \quad , \quad v=\frac{m}{m+n},$$
$$\mu_1=\frac{\sum_{i=1}^{n} y_i}{n}, \mu_2=\frac{\sum_{i=n+1}^{n+m} y_i}{m} \quad , \quad s_1^2=\frac{\sum_{i=1}^{n} y_i^2}{n}, s_2^2=\frac{\sum_{i=n+1}^{n+m} y_i^2}{m}. \tag{6.33}$$

Substituindo (6.33) em (6.32), tem-se:

$$\widehat{\mu^3}+\widehat{\mu^2}(-2v\mu_1 - v\mu_2 - 2w\mu_2 - w\mu_1) + \widehat{\mu}\left(\frac{v}{n}\sum_{i=1}^{n} y_i^2 + 2v\mu_1\mu_2 + 2w\mu_1\mu_2 + \frac{w}{m}\sum_{i=n+1}^{n+m} y_i^2\right) - \left(\frac{w}{m}\mu_1\sum_{i=n+1}^{n+m} y_i^2 + \frac{v}{n}\mu_2\sum_{i=1}^{n} y_i^2\right) = 0$$

Efetuando-se as operações necessárias tem-se:

$$\mu^3 + \mu^2\left(-\mu_1(2v+w) - (v+2w)\mu_2\right) + \mu\left(vs_1^2 + 2v\mu_1\mu_2 + 2w\mu_1\mu_2 + ws_2^2\right) - \left(\mu_1 ws_2^2 + v\mu_2 s_1^2\right) = 0$$

$$\Downarrow$$

$$\widehat{\mu^3} + \widehat{\mu^2}\left(-\mu_1\left(\frac{m+m+n}{m+n}\right) - \left(\frac{m+n+n}{m+n}\right)\mu_2\right) + \widehat{\mu}\left(vs_1^2 + 2v\mu_1\mu_2 + 2w\mu_1\mu_2 + ws_2^2\right) - \left(\mu_1 ws_2^2 + v\mu_2 s_1^2\right) = 0$$

$$\Downarrow$$

$$\widehat{\mu^3} + \widehat{\mu^2}\left(-\mu_1\left(1+\frac{m}{m+n}\right) - \left(1+\frac{n}{m+n}\right)\mu_2\right) + \widehat{\mu}\left(2\mu_1\mu_2\left(\frac{m}{m+n}+\frac{n}{m+n}\right) + vs_1^2 + ws_2^2\right) - \left(\mu_1 ws_2^2 + v\mu_2 s_1^2\right) = 0$$

$$\Downarrow$$

$$\widehat{\mu^3} + \widehat{\mu^2}\left(-\mu_1(1+v) - (1+w)\mu_2\right) + \widehat{\mu}\left(2\mu_1\mu_2 + vs_1^2 + ws_2^2\right) - \left(\mu_1 ws_2^2 + v\mu_2 s_1^2\right) = 0$$

Fazendo

$$A = -\left(\mu_1\left(1+v\right)+\left(1+w\right)\mu_2\right) \tag{6.34}$$

$$B = \left(2\mu_1\mu_2 + vs_1^2 + ws_2^2\right) \tag{6.35}$$

$$C = -\left(\mu_1 w s_2^2 + v\mu_2 s_1^2\right) \tag{6.36}$$

Tem-se $\widehat{\mu^3} + A\widehat{\mu^2} + B\widehat{\mu} + C = 0$ que é uma equação do terceiro grau cuja solução pode ser obtida através da fórmula de Cardano (Garbi, 1997) dada a seguir.

A fórmula de Cardano

Toda equação cúbica

$$ax^3 + bx^2 + cx + d = 0$$

com $a \neq 0$ pode ser reduzida à forma

$$y^3 + py + q = 0$$

em que $x = y - \dfrac{b}{3a}$ $p = \dfrac{\left(3ac - b^2\right)}{3a^2}$ e $q = \dfrac{1}{27a^3}\left(\left(3-a\right)b^3 - 9abc + 27a^2d\right)$, sendo que suas soluções são dadas por

$$y = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}.$$