Universidade de Évora

Mestrado em Matemática Aplicada

Ajustamento de distribuições a dados de captura de pescado

Dissertação apresentada para obter o grau de Mestre em Matemática Aplicada pela Universidade de Évora

Carlos Ferreira do Carmo de Sousa

ÉVORA, 2000

Universidade de Évora

Mestrado em Matemática Aplicada

Ajustamento de distribuições a dados de captura de pescado

Dissertação apresentada para obter o grau de Mestre em Matemática Aplicada pela Universidade de Évora

Carlos Ferreira do Carmo de Sousa

ÉVORA, 2000

Universidade de Évora

Mestrado em Matemática Aplicada

Ajustamento de distribuições a dados de captura de pescado

Dissertação apresentada para obter o grau de Mestre em Matemática Aplicada pela Universidade de Évora

Carlos Ferreira do Carmo de Sousa

ÉVORA, 2000

Resumo

As distribuições de comprimentos à idade referentes à captura de pescado, apresentam variações de assimetria, que podem ser devidas a predação, mortalidade natural acrescida para indivíduos mais pequenos ou outros factores. Não obstante esta variação de assimetria, em biologia das pescas assume-se que a distribuição normal descreve o comportamento deste tipo de dados. Pretende-se com este trabalho, usando dados do "National Marine Fisheries Service" (Woods Hole, Estados Unidos) sobre a pescada (*Merluccius bilinearis*), ajustar outras distribuições, que permitam variações de assimetria, alternativas à distribuição normal, que possam constituir um modelo estatístico mais adequado. Para isso os parâmetros das diferentes distribuições ensaiadas (normal, lognormal, gamma, Burr e misturas) são estimados por máxima verosimilhança (ou, quando tal não seja viável, pelo método dos momentos), comparados os valores da função de log-verosimilhança e do critério de informação de Akaike e aplicados testes de qualidade de ajuste. Os resultados obtidos analiticamente serão comparados com os gráficos dos referidos ajustamentos.

Com o objectivo de se validar um algoritmo de estimação (por máxima verosimilhança) viável para as diversas distribuições ensaiadas, fez-se primeiro, usando os dados referidos e tendo como referência a distribuição normal, um estudo comparativo de diferentes algoritmos iterativos (métodos de Newton, algoritmo EM e algoritmo EM com simulação). A aplicação do algoritmo EM tem a vantagem de se poderem considerar os efeitos do agrupamento dos dados e de simplificar o processo de obtenção das estimativas de máxima verosimilhança para a mistura de distribuições.

Todos os modelos alternativos dão melhores resultados de ajustamento do que a distribuição normal.

Abstract

Fish length-at-age distributions are characterised by varying degrees of asymmetry that may be due to predation, increased mortality in smaller sized individuals and other size dependent factors. Despite the variation in asymmetry, the normal distribution is usually used to describe these types of fisheries biology data. The objective of this study was to fit other distributions, which are alternatives to the normal distribution to data from the National Marine Fisheries Service (Woods Hole, U.S.A.) for the hake (*Merluccius bilinearis*). These models, that permit variation in asymmetry, may give better fits and prove more adequate from a statistical point of view. The parameters of the different models (log-normal, gamma, Burr and mixture) tested were estimated by maximum likelihood or alternatively by the method of moments. Values of maximum likelihood as well as the information criterion of Akaike were compared and goodness of fit tests carried out. The results obtained analytically were compared with the fits.

In order to validate a viable estimation algorithm (by maximum likelihood) for the different distributions tested, a preliminary comparative study of different iterative algorithms (Newton method, EM algorithm and EM algorithm with simulation) were carried out with the above mentioned data and with the normal distribution as a baseline reference. The EM algorithm has the advantage of allowing the consideration of grouped data and the simplification of the process of maximum likelihood parameter estimation for mixtures of distributions. All the models tested gave better fits to the data than the normal.

Agradecimentos

À minha família, a eles dedico este trabalho.
Ao Professor Doutor Carlos Braumann.
À Professora Doutora Margarida Castro.
Ao Professor Doutor Karim Erzini.
Ao José Mendinhos.
Ao João Leitão.

Índice

1. Introdução

Em estatística aplicada, nomeadamente na investigação biológica e ecológica, é muitas vezes necessário o ajustamento de modelos ou distribuições teóricas a um conjunto de dados. Este ajustamento tem como principal objectivo a interpretação dos dados empíricos. Assim, considerando um conjunto de modelos que pareçam biologicamente razoáveis, o problema central da análise de dados será a selecção de um modelo apropriado para um conjunto de dados como base para a inferência.

Em biologia das pescas surge, muitas vezes, este problema prático quando se pretende estudar a evolução da distribuição do comprimento de uma espécie de peixes ao longo do tempo. O objectivo deste trabalho foi o desenvolvimento de uma metodologia que permita estudar este problema.

Para as espécies de peixes, a distribuição do comprimento à idade pode apresentar assimetria negativa ou positiva, devido a fenómenos de mortalidade que afectam, de forma diferente, os extremos da distribuição. Assimetrias negativas podem surgir quando a predação afecta sobretudo os indivíduos de maiores dimensões de um determinado grupo etário. A assimetria positiva resulta de situações em que a mortalidade natural é superior em indivíduos mais pequenos. A evolução da assimetria ao longo do tempo permite obter indicação da evolução dos mecanismos de mortalidade, natural e devida à predação, que é de grande interesse ecológico. O conhecimento de como actuam os factores de mortalidade numa população pode ser utilizado em modelos de previsão da evolução da estrutura de comprimentos, sendo relevante para o planeamento da actividade pesqueira, através de legislação relacionada com o efeito das artes de pesca (por exemplo, tamanhos de malha da rede e/ou tamanhos mínimos de captura). A utilização de um modelo que permita variações de assimetria, permitirá a caracterização da distribuição do comprimento à idade em termos de localização, dispersão e assimetria, contrariamente à utilização do modelo normal que não permite variações de assimetria.

Várias distribuições e situações amostrais podem ser consideradas para representar a forma da população subjacente amostrada. Raramente se pode esperar descobrir o verdadeiro modelo; em vez disto, o objectivo é seleccionar o modelo biologicamente significativo que seja completamente suportado pelo conjunto de dados específico, tendo-se em conta as limitações da amostragem. Uma selecção conveniente do modelo rejeita o modelo que esteja longe da realidade e tende a identificar um modelo no qual o erro de

aproximação e o erro devido a flutuações aleatórias estejam bem balanceados (Shibata 1989).

Para ilustrar esta problemática foram utilizados dados do "National Marine Fisheries Service", Woods Hole, Estados Unidos. Foram seleccionadas 4 matrizes de dados, agrupados em classes, que representam a distribuição de comprimentos dos peixes (a pescada, *Merluccius bilinearis*), para uma série de idades presentes na população em determinado momento. Na generalidade, aceita-se, por simplificação, que a distribuição normal descreve adequadamente o comportamento deste tipo de dados; mas, por razões já mencionadas, este conjunto de dados pode apresentar observações extremas com uma ocorrência muitas vezes superior à esperada numa população normal. Em particular, neste estudo 83% das amostras apresentam assimetria positiva, ou seja, é de esperar que um modelo que permita variações na assimetria se ajuste melhor a este conjunto de dados do que a distribuição normal.

Considerar-se-ão algumas distribuições alternativas que variam na forma e no peso das caudas, desde caudas neutras até caudas alongadas, isto é, que permitam variações na assimetria. Espera-se que estas distribuições cubram a maioria das situações razoáveis. Foram estudadas, para além da distribuição normal, a distribuição lognormal (por o processo de crescimento dos peixes se tratar de um fenómeno de carácter multiplicativo a nível celular), a distribuição gama (pela sua flexibilidade e por ser distribuição assintótica de certos modelos de crescimento em ambientes aleatórios) e as distribuições de Burr (Burr, 1954), que foram sugeridas por investigadores da área de biologia e pescas (apesar de nunca terem sido estudadas neste campo e não se conhecerem referências, são consideradas úteis por permitirem assimetria tanto negativa como positiva). Estudaram-se também misturas de distribuições normais e lognormais (por poder haver mistura de subpopulações e estarem bem descritas na literatura da especialidade, sendo utilizadas como modelos biológicos).

A estimação dos parâmetros das diferentes distribuições teóricas é feita utilizando, quando possível, o método da máxima verosimilhança (por este apresentar óptimas propriedades) e métodos iterativos para o cálculo das estimativas de máxima verosimilhança (métodos de Newton, algoritmo EM ou de estimação-maximização e algoritmo EM com simulação). A aplicação do algoritmo EM tem a vantagem de permitir considerar os efeitos do agrupamento dos dados e de simplificar o processo de obtenção das estimativas de máxima verosimilhança para a mistura de distribuições. O método dos

momentos é utilizado para se obterem valores iniciais para a aplicação dos algoritmos iterativos e sempre que a utilização de métodos de máxima verosimilhança seja impossível.

Comparámos, para cada parâmetro, as estimativas de máxima verosimilhança obtidas pelos diferentes métodos usando como referência a distribuição normal, para a qual é possível aplicar com relativa facilidade todos os métodos considerados neste trabalho. Verificámos que o algoritmo EMMC (algoritmo EM com simulação de Monte Carlo) com $M = 30$ valores simulados por iteração e com utilização do método de Newton-Raphson produzia resultados bastante bons. Como esse método, ao contrário de alguns outros, é de fácil aplicação para todas as distribuições em estudo, foi o utilizado para obtenção das estimativas dos parâmetros dos vários modelos considerados.

A selecção dos modelos é baseada em testes de qualidade de ajuste e, paralelamente, através da comparação dos valores obtidos para a função de verosimilhança e para o *critério de informação de Akaike* (*AIC*) (Akaike 1973, 1985), que se baseia na verosimilhança. Este último reduz a selecção do modelo a um problema de optimização unidimensional dando ênfase ao princípio da parcimónia, parecendo oferecer bons resultados na prática, e é simples de calcular e interpretar.

A inferência estatística a partir de um conjunto de dados, conhecido o modelo, está bem desenvolvida e suportada pela teoria, sendo utilizada, por exemplo, na resolução de problemas biológicos. A questão mais pertinente, será perguntar de onde vem o modelo, o que o justifica, e como a inferência é afectada devido à incerteza do conhecimento deste?

2. Distribuições de probabilidade

2.1 Momentos. Função geradora de momentos

As distribuições associadas às populações, que se supõem absolutamente contínuas e definidas pela função densidade de probabilidade (f.d.p.) ou pela função distribuição (f.d.), podem ser caracterizadas por um conjunto de parâmetros. Assim, quando se fala de «parâmetro desconhecido θ», este símbolo pode representar um vector com uma ou mais componentes, isto é $\theta = (\theta_i)$ com $i = 1,\ldots,p \in IN$. O conjunto de valores possíveis de θ designa-se por espaço de parâmetros. De um modo geral, o espaço de parâmetros de um modelo aleatório com p parâmetros desconhecidos é um subconjunto de IR^p e será representado por Ω. Se Y é uma variável aleatória (v.a.) com f.d.p. $f(y)$ que envolve o parâmetro θ, então, para se pôr esse facto em evidência, escreve-se $f(y \mid \theta)$. Consideram-se sempre v.a.'s absolutamente contínuas, isto é, com f.d.p..

Para uma dada população, os parâmetros das distribuições de probabilidade são fixos, em contraste com as estatísticas, que variam de amostra para amostra (são função da amostra). Contudo, a designação do parâmetro é, em geral, idêntica à da estatística, podendo a distinção entre eles ser estabelecida, quando necessário, apelidando o parâmetro de «populacional» e a estatística de «amostral».

Apesar da f.d. (ou da f.d.p.) caracterizar completamente uma v.a., uma descrição ou caracterização desta pode obter-se, ainda, através de certos parâmetros característicos. Os parâmetros média μ e desvio-padrão σ, constituem medidas numéricas descritivas relevantes localizando o centro e descrevendo a dispersão de $f(y \mid \theta)$, respectivamente. Todavia, é possível que distribuições diferentes possuam médias e desvios-padrão comuns, o que significa que μ e σ, só por si, não caracterizam completamente, com unicidade, a distribuição. Este problema de unicidade leva a procurar um conjunto mais alargado de medidas numéricas descritivas que determine unicamente a f.d.p., sob certas condições gerais.

Por outro lado, deve observar-se que, uma vez que cada distribuição determina um conjunto de momentos (caso estes existam), é condição necessária, para que duas distribuições sejam iguais, que tenham a mesma sequência de momentos populacionais.

Existem dois tipos de momentos populacionais, os momentos ordinários e os momentos centrados. No entanto, em geral, esta condição não é suficiente, pois uma sequência de momentos não determina univocamente uma distribuição. Para que tal suceda, há que garantir a existência de uma função que se designa por função geradora de momentos.

Definição 2.1 - Seja Y uma v.a. qualquer. Se existir um número positivo h tal que se possa definir a função

$$M(t) = E\left[e^{tY}\right] \tag{2.1}$$

para qualquer $t \in \left]-h, h\right[$, tal função designa-se por função geradora de momentos (f.g.m) da variável Y.

Pode demonstrar-se que, se esta função existir, então é contínua e diferenciável em torno de $t = 0$.

Em geral, os momentos ordinários de ordem i $(i = 1,2\ldots)$ são dados por

$$M^{(i)}(0) = E\left[Y^i\right] = \mu_i' . \tag{2.2}$$

Note-se que, caso Y tenha f.d.p. $f(y \mid \theta)$, tem-se

$$\mu_i' = \int_D y^i f(y \mid \theta)\, dy , \tag{2.3}$$

sendo D o suporte da v.a. Y. A partir desta expressão e da definição de valor esperado, conclui-se que este parâmetro corresponde ao momento populacional ordinário de primeira ordem, ou seja, $\mu_1' = \mu = E[Y]$.

Apesar de nem toda a distribuição ter f.g.m., a importância desta, quando existe, deve-se ao facto de ser única e determinar completamente a distribuição da v.a.. A função $M(t)$ deve a sua designação à propriedade (2.2), pois, a partir dela, é possível gerar todos os momentos ordinários de uma qualquer v.a. Y e, com base nestes, calcular os momentos

centrados. De facto, como é sabido, qualquer momento centrado de ordem i pode exprimir-se em função dos momentos ordinários de ordem não superior a i (ver por exemplo Stuart e Ord, 1987). Para distribuições com f.d.p. $f(y \mid \theta)$, o momento centrado de ordem i $(i = 1,2,\ldots)$ é

$$\mu_i = \int_D (y-\mu)^i f(y \mid \theta)\, dy, \tag{2.4}$$

sendo D o suporte da v.a. Y. A partir desta definição resulta que o momento centrado de primeira ordem é sempre nulo, $\mu_1 = 0$ e que o momento centrado de segunda ordem é a variância populacional, $\mu_2 = \sigma^2 = V[Y]$.

Na prática, raramente se calculam momentos para ordens superiores a 4, pois, por um lado, tais momentos são difíceis de obter e de interpretar e, por outro, a igualdade dos momento de ordem não superior a quatro é, normalmente, uma condição para que duas distribuições sejam aproximadamente iguais.

Os momentos centrados de terceira e quarta ordem servem para medir, respectivamente, a assimetria e o achatamento (curtose) das distribuições populacionais. A partir destes momentos, podem definir-se os coeficientes populacionais de assimetria e achatamento, que podem ser interpretados como parâmetros de forma.

Define-se o coeficiente de assimetria por

$$\gamma_1 = \frac{\mu_3}{\sigma^3}. \tag{2.5}$$

Se $\gamma_1 > 0$, a distribuição diz-se assimétrica à direita e $\mu > me$ (onde me representa a mediana). Se $\gamma_1 < 0$, a distribuição diz-se assimétrica à esquerda e $me > \mu$. Se a distribuição for simétrica, $\gamma_1 = 0$.

Define-se o coeficiente de achatamento por

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{2.6}$$

As distribuições com $\gamma_2 = 0$ (caso da distribuição normal) são designadas por mesocúrticas, com $\gamma_2 > 0$ leptocúrticas (a curva é menos achatada) e com $\gamma_2 < 0$ platicúrticas (a curva é mais achatada).

2.2 Distribuições teóricas utilizadas

Na maior parte dos casos é assumido, por simplificação, que a distribuição do comprimento à idade de dados referentes à captura de pescado, segue uma distribuição normal. Os desvios, relativamente à normalidade, não são, pois, considerados significativos. No entanto, a generalidade das distribuições amostrais apresenta desvios a este modelo, em geral apresentando assimetrias positivas que são significativas do ponto de vista biológico, sobretudo se se alterarem ao longo do crescimento. A explicação destas assimetrias envolve fenómenos de mortalidade diferenciada para diferentes tamanhos da mesma classe etária e tem importantes consequências do ponto de vista biológico.

O objectivo principal deste estudo é, pois, o ajustamento de distribuições teóricas alternativas à distribuição normal que permitam variações de assimetria (e curtose) na distribuição do comprimento à idade de dados referentes à captura de pescado.

Um modelo "adequado" (a distribuição que melhor se ajusta aos dados) deve ter estrutura e parâmetros suficientes para ter em conta, adequadamente, a variabilidade significativa dos dados. O número de parâmetros num modelo corresponde à quantidade e ao tipo de estrutura que o modelo tem para descrever (i.e. ajustar) os dados. Pode-se ordenar os modelos por graus de estrutura em termos do número de parâmetros destes, existindo um número máximo de parâmetros que pode ser "suportado" por qualquer conjunto de dados. Este suporte é uma função, quer do verdadeiro modelo quer do tamanho da amostra. Assim, deve-se considerar alternativas entre dois indesejáveis extremos: o subajustamento se o modelo tiver poucos parâmetros ou uma estrutura fraca, caso em que o resultado pode ser o enviesamento dos estimadores; o sobreajustamento para um modelo com muitos parâmetros ou uma estrutura forte, caso em que se perde precisão dos estimadores. O princípio da parcimónia (Goodman 1984:34-36, McCullangh e Nelder 1989:6) providencia um "acordo" entre estes dois extremos

Foram consideradas as seguintes distribuições biologicamente significativas para este conjunto de dados:

(1) distribuição normal, por ser a distribuição habitualmente utilizada para descrever a distribuição do comprimento à idade em biologia das pescas e devido ao objectivo principal deste estudo ser a comparação dos resultados obtidos para esta distribuição com as distribuições alternativas que se seguem;
(2) distribuição lognormal, por o processo de crescimento dos peixes se tratar de um fenómeno de carácter multiplicativo a nível celular;
(3) distribuição gama, pela sua flexibilidade e por ser distribuição assintótica de certos modelos de crescimento em ambientes aleatórios;
(4) distribuições de Burr (denominadas por Burr1 e Burr2), por nunca terem sido estudas neste campo (não se conhecem quaisquer referências na literatura da especialidade) foram sugeridas por investigadores da área de biologia e pescas, uma vez que permitem assimetria tanto negativa como positiva;
(5) mistura de distribuições normais e lognormais, por poder haver mistura de subpopulações e por estarem bem descritas na literatura da especialidade, sendo utilizadas como modelos biológicos.

2.2.1 Distribuição normal

A distribuição normal é uma das mais importantes, senão a mais importante, distribuição contínua. Em virtude do teorema do limite central, muitas v.a.'s, nomeadamente aquelas que dizem respeito a fenómenos físicos, obedecem à lei de probabilidade normal; muitas outras têm distribuição que, se não é normal, aproxima-se muito desta, nomeadamente as que dizem respeito a fenómenos biométricos.

Uma v.a. Y tem distribuição normal com parâmetros μ e σ^2, $Y \sim N(\mu, \sigma^2)$, quando a sua f.d.p. for da forma,

$$f(y \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma} \right)^2 \right), y \in IR, \qquad (2.7)$$

onde $\theta = (\mu, \sigma)$, com $\mu \in IR$ o parâmetro de localização e $\sigma > 0$ o parâmetro de escala.

Fazendo $Z = \frac{Y - \mu}{\sigma}$, diz-se que a variável Z tem distribuição normal reduzida, $Z \sim N(0,1)$. A f.d.p. de Z é

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \; z \in IR, \tag{2.8}$$

e a correspondente f.d. será dada por

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt, \tag{2.9}$$

designada por função de Laplace, e representa-se, convencionalmente, por $\Phi(z)$. Não há uma expressão explícita para esta função; os seus valores encontram-se, por isso, tabelados.

Por analogia à normal reduzida, ter-se-á para f.d. de uma distribuição $N(\mu, \sigma^2)$

$$F(y \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt = \Phi\left(\frac{y-\mu}{\sigma}\right). \tag{2.10}$$

Conhecidos os parâmetros μ e σ, o cálculo das probabilidades reduz-se à manipulação da tabela de $\Phi(z)$, através do resultado anterior.

Considerando (2.7), a f.g.m. será dada por

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \tag{2.11}$$

vindo

$$E[Y] = \mu = me = mo, \; V[Y] = \sigma^2, \; \gamma_1 = 0 \text{ e } \gamma_2 = 0$$

Esta distribuição é, pois, simétrica em relação ao valor médio respectivo, μ, e o seu coeficiente de assimetria será uma referência para classificar outras distribuições, falando-se de:

(i) assimetria à direita ou assimetria positiva se $\gamma_1 > 0$;

(ii) assimetria à esquerda ou assimetria negativa se $\gamma_1 < 0$.

Assim, também o coeficiente de achatamento, será uma referência no que concerne ao achatamento das distribuições:

(i) $\gamma_2 > 0$ significará que as distribuições têm caudas menos longas ou mais pesadas que a normal, menos achatadas (distribuições leptocúrtícas);

(ii) $\gamma_2 < 0$ significará que as distribuições têm caudas mais longas ou menos pesadas que a normal, mais achatadas (distribuições platicúrticas).

2.2.2 Distribuição lognormal

Paralelamente às situações da distribuição normal, é útil, por vezes, considerar a distribuição de um fenómeno que se revela como o resultado de um mecanismo multiplicativo actuando sobre um número de factores. Envolvendo o crescimento de um organismo a multiplicação celular, poderá ser um modelo adequado à descrição de variáveis como o comprimento dos peixes.

Se $Z = \ln Y$ tiver distribuição normal $N(\lambda, \delta^2)$, então a distribuição de Y diz-se lognormal, $Y \sim LN(\lambda, \delta^2)$, e a sua f.d.p. é dada por

$$f(y \mid \boldsymbol{\theta}) = \frac{1}{\delta y \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln y - \lambda}{\delta} \right)^2 \right), \qquad (2.12)$$

com $\boldsymbol{\theta} = (\lambda, \delta^2)$, $y > 0, \delta > 0$ e $\lambda \in IR$. Repare-se que, se $Y \sim LN(\lambda, \delta^2)$, então $Z = \ln Y \sim N(\lambda, \delta^2)$. A f.d. é

$$F(y \mid \theta) = \int_{-\infty}^{y} \frac{1}{\delta t \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln t - \lambda}{\delta} \right)^2 \right) dt = \Phi\left(\frac{\ln y - \lambda}{\delta} \right). \qquad (2.13)$$

Esta é a forma com dois parâmetros da distribuição lognormal, com λ como parâmetro de localização e δ parâmetro de escala.

Neste caso, pode-se, para o cálculo dos momentos, fazer a mudança de variável $z = \frac{\ln y - \lambda}{\delta}$ nos integrais que os definem, vindo

$$E[Y^i] = \left(e^{\lambda + \frac{i\delta^2}{2}} \right)^i \qquad (2.14)$$

para a expressão dos momentos ordinários de ordem i.

Ter-se-á, com $a = \exp(\delta^2)$,

$E[Y] = \sqrt{a}e^{\lambda}$, $V[Y] = a(a-1)e^{2\lambda}$, $\gamma_1 = (a-1)^{\frac{1}{2}}(a+2) > 0$ e $\gamma_2 = a^4 + 2a^3 + 3a^2 - 6$.

(Repare-se que γ_1 e γ_2 não dependem de λ).

Heyde (Gumbel, 1962) mostrou que a sequência destes momentos não pertence apenas ás distribuições lognormais, i.e., a distribuição não pode ser definida pelos seus momentos.

Através de γ_1 verifica-se que todos os membros da família lognormal apresentam assimetria positiva, isto é, $E[Y] > me > mo$, sendo o grau de assimetria tanto maior quanto maior for δ. Para valores baixos de δ, a distribuição lognormal aproxima-se da distribuição normal.

A moda e mediana são, respectivamente, $mo = e^{\lambda - \delta^2}$ e $me = e^{\lambda}$.

No domínio das aplicações deste modelo estão, por exemplo, a distribuição de partículas em agregados naturais, estudos de tempos de vida, etc.. Características como o

peso, altura, densidade, são melhor representadas pela lognormal do que pela normal, dado que tais quantidades são sempre não-negativas.

2.2.3 Distribuição gama

Algumas v.a. são sempre não-negativas produzindo distribuições de dados assimétricos à direita, ou seja, a maior parte da área limitada pela função densidade localiza-se nas proximidades da origem, à direita, com a função densidade a decrescer gradualmente, conforme x cresce.

A v.a. Y tem distribuição gama, $Y \sim G(\lambda, \delta, \rho)$, se a sua f.d.p. for

$$f(y \mid \boldsymbol{\theta}) = \frac{1}{\delta \Gamma(\rho)} \left(\frac{y-\lambda}{\delta} \right)^{\rho-1} e^{-\frac{y-\lambda}{\delta}}, \tag{2.15}$$

sendo $\boldsymbol{\theta} = (\lambda, \delta, \rho)$, com $\delta, \rho > 0$ e $y > \lambda$. É uma distribuição do tipo III do sistema de Pearson. A função $\Gamma(z)$ é conhecida como a função gama, definida pelo integral de Euler

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt. \tag{2.16}$$

Para z inteiro positivo, vem $\Gamma(z) = (z-1)!$.

λ e δ são, respectivamente, os parâmetros de localização e escala.

A f.g.m. é

$$M(t) = \frac{e^{\lambda t}}{(1-\delta t)^{\rho}}, \tag{2.17}$$

donde

$$E[Y] = \lambda + \delta\rho, \; V[Y] = \delta^2 \rho, \; \gamma_1 = \frac{2}{\sqrt{\rho}} > 0 \text{ e } \gamma_2 = \frac{6}{\rho} > 0.$$

Repare-se que os coeficientes de assimetria e achatamento não dependem quer de δ quer de λ (ρ é considerado como o parâmetro de forma). A assimetria é sempre positiva e o gráfico desta distribuição é sempre menos achatado que o da normal.

A moda é dada por $mo = \lambda + \delta(\rho - 1)$.

Existe uma número extremamente variado de formas desta família. Todavia, todas são assimétricas à direita e mais alongadas que a normal. À medida que ρ aumenta, a forma do gráfico desta f.d.p. torna-se similar ao gráfico da f.d.p. da normal.

2.2.4 Distribuição Burr1

Uma v.a. Y, segue uma distribuição Burr1, $Y \sim B1(\lambda, \delta, \rho)$, quando a sua f.d.p. é

$$f(y \mid \boldsymbol{\theta}) = \frac{1}{\delta} \frac{\rho e^{-\frac{y-\lambda}{\delta}}}{\left(1 + e^{-\frac{y-\lambda}{\delta}}\right)^{\rho+1}}, \quad -\infty < y < +\infty \tag{2.18}$$

onde $\boldsymbol{\theta} = (\lambda, \delta, \rho)$, sendo $\lambda \in IR$ o parâmetro de localização e $\delta, \rho > 0$, com δ o parâmetro de escala.

A respectiva f.d. será

$$F(y \mid \boldsymbol{\theta}) = \left(e^{-\frac{y-\lambda}{\delta}} + 1\right)^{-\rho}. \tag{2.19}$$

Fazendo, $\frac{1}{1 + e^{-\frac{y-\lambda}{\delta}}} = z$ em (2.18) vê-se que a f.g.m. é

$$M(t) = \rho e^{\lambda t} B(1 - \delta t, \rho + \delta t), \tag{2.20}$$

onde $B(x, y) = \int_0^1 z^{x-1}(1-z)^{y-1} dz$ é designada por função beta (integral de Euler de 1ª espécie) e

$$E[Y^i] = \rho \int_0^{+\infty} \frac{(\lambda - \delta \ln u)^i}{(1+u)^{\rho+1}} du . \tag{2.21}$$

Logo

$$E[Y] = \lambda + \delta(C + \Psi(\rho)),\ V[Y] = \delta^2(\Psi'(1) + \Psi'(\rho)),\ \gamma_1 = \frac{\Psi''(\rho) - \Psi''(1)}{\sqrt{(\Psi'(1) + \Psi'(\rho))^3}} \text{ e}$$

$$\gamma_2 = \frac{5{,}4\Psi'^2(1) + \Psi'(\rho)\left[\pi^2 + 3\Psi'(\rho)\right] + \Psi'''(\rho)}{(\Psi'(1) + \Psi'(\rho))^2} - 3 .$$

Sendo $\Psi(z)$ a função digama (definida na secção 3.3.3) tem-se

$$\Psi(1) = -C \approx 0{,}5772 , \Psi'(1) = \frac{\pi^2}{6} \text{ e } \Psi''(1) = -2Z(3) \approx -2{,}4041$$

com

$$Z(s) = \sum_{k=1}^{+\infty} k^{-s} \text{ a função Zeta de Riemann.}$$

A mediana é dada por $me = \lambda + \delta \ln \rho$, enquanto a moda é $mo = \lambda - \delta \ln\left(2^{\frac{1}{\rho}} - 1\right)$.

Um caso particular desta distribuição é quando $\rho = 1$, vindo

$$F(y \mid \boldsymbol{\theta}) = \left(e^{-\frac{y-\lambda}{\delta}} + 1\right)^{-1} \tag{2.22}$$

e

$$f(y \mid \boldsymbol{\theta}) = \frac{1}{\delta} \frac{e^{-\frac{y-\lambda}{\delta}}}{\left(1 + e^{-\frac{y-\lambda}{\delta}}\right)^2} , \quad -\infty < x < +\infty , \tag{2.23}$$

vulgarmente conhecida como distribuição logística.

Facilmente se obtém $E[Y] = me = mo = \lambda$, $V[Y] = 2\delta^2\Psi'(1)$, $\gamma_1 = 0$ e $\gamma_2 = 1{,}2 < 3$. Esta distribuição é simétrica e pode-se concluir que as suas caudas são mais longas ou menos pesadas do que no caso da distribuição normal. Serve para modelar dados que não

revelem características assimétricas (ver figura3.1 para $\rho = 1$). Tem sido utilizada para representar funções de crescimento (com y a representar o tempo) (Johnson 1969).

O aspecto gráfico da f.d.p. é o seguinte

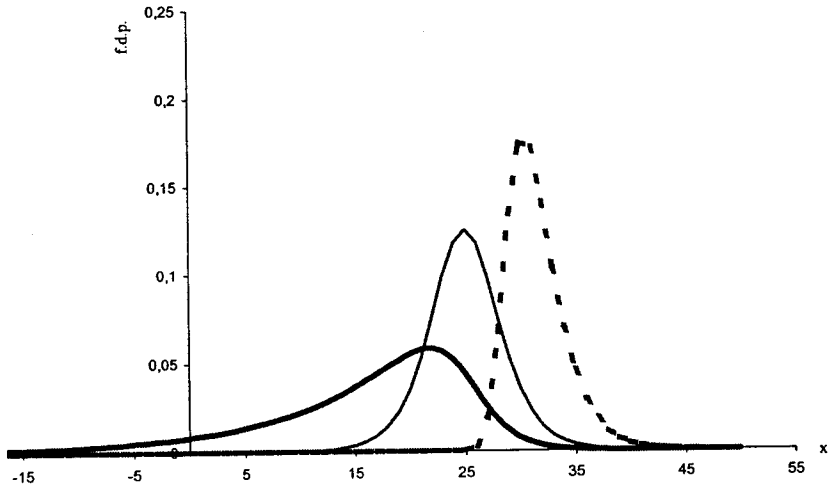

Figura2.1 – f.d.p. Burr1 para λ=25 e δ=2. As curvas correspondem, respectivamente: negrito ρ=0,2; simples ρ=1 e tracejado ρ=15.

À medida que ρ vai aumentando, a assimetria da f.d.p. vai de negativa a positiva passando por um estado de simetria ($\rho = 1$), e a curtose vai diminuindo, isto é, a forma das curvas vai-se tornando mais achatada. Assim, ρ pode ser interpretado como um parâmetro de forma.

2.2.5 Distribuição Burr2

Diz-se que uma v.a., Y, segue uma distribuição Burr2, $Y \sim B2(\lambda, \delta, \kappa, \rho)$, quando a sua f.d.p. é dada por

$$f(y \mid \theta) = \frac{2\kappa\rho e^{\frac{y-\lambda}{\delta}}\left(e^{\frac{y-\lambda}{\delta}}+1\right)^{\rho-1}}{\delta\left(2-\kappa+\kappa\left(e^{\frac{y-\lambda}{\delta}}+1\right)^{\rho}\right)^{2}}, -\infty < y < +\infty, \qquad (2.24)$$

com $\theta=(\lambda,\delta,\kappa,\rho)$, $\delta,\kappa,\rho>0$ e $\lambda\in IR$, sendo λ o parâmetro de localização e δ o parâmetro de escala. A f.d. é

$$F(y\mid\theta)=1-\frac{2}{\kappa\left(\left(e^{\frac{y-\lambda}{\delta}}+1\right)^{\rho}-1\right)+2} \tag{2.25}$$

Não foi possível obter uma expressão explícita para a f.g.m.. A expressão para os momentos ordinários de ordem i é

$$E\left[Y^{i}\right]=2\kappa\rho\int_{0}^{+\infty}\frac{(\lambda+\delta\ln(u))^{i}(1+u)^{\rho-1}}{\left(2-\kappa+\kappa[1+u]^{\rho}\right)^{2}}du. \tag{2.26}$$

A mediana é dada por $me=\lambda-\delta\ln\left[\left(\frac{2+\kappa}{\kappa}\right)^{\frac{1}{\rho}}-1\right]^{-1}$

Graficamente a f.d.p. tem o seguinte aspecto

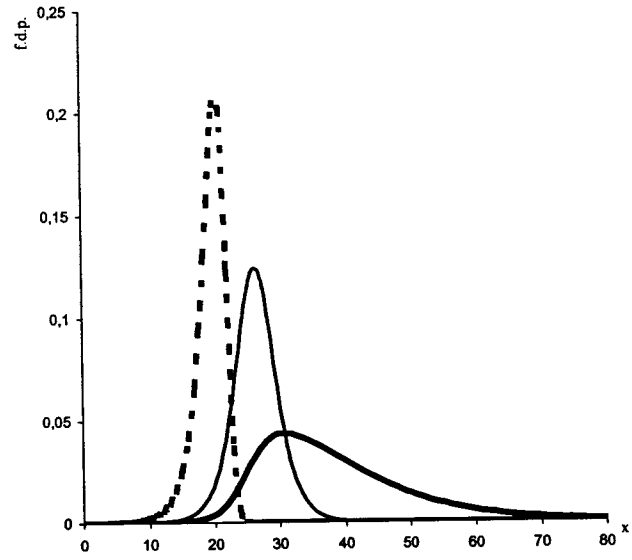

Figura2.2 – f.d.p Burr2 para λ=25, δ=2 e κ=1. As curvas correspondem, respectivamente: tracejado ρ=15, simples ρ=1 e negrito ρ=0,2.

À medida que ρ vai diminuindo, a f.d.p. passa de assimétrica positiva a assimétrica negativa e a curtose vai diminuindo.

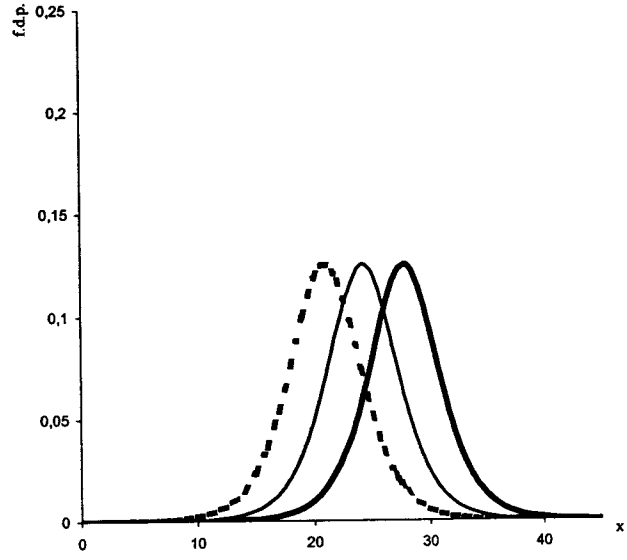

Figura2.3 – f.d.p Burr2 para λ=25, δ=2 e ρ=1. As curvas correspondem, respectivamente: tracejado κ=15, simples κ=3 e negrito κ=0,5.

Fixando λ, δ e $\rho = 1$, para qualquer valor de κ as distribuições são simétricas e têm a mesma curtose. Assim, nestas condições, κ funciona como um parâmetro de localização.

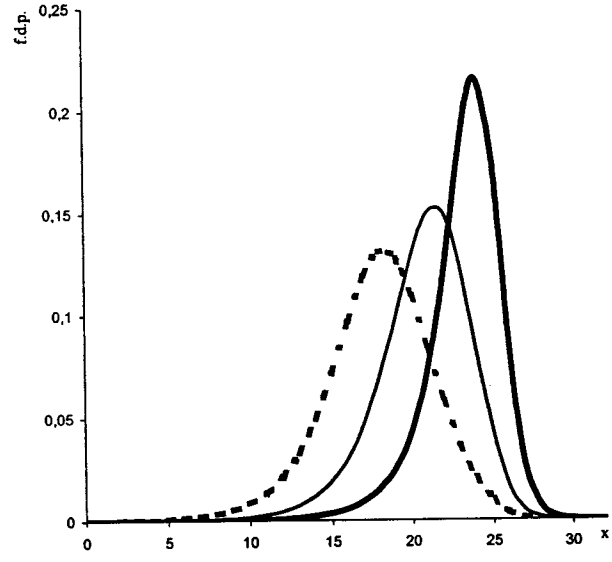

Figura2.4 – f.d.p Burr2 com λ=25, δ=2 e ρ=4. As curvas correspondem, respectivamente: tracejado κ=15, simples κ=3 e negrito κ=0,5.

Se $\rho > 1$, a assimetria é sempre negativa e à medida que κ vai diminuindo, a curtose vai aumentando.

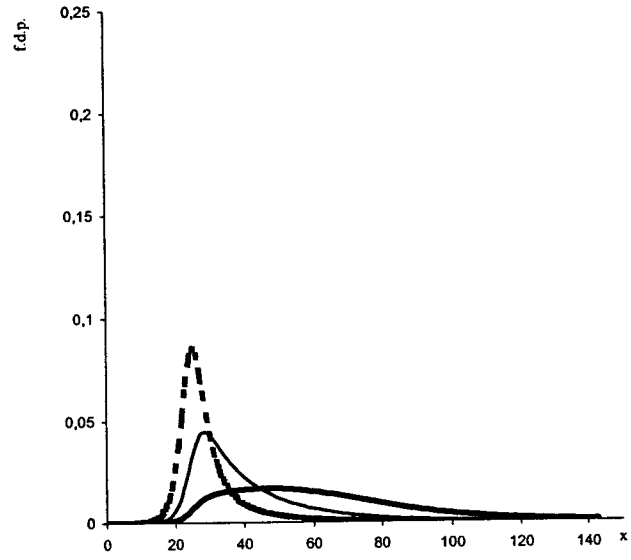

Figura2.5 – f.d.p Burr2 para λ=25, δ=2 e ρ=0,1. As curvas correspondem, respectivamente: tracejado κ=15; simples κ=3 e negrito κ=0,5.

Para $\rho < 1$, a simetria é sempre positiva e à medida que κ vai diminuindo, a curtose vai diminuindo.

Note-se que a curva referente a esta distribuição apresenta várias formas, abrangendo uma série de situações.

Da observação dos gráficos apresentados, pode-se concluir que κ e ρ podem ser interpretados como parâmetros de forma, onde κ afecta essencialmente a curtose e ρ a assimetria.

2.2.6 Misturas finitas

As misturas finitas de distribuições vêm, desde há muito, merecendo a atenção dos estatísticos, não só pelo seu interesse teórico como também pelo grande número de aplicações que encontram na prática. Ciências como Medicina, Biologia, Ecologia, Engenharia, Física, Química e Psicologia são algumas áreas onde essas aplicações podem

ocorrer. Por exemplo, na Investigação Pesqueira, muitas vezes é fornecido o comprimento dos peixes, mas o sexo é desconhecido ou provêm de gerações diferentes.

A análise estatística de uma amostra deste tipo não é tarefa fácil devido a duas razões fundamentais. Em primeiro lugar, porque os estimadores dos vários parâmetros não apresentam, em geral, uma forma explícita, sendo necessário recorrer a métodos iterativos. Em segundo lugar, as dificuldades teóricas que surgem em alguns aspectos de análise estatística revelam que alguns problemas comuns em mistura não são *standard*

Neste trabalho, baseado no estudo efectuado por Cabral (1987a), faz-se uma breve referência ao teste de existência de misturas em escala e em localização de duas populações desconhecidas que se irá utilizar.

2.2.6.1 Definições básicas e conceitos

Considera-se a existência de misturas, absolutamente contínuas, em escala e em localização de populações conhecidas com a mesma família paramétrica.

Suponha-se que uma v.a. Y toma valores num espaço S e tem f.d.p. dada por

$$f(y \mid \mathbf{\Pi}) = \sum_{i=1}^{d} \varphi_i f_i(y \mid \theta) \tag{2.27}$$

onde $\mathbf{\Pi} = (\theta, \varphi)$ é constituído por todos os parâmetros, $\varphi = (\varphi_1, \ldots, \varphi_d)$ são as proporções ou os parâmetros da mistura, com $0 < \varphi_i < 1 \ \ (i = 1, \ldots, d)$ e $\sum_{i=1}^{d} \varphi_i = 1$, e $f_i(y \mid \theta)$ $(i = 1, \ldots, d)$ são f.d.p.'s, as componentes da mistura (que neste caso pertencem à mesma família paramétrica). Pode-se, então, considerar φ_i como sendo a probabilidade de que X provenha da população com f.d.p. $f_i(y \mid \theta)$.

Definição 2.2 A v.a. Y cuja f.d.p. é dada por (2.27) diz-se uma *mistura - φ* ou uma *mistura finita.*

Como só se faz referência a misturas finitas, omite-se a palavra finita, em tudo o que se segue.

Definição 2.3 Uma mistura - φ diz-se uma *mistura em escala* se

$$f(y \mid \mathbf{\Pi}) = \sum_{i=1}^{d} \varphi_i \frac{1}{\delta_i} f\left(\frac{y}{\delta_i}\right),\ \delta_1,\ldots,\delta_d > 0 \tag{2.28}$$

Definição 2.4 Uma *mistura* - φ diz-se uma *mistura em localização* se

$$f(y \mid \mathbf{\Pi}) = \sum_{i=1}^{d} \varphi_i f(y - \lambda_i),\ \lambda_1,\ldots,\lambda_d \in IR \tag{2.29}$$

Propriedade 2.1 Se Y é uma *mistura*- φ, então a v.a. $Z = g(Y)$, onde g é uma transformação contínua e monótona, é também uma *mistura* - φ.

Apresenta-se de seguida um conceito extremamente importante e que surge sempre que se tem um problema de estimação ou se pretende fazer um teste, na medida em que garante uma única caracterização para qualquer uma das classes de modelos considerados, a *identificabilidade*.

Definição 2.5 Uma mistura com f.d.p. dada por (2.27) diz-se identificável se

$$f(y \mid \Pi) = \sum_{i=1}^{d} \varphi_i f_i(y \mid \boldsymbol{\theta}) = \sum_{i=0}^{d'} \varphi_i' f_i'(y \mid \boldsymbol{\theta}) \tag{2.30}$$

implicar $d = d'$ e que, para todo o i, existe algum j tal que $\varphi_j = \varphi_i'$ e $f_j(y \mid \boldsymbol{\theta}) = f_i(y \mid \boldsymbol{\theta})$.

Este assunto foi primeiramente investigado por Teicher (1963) que obteve alguns teoremas gerais contendo condições necessárias e suficientes para a identificabilidade de misturas finitas. Em particular, os seus resultados implicam que as misturas de distribuições normais e distribuições gama são identificáveis. Em Yakowitz e Spragins (1968) encontram-se resultados mais gerais.

Neste contexto Behboodian (1975) provou que:

Propriedade 2.2 As misturas em localização são identificáveis.

Propriedade 2.3 As misturas em escala são identificáveis se a f.d.p. $f(y \mid \mathbf{\Pi})$ tem r-ésimo momento para algum $r > 0$.

2.2.6.2 Teste para a identificação da existência de misturas em populações desconhecidas

Os testes de existência ou não existência de mistura são um assunto pouco trabalhado devido à sua complexidade. Para dar uma noção disso, considere-se o problema de testar a hipótese

$$H_0 : f(\cdot \mid \boldsymbol{\theta}) = N(\mu, \sigma^2) \quad \text{vs} \quad H_1 : f(\cdot \mid \boldsymbol{\theta}) = \varphi N(\mu_1, \sigma_1^2) + (1-\varphi) N(\mu_2, \sigma_2^2), \qquad (2.31)$$

onde os parâmetros, em ambas as hipóteses, são desconhecidos. Por outras palavras, pretende-se saber se se tem uma única população normal ou a mistura de duas populações normais. A ideia mais natural será recorrer ao teste de razão de verosimilhança e utilizar a distribuição assintótica da estatística de teste, se a dimensão o permitir, que se esperaria ser $\chi^2_{(r)}$, onde r é o número de restrições impostas a H_1 para obter H_0.

Neste caso tudo se complica. Basta ver que não é única a maneira de se obter H_0 de H_1. Com efeito pode-se ter:

$$\varphi = 0 \qquad \text{(uma restrição)}$$

ou

$$\mu_1 = \mu_2, \sigma_1 = \sigma_2 \qquad \text{(duas restrições)}$$

Perante tal situação foram desenvolvidos vários métodos para se resolver o problema anterior, quer se trate de populações normais ou não, com ou sem componentes conhecidas. Esses métodos podem ser divididos em duas classes: a primeira contendo

técnicas informais, tal como o simples exame de um histograma; a segunda técnicas formais, isto é, testes estatísticos, incluindo, por exemplo, os clássicos testes de hipóteses, obtidos pelas mais variadas técnicas, e os testes de ajustamento.

Em Cabral (1987a) pode-se ter uma visão mais detalhada dos trabalhos efectuados até à data, assim como do teste para a existência de misturas em escala e em localização de populações desconhecidas, que se passa a apresentar de uma maneira sucinta.

Em muitas situações que ocorrem na prática a amostra correspondente à v.a. X é suposta ser proveniente de uma mistura de d populações com f.d.p.

$$f(y \mid \mathbf{\Pi}) = \sum_{i=1}^{d} \varphi_i f\left(\frac{y-\lambda_i}{\delta_i}\right) \tag{2.32}$$

com $i = 1,\ldots,d$, $\mathbf{\Pi} = (\varphi_1, \lambda_1, \delta_1, \ldots \varphi_d, \lambda_d, \delta_d)$, onde $\delta_i > 0$ e $\lambda_i \in IR$ se supõem desconhecidos. Supõe-se também $\sum_{i=1}^{d} \varphi_i = 1$ e $\varphi_i \in]0,1]$, a fim de garantir a identificabilidade em ambas as hipóteses, visto que as distribuições são do mesmo tipo e se desconhece os parâmetros que as distinguem. Por isso, a inclusão dos dois extremos do intervalo levaria à situação de, quando não existe mistura, a amostra poder ter tido origem em duas populações distintas, ou seja, a uma situação de não identificabilidade (Cabral 1987a).

O problema que se coloca é o de, com base na amostra, se decidir se se trata de uma mistura ou não.

Consideramos as seguintes condições:

(i) $\lambda_1 = \lambda_2 = \ldots = \lambda_d = \lambda$ conhecido, com $\lambda = 0$ por conveniência (2.33a)

δ_i desconhecidos (mistura em escala)

(ii) $\delta_1 = \delta_2 = \ldots = \delta_d = \delta$ conhecido, com $\delta = 1$ por conveniência (2.33b)

λ_i desconhecidos (mistura em localização)

Para cada um dos casos (i) ou (ii), o objectivo é, com base numa sucessão amostral $\{Y_1,\ldots,Y_n\}$ de v.a.'s i.i.d., testar as hipóteses:

$$H_0 : \text{não há mistura} \qquad \text{vs} \qquad H_1 : \text{há mistura} \tag{2.34}$$

Em qualquer das situações (i) ou (ii), os parâmetros desconhecidos das componentes funcionam como parâmetros perturbadores, os quais se opta por eliminar. Apesar deste facto conduzir a uma perca de eficiência, simplifica bastante o problema. A maneira de fazer "desaparecer" esses parâmetros é através da introdução de uma função $g(Y)$ a partir da qual se pode construir uma função que dependa apenas de φ, sensível à existência ou não de mistura.

Os parâmetros perturbadores poderiam ser estimados; contudo, a extensa literatura dedicada ás soluções deste tipo mostra a necessidade de cálculos extremamente laboriosos, além de resultados pouco satisfatórios.

Seja Y uma v.a. cuja f.d.p. é dada por (2.28) ou (2.29) e g uma função contínua. Os momentos de ordem i de $g(Y)$, $\mu_i' = E\left[g^i(Y)\right]$, são dados por (designando por F a f.d. correspondente às f.d.p. que aparecem nas expressões (2.28) e (2.29)) :

$$\mu_i' = \varphi_1 \int_S g^i(z\delta_1)dF(z) + \ldots + \varphi_d \int_S g^i(z\delta_d)dF(z) \text{ (mistura em escala)} \tag{2.35a}$$

$$\mu_i' = \varphi_i \int_S g^i(z+\lambda_i)dF(z) + \ldots + \varphi_d \int_S g^i(z+\lambda_d)dF(z) \text{ (mistura em localização),} \tag{2.35b}$$

onde S designa o suporte da mistura, devendo ser μ_i', em ambos os casos, finito.

O primeiro objectivo será a determinação de uma função $g(Y)$, tal que:

$$g(\delta z) = g(\delta)g(z) \text{ (mistura em escala)} \tag{2.36a}$$

$$g(z+\lambda) = g(z)g(\lambda) \text{ (mistura em localização)} \tag{2.36b}$$

Seja $\phi(\varphi)=\frac{\mu_2'}{\mu_1'^2}$ (Cabral, 1987a,b), que se pretende independente dos parâmetros perturbadores. Então $g(Y)$ é dada por (Cabral 1987b):

$$g(Y)=|Y|^a \text{ (mistura em escala)} \tag{2.37a}$$

$$g(Y)=\exp(-bY) \text{ (mistura em localização),} \tag{2.37b}$$

com a e b escolhidos convenientemente. Estes foram objecto de estudo (em Cabral 1987a), quando o número de componentes da mistura é dois. O sinal de – no expoente de (2.37b) prende-se a factores de ordem prática. Obteve-se

(i) mistura em escala

$$\phi(\varphi)=\frac{\left[\varphi_1\delta_1^2+\ldots+\varphi_d\delta_d^2\right]}{\left[\varphi_1\delta_1+\ldots+\varphi_d\delta_d\right]^2}\times A, \tag{2.38a}$$

onde $A=\frac{\mu(2a)}{\mu^2(a)}$, sendo $\mu(t)=\int_{-\infty}^{+\infty}|z|^t\,dF(z)>0, t>0$, que se supõe existir e ser finito;

(ii) mistura em localização

$$\phi(\varphi)=\frac{\left[\varphi_1\lambda_1^2+\ldots+\varphi_d\lambda_d^2\right]}{\left[\varphi_1\lambda_1+\ldots+\varphi_d\lambda_d\right]^2}\times A, \tag{2.38b}$$

onde $A=\frac{\mu(2b)}{\mu^2(b)}$, com $\mu(t)=\int_{-\infty}^{+\infty}\exp(-zt)dF(z)$ definida pelo menos numa vizinhança da origem.

Designando

$$\Lambda(\varphi)=\frac{\phi(\varphi)}{A}, \tag{2.39}$$

vem

$$\begin{cases} \Lambda(\varphi)=1 & \text{se} \quad \exists i:\varphi_i=1 & \text{não há mistura} \\ \Lambda(\varphi)>1 & \text{se} \quad \exists i,j\ (i\neq j):0<\varphi_i,\varphi_j<1 & \text{há mistura} \end{cases} \tag{2.40}$$

em qualquer dos casos (i) ou (ii). Obtém-se assim uma função sensível à existência ou não de mistura que não depende dos parâmetros desconhecidos, δ_i ou λ_i respectivamente quando se tem uma mistura em escala ou uma mistura em localização, com $i=1,\ldots,d$.

O teste (2.34) poder-se-á então escrever da forma:

$$H_0:\Lambda(\varphi)=1, \text{ não há mistura} \quad \text{vs} \quad H_1:\Lambda(\varphi)>1, \text{ há mistura} \tag{2.41}$$

Efectuando as transformações (2.37a) e (2.37b) respectivamente. Para a amostra aleatória $(Y_1,\ldots,Y_n)$ a estatística de teste correspondente é dada por:

$$T_n=\frac{t_2'}{t_1'^2}\times A^{-1}, \tag{2.42}$$

onde

$$t_k'=\begin{cases} \dfrac{1}{n}\displaystyle\sum_i^n\left(|Y_i|^a\right)^k & \text{mistura em escala} \\ \dfrac{1}{n}\displaystyle\sum_i^n \exp\left(-kbY_i\right) & \text{mistura em localização.} \end{cases} \tag{2.43}$$

T_n estima $\Lambda(\varphi)$, sendo a regra de decisão do teste unilateral definido em (2.41) dada por:

- $T_n \le c_n$, decide-se que não há mistura
- $T_n > c_n$, decide-se que há mistura,

(2.44)

com c_n dado por

$$\text{Prob}\{\text{rejeitar}H_0 \mid \exists i:\varphi_i=1\}=\alpha, \tag{2.45}$$

sendo α o nível se significância do teste.

A determinação do ponto crítico c_n faz-se com base na distribuição assintótica de T_n.

A v.a.

$$T'_n = \frac{h(t'_1,t'_2)-h(\mu'_1,\mu'_2)}{\sigma(\varphi)}\sqrt{n} \tag{2.46}$$

é assintoticamente normal com valor médio zero e variância um, $N(0,1)$ (Cabral, 1987a), com

$$h(\mu'_1,\mu'_2)=\frac{\mu'_2}{(\mu'_1)^2}\times A^{-1} \text{ e } h(t'_1,t'_2)=T_n.$$

Sendo

$$\sigma^2(\varphi)\cong \sigma_{11}\left(\frac{\partial h}{\partial \mu'_1}\right)^2 + 2\sigma_{12}\left(\frac{\partial h}{\partial \mu'_1}\right)\left(\frac{\partial h}{\partial \mu'_2}\right)+\sigma_{22}\left(\frac{\partial h}{\partial \mu'_2}\right)^2 \tag{2.47}$$

com

$$\sigma_{11}=\mu'_2-\mu'^2_1,\ \sigma_{12}=\mu'_3-\mu'_1\ \mu'_2 \text{ e } \sigma_{22}=\mu'_4-\mu'^2_2$$

e

$$\mu'_j=\begin{cases}\left[\varphi_1\delta_1^j+\ldots+\varphi_d\delta_d^j\right]\mu(ja) & \text{mistura em escala}\\ \left[\varphi_1\lambda_1^j+\ldots+\varphi_d\lambda_d^j\right]\mu(jb) & \text{mistura em localização}\end{cases} \tag{2.48}$$

vem

$$\sigma^2(\varphi)\cong A^{-2}\mu'^{-6}_1\left(4\mu'^3_2-4\mu'_1\mu'_2\mu'_3+\mu'_4\mu'^2_1-\mu'^2_1\mu'^2_2\right). \tag{2.49}$$

Designando por $P_R(\varphi)$ a função potência do teste, ou seja, $P_R(\varphi)=P\{\text{decidir mistura}\mid\varphi\}$, deve-se ter $P_R(1)=P\{\text{decidir mistura}\mid\exists i:\varphi_i=1\}=\alpha$, cuja solução assintótica é dada por

$$c_n=1+V_\alpha\frac{\sigma(1)}{\sqrt{n}} \tag{2.50}$$

onde V_α é o quantil $(1-\alpha)$ da $N(0,1)$, ou seja, é a solução da equação $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt = 1-\alpha$ e $\sigma(1)$ é o valor de $\sigma(\varphi)$ quando não existe mistura.

Pode-se observar que $c_n \underset{n\to\infty}{\to} 1$.

Em Cabral (1987a) prova-se que o teste é consistente.

2.2.6.3 Aplicação do teste para a existência ou não de mistura

Pretende-se, com base na amostra $(y_1,\ldots,y_n)$, testar as hipóteses:

H_0 : não há mistura vs H_1 : há mistura

utilizando a seguinte regra de decisão

- $T_n \le c_n$, decide-se que $\varphi = 1$ (não há mistura);
- $T_n > c_n$, decide-se que : $0 < \varphi < 1$ (há mistura).

2.2.6.3.1 Mistura em escala de duas distribuições normais

Considere-se a v.a. Y cuja f.d. dada por

$$f(y \mid \theta) = \frac{\varphi}{\delta_1\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\delta_1^2}\right) + \frac{(1-\varphi)}{\delta_2\sqrt{2\pi}} \exp\left(-\frac{y^2}{2\delta_2^2}\right), \quad 0 < \varphi \le 1,\ y \in IR \text{ e}$$

δ_1, $\delta_2 > 0$ desconhecidos (mistura em escala de duas distribuições normais, com $\lambda = 0$ por conveniência (2.33a)). Repare-se que, considerando $\ln Y$ em vez de Y (prop. 2.1 pág.23), obtém-se uma mistura em escala de duas distribuições lognormais.

Uma vez que $\mu(t) = \int_{-\infty}^{+\infty} |z|^t \, dF(z) = 2\int_0^{+\infty} \frac{z^t}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \frac{2^{\frac{t}{2}}}{\sqrt{\pi}} \Gamma\left[\frac{t+1}{2}\right]$,

vem $A = \sqrt{\pi} \frac{\Gamma\left[\frac{2a+1}{2}\right]}{\Gamma^2\left[\frac{a+1}{2}\right]}$.

De (2.48), $\mu'_j = \left[\varphi\delta_1{}^j + (1-\varphi)\delta_2{}^j\right] \mu(ja)$, vindo, de (2.49),

$$\sigma^2(1) = A^{-2}\mu^{-6}(a)\left(4\mu^3(2a) - 4\mu(a)\mu(2a)\mu(3a) + \mu(4a)\mu^2(a) - \mu^2(a)\mu^2(2a)\right).$$

Prova-se que, para uma mistura de duas distribuições normais com parâmetros de dispersão desconhecidos, $a = 1{,}75$ (Cabral 1987a). Conhecendo a, calcula-se $A = 2{,}5415$ e $\sigma(1) = 1{,}2657$. O ponto crítico é, então, definido por

$$c_n = 1 + V_\alpha \frac{1{,}2657}{\sqrt{n}},$$

e a estatística de teste será

$$T_n = \frac{t'_2}{(t'_1)^2} \times 0{,}3935,$$

onde

$$t'_1 = \frac{1}{n}\sum_{i=1}^{n}\left(|Y_i|^{1,75}\right) \text{ e } t'_2 = \frac{1}{n}\sum_{i=1}^{n}\left(|Y_i|^{1,75}\right)^2.$$

2.2.6.3.2 Mistura em localização de duas distribuições normais

Considere-se agora a v.a. Y com f.d.p. dada por

$$f(y \mid \theta) = \frac{\varphi}{\sqrt{2\pi}} \exp\left(-\frac{(y-\lambda_1)^2}{2}\right) + \frac{(1-\varphi)}{\sqrt{2\pi}} \exp\left(-\frac{(y-\lambda_2)^2}{2}\right), \quad 0 < \varphi \leq 1, \; y \in IR$$

e λ_1, $\lambda_2 \in IR$ desconhecidos (mistura em localização de duas distribuições normais, com $\delta = 1$ por conveniência (2.33b)). Considerando $\ln Y$ em vez de Y (prop. 2.1 pág.23) obtém-se uma mistura em localização duas distribuições lognormais.

De $\mu(t) = \int_{-\infty}^{+\infty} \exp(-zt) dF(z)$, vem $\mu(t) = \exp\left(\frac{t^2}{2}\right)$, o que permite escrever $A = \exp(b^2)$.

De (2.48), $\mu'_j = \left[\varphi\lambda_1^j + (1-\varphi)\lambda_2^j\right] \mu(jb)$, então, de (2.49),

$$\sigma^2(1) = A^{-2}\mu^{-6}(b)\left(4\mu^3(2b) - 4\mu(b)\mu(2b)\mu(3b) + \mu(4b)\mu^2(b) - \mu^2(b)\mu^2(2b)\right).$$

No caso de uma mistura de duas distribuições normais com parâmetros de localização desconhecidos, $b = 0{,}05$ (Cabral 1987a), donde $A = 1{,}0025$ e $\sigma(1) = 0{,}00354$. O ponto crítico será

$$c_n = 1 + V_\alpha \frac{0{,}00354}{\sqrt{n}}$$

e

$$T_n = \frac{t'_2}{(t'_1)^2} \times 0{,}9975,$$

com

$$t'_2 = \frac{1}{n}\sum_{i=1}^{n} \exp(-0{,}1Y_i) \text{ e } t'_1 = \frac{1}{n}\sum_{i=1}^{n} \exp(-0{,}05Y_i).$$

2.2.6.4 Determinação do número de componentes da mistura

Se a hipótese H_0 não é rejeitada, o problema fica resolvido, $d = 1$ ou seja, decide-se que não existe mistura. No caso de se rejeitar H_0, admite-se que existe mistura; no entanto, não se sabe qual o valor de *d*, isto é, o número de componentes da mistura. Atendendo a que se está a trabalhar com misturas identificáveis, o problema da identificação das componentes das misturas é equivalente ao problema de identificação dos

"clusters" (Boes, 1966). Para este caso, demonstra-se (Yakowitz, 1969) que existe um estimador consistente que resolve o problema da identificação.

Baseado neste facto e na técnica de misturas de normais utilizada em análise de "clusters", cujo critério se baseia na maximização do logaritmo da função verosimilhança da amostra, desenvolveu-se a técnica iterativa que se apresenta de seguida através do seguinte algoritmo (Cabral, 1993):

(1) Dada a amostra $(y_1,\ldots,y_n)$ testa-se a hipótese (2.34), com base na estatística T_n dada por (2.42). Se a hipótese H_0 não é rejeitada, $d=1$. FIM

Se H_0 é rejeitada passa-se ao passo (2).

(2) Começa-se por assumir $d=2$ e obtêm-se os estimadores de MV dos parâmetros da distribuição.

(3) Divide-se a amostra $(y_1,\ldots,y_n)$ em duas subamostras $P_1=(y_{11},\ldots,y_{1n_1})$ e $P_2=(y_{21},\ldots,y_{2n_2})$ com $n_1+n_2=n$ de acordo com a regra:

$$\begin{cases} y_i \in P_1 & \text{se} \quad P(1\mid y_i)\geq P(2\mid y_i) \\ y_i \in P_2 & \text{se} \quad P(1\mid y_i)< P(2\mid y_i) \end{cases}$$

onde

$$P(1\mid y_i)=\frac{\hat{\varphi}f(y_i\mid\theta)}{\hat{\varphi}f(y_i\mid\theta)+(1-\hat{\varphi})\hat{\varphi}f(y_i\mid\theta)},$$

e $P(2\mid y_i)=1-P(1\mid y_i)$, sendo $P(j\mid y_i)$ a probabilidade da y_i-ésima observação pertencer à j-ésima população com $j=1,2$.

(4) Para cada uma das subamostras obtidas em (3) repete-se os passos (1), (2) e (3) até se decidir $d=1$ para cada uma das subamostras assim obtidas. O valor de d é finalmente dado pelo número de subamostras para as quais se decidiu $d=1$.

3. Estimação pontual

3.1 Estimadores e estimativas

Assumindo que um dado fenómeno aleatório se comporta de acordo com um modelo específico, envolvendo vários parâmetros, uma questão básica é a de obter estimativas dos parâmetros do modelo que sejam compatíveis com as observações, isso é, com a informação contida na amostra. Pode-se, assim, considerar que se tem o objectivo de caracterizar a população a partir da qual a amostra foi retirada, procurando designadamente estimar os parâmetros desta população.

A cada modelo que se utilize para representar o comportamento aleatório da população corresponde o seu próprio problema de estimação. Certas estatísticas podem ser utilizadas para se obter estimativas dos valores dos parâmetros do modelo de comportamento aleatório da população.

Definição 3.1 *Conceito de estimador.* Considere-se uma população cujo comportamento aleatório é dado por um modelo que envolve o parâmetro θ. Seja $(Y_1,\ldots,Y_n)$ uma amostra aleatória de Y e $T = f(Y_1,\ldots,Y_n)$ uma estatística. Diz-se que T é um estimador de θ quando as realizações de T são utilizadas como estimativas do valor desconhecido de θ.

De um modo geral, o parâmetro a estimar é uma constante. O estimador usado para estimar o parâmetro, sendo uma estatística, é uma v.a., possuidora da sua própria distribuição amostral. Apesar de tal facto, utilizar-se-á o símbolo $\hat{\theta}$ para representar o estimador do parâmetro θ. As realizações do estimador são as estimativas.

Observe-se que o mesmo parâmetro desconhecido, θ, pode ter várias estimativas, obtidas utilizando o mesmo estimador sobre observações diferentes, ou utilizando estimadores diferentes. O que põe o problema de decidir que critério utilizar para escolher um estimador, de entre as possibilidades existentes.

Nas secções seguintes começar-se-á por analisar um conjunto de características desejáveis dos estimadores pontuais (considerando o caso unidimensional) e, posteriormente, apresentar-se-ão diferentes métodos de estimação (isto é, métodos com base nos quais tais estimadores podem ser obtidos).

3.2 Algumas propriedades dos estimadores pontuais

3.2.1 Não-enviesamento

Sendo o estimador $\hat{\theta}$ uma v.a., as suas realizações têm um certa distribuição amostral. Interessa que as estimativas estejam, o mais frequentemente possível, próximas do valor a estimar. À v.a. $R = \hat{\theta} - \theta$ dá-se o nome de erro de estimação.

Definição 3.2 *Estimador não enviesado.* Diz-se que um estimador $\hat{\theta}$ de um parâmetro θ é não enviesado (centrado) quando $E[\hat{\theta}] = \theta$ e, consequentemente, $E[R] = 0$. Quando $E[\hat{\theta}] \neq \theta$, o estimador diz-se enviesado, sendo $E[R] = E[\hat{\theta}] - \theta$ o enviesamento de T.

Um estimador diz-se, portanto, não enviesado quando o seu enviesamento for nulo, e enviesado no caso contrário.

Se bem que desejável, a propriedade de ser centrado não pode ser o único critério de escolha de um estimador. Face a dois estimadores centrados do mesmo parâmetro θ deve preferir-se o mais preciso; aquele cuja variância é menor.

3.2.2 Eficiência

Para a adopção de um critério de eficiência deve-se ter em conta, para além da variância de cada estimador, também uma medida de dispersão do estimador em redor do parâmetro estimado. A eficiência de um estimador $\hat{\theta}$ (que reflecte a sua precisão potencial) pode ser medida através do erro quadrático médio, definido pela expressão seguinte

$$E\left[\left(\hat{\theta} - \theta\right)^2\right] = E\left[\left(\hat{\theta} - \mu_{\hat{\theta}} - \theta + \mu_{\hat{\theta}}\right)^2\right] = E\left[\left(\hat{\theta} - \mu_{\hat{\theta}}\right)^2\right] + \left(\theta - \mu_{\hat{\theta}}\right)^2, \tag{3.1}$$

onde $\mu_{\hat{\theta}}$ representa a valor esperado de $\hat{\theta}$.

Note-se que a primeira parcela do segundo membro representa a variância do estimador e que a segunda é igual ao quadrado do enviesamento.

Para se comparar a eficiência entre diferentes estimadores recorre-se ao conceito de eficiência relativa. Para dois estimadores quaisquer $\hat{\theta}_1$ e $\hat{\theta}_2$ de um mesmo parâmetro θ, a eficiência do primeiro relativamente ao segundo é dada pela razão

$$\frac{E\left[\left(\hat{\theta}_1-\theta\right)^2\right]}{E\left[\left(\hat{\theta}_2-\theta\right)^2\right]}. \tag{3.2}$$

Para um parâmetro qualquer, se existir algum estimador que seja mais eficiente do que qualquer outro, então aquele estimador diz-se eficiente (ou, se se preferir, absolutamente eficiente).

3.2.3 Consistência

Seja $\hat{\theta}$ um estimador do parâmetro θ e n a dimensão da amostra com base na qual $\hat{\theta}$ é calculado. Este estimador diz-se consistente quando, para qualquer valor positivo ξ, se verifica a condição seguinte:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|<\xi\right)=1. \tag{3.3}$$

Esta condição - a convergência em probabilidade do estimador para o parâmetro - significa que, quando a dimensão da amostra tende para infinito, o estimador consistente se concentra sobre o seu alvo, isto é, toma valores próximos do parâmetro estimado com probabilidade cada vez maior. Mostra-se que, se o enviesamento e a variância de um estimador tenderem para zero - e, portanto, tender para zero o valor esperado de $\left(\hat{\theta}-\theta\right)^2$ (convergência em média quadrática de $\hat{\theta}$ para θ) - quando a dimensão da amostra tender para infinito, então o estimador será consistente. Em notação simbólica, estas condições suficientes para a consistência de um estimador podem ser expressas nos termos seguintes

$$\lim_{n\to\infty}\left(\mu_{\hat{\theta}}-\theta\right)=0,\ \lim_{n\to\infty}\sigma^2_{\hat{\theta}}=0 \text{ ou } \lim_{n\to\infty}E\left[\left(\hat{\theta}-\theta\right)^2\right]=0 \tag{3.4}$$

3.2.4 Suficiência

Quando se resume uma amostra através de uma estatística, há perda de informação. Com efeito, dispondo das observações, pode calcular-se o valor de qualquer estatística. O recíproco não é verdadeiro; conhecendo o valor de uma estatística, não se pode, em geral, reproduzir os dados em que o seu cálculo se baseou. As estatísticas «resumem» as observações em certa perspectiva; mas os resumos contêm menos informação do que os dados originais.

No âmbito da estimação pontual, a suficiência de uma estatística traduz a capacidade que ela tem de condensar toda a informação que, relativamente ao parâmetro estimado, esteja contida no conjunto das observações que integram a amostra. Por outras palavras, uma amostra (constituída por n observações) não contém mais informação relativamente ao parâmetro estimado do que um estimador suficiente calculado a partir dela.

Convém dispor de um método objectivo que permita reconhecer, mecanicamente se, perante um candidato a estimador de certo parâmetro, ele é, ou não, suficiente.

A informação contida numa amostra acerca de um parâmetro a estimar obtém-se à custa da chamada função de verosimilhança da amostra, conceito a desenvolver na secção 3.3.2.

Considere-se uma população caracterizada por uma v.a contínua Y com f.d.p. $f(y \mid \theta)$. Dada uma amostra aleatória $(Y_1,\ldots,Y_n)$, a sua f.d.p. conjunta é

$$f(y_1 \mid \theta)\times\ldots\times f(y_n \mid \theta)=\prod_{i=1}^{n} f(y_i \mid \theta) \tag{3.5}$$

pelo facto de $Y_1,\ldots,Y_n$ serem independentes.

Note-se que esta expressão dá a densidade de probabilidade, expressa em função do parâmetro desconhecido θ, de se obterem as observações $(y_1,\ldots,y_n)$ quando o modelo de comportamento aleatório assumido para a população é especificado por $f(y \mid \theta)$. Quando considerada como função de θ, designa-se por função de verosimilhança da amostra, $L(\theta \mid y_1,\ldots,y_n)$.

Definição 3.3 *Estatísticas suficientes.* Uma estatística $T = t(y_1,\ldots,y_n)$ é suficiente para θ sse a função de verosimilhança é da forma: $L(\theta \mid y_1,\ldots,y_2) = g(t(y_1,\ldots,y_n) \mid \theta)$.

A definição anterior significa que a função de verosimilhança apenas envolve as observações $(y_1,\ldots,y_n)$ através da expressão $t(y_1,\ldots,y_n)$ do estimador e o próprio parâmetro θ.

3.2.5 Robustez

Duas propriedades exigidas a um estimador robusto são a resistência e a robustez de eficiência, discutidas, por exemplo, em Huber (1972), Hampel (1974) e Mosteller e Tukey (1977).

Um estimador é resistente se é afectado apenas até um certo limite quer por um pequeno número de erros grosseiros, quer por qualquer número de pequenos arredondamentos e erros de agrupamento. Um estimador é resistente a erros grosseiros se um pequeno subconjunto da amostra não poder ter um efeito desproporcionado na estimativa. Um estimador é resistente a erros de arredondamento e agrupamento se responde continuamente a pequenos erros e, para além disso, se a estimativa é pouco afectada por arredondamento ou agrupamento de uma pequena parte das observações. Normalmente teme-se mais os efeitos dos erros grosseiros do que os dos erros de arredondamento ou agrupamento.

Um estimador tem robustez de eficiência relativamente a um conjunto de distribuições se a sua variância (ou, se o estimador for enviesado, o seu erro quadrático médio) não se afastar muito da do estimador de variância mínima para cada distribuição.
A robustez de eficiência garante que o estimador é bom quando são recolhidas amostras repetidas de uma distribuição que não é conhecida com exactidão. Além disso, a estimativa deve sofrer apenas uma pequena alteração quando a amostra estiver contaminada. A contaminação pode ocorrer quer por erros grosseiros (outliers), quer por erros de arredondamento ou agrupamento entre as observações.

Definidas algumas das propriedades desejáveis dos estimadores, a questão que se coloca seguidamente é a de saber como os definir. Não existe um método geral único que permita especificar estimadores ideais em todas as circunstâncias. Nas secções que se seguem serão analisados alguns métodos alternativos de estimação pontual.

3.3 Métodos de estimação

Quando se tem por objectivo obter estimadores dos parâmetros de uma distribuição pode-se utilizar o método dos momentos, que, apesar das suas limitações, apresenta grande simplicidade em termos de cálculo. Outro dos processos mais utilizados é o método da máxima verosimilhança (MV). Contudo, em muitas situações, não é possível alcançar expressões explícitas para os estimadores MV (EMV), o que leva à utilização de métodos iterativos, tais como os métodos de Newton. Estes métodos, em certos casos, podem tornar-se algo complicados, sendo vantajoso, nesta situação, o algoritmo de Esperança-Maximização (EM) como uma aproximação largamente aplicável à computação iterativa das estimativas de MV.

As situações onde o algoritmo EM é aplicado com vantagem podem ser descritas como problemas de dados incompletos, onde a aplicação da estimação por MV se torna difícil pela ausência de parte dos dados, por estes, por exemplo, se apresentarem agrupados em classes, estarem censurados, envolverem modelos com distribuições truncadas, etc. Mas também é útil numa série de situações onde o facto dos dados serem incompletos não é natural ou evidente, o que inclui modelos estatísticos tais como efeitos aleatórios, misturas, convoluções, modelos log-lineares, classes latentes e estruturas de classes latentes. Assim o algoritmo EM tem aplicação em quase todos os campos onde as técnicas estatísticas tem vindo a ser aplicadas.

Apesar do desenvolvimento do algoritmo EM e consequente metodologia, juntamente com o rápido desenvolvimento dos meios informáticos, tornarem a análise dos problemas de dados incompletos mais tratáveis do que no passado, pode, em certas situações, a convergência ser muito lenta ou a sua implementação ser impossível. Tal deu origem ao desenvolvimento de versões modificadas do algoritmo, tais como métodos baseados em simulação, por exemplo o algoritmo EM de Monte Carlo (EMMC) , e outras extensões deste.

3.3.1 Método dos momentos

Foi no virar do século que Karl Pearson defendeu e desenvolveu o método dos momentos.

Considere-se uma população representada pela v.a. Y cuja distribuição é conhecida a menos de p parâmetros, $\boldsymbol{\theta} = (\theta_i)^T, i = 1,\ldots,p \in IN$. Em geral, os momentos ordinários da população são funções conhecidas dos parâmetros a estimar, facto que se pode expressar da forma seguinte

$$\mu_i' = \mu_i'(\boldsymbol{\theta}). \tag{3.6}$$

Seja $(Y_1, Y_2, \ldots, Y_n)$ uma amostra aleatória obtida a partir daquela população e denotem-se os momentos amostrais ordinários com base naquela amostra por

$$M_i' = \frac{1}{n}\sum_{j=1}^{n}(Y_j)^i \ . \tag{3.7}$$

Ao representar os momentos amostrais com a letra maiúscula pretende-se realçar que tais momentos são encarados como v.a.'s.

De acordo com o método dos momentos, o estimador $\hat{\boldsymbol{\theta}}$ do parâmetro $\boldsymbol{\theta}$ é obtido igualando os momentos populacionais aos momentos amostrais, isto é, fazendo

$$M_i' = \mu_i'(\boldsymbol{\theta})\,(i = 1,2,\ldots,p) \tag{3.8}$$

e resolvendo este sistema de equações em ordem a $\theta_1, \theta_2, \ldots, \theta_p$. Tais estimadores são, portanto, calculados por substituição dos momentos da amostra nas expressões que representam os momentos da população em termos dos parâmetros.

Os estimadores obtidos pelo método dos momentos – como os decorrentes de qualquer outro processo – têm de ser julgados em função das propriedades.

Pode demonstrar-se que, sob condições bastante gerais, esses estimadores são consistentes e assintoticamente normais. No entanto, os estimadores obtidos pelo método

dos momentos não são, em regra, assintoticamente eficientes, aspecto em que são inferiores aos EMV (para amostras de grandes dimensões são, em geral, menos eficientes do que os estimadores obtidos por este último método, que será analisado na próxima secção). Por outro lado, há que destacar a maior simplicidade por vezes conseguida (em vários casos os estimadores de um e outro até coincidem) e a possibilidade de cálculo em situações em que não existem EMV. Também podem ser utilizados como valores iniciais para os algoritmos iterativos de MV.

Em relação ao método que acabou de ser apresentado, deve notar-se que ele admite algumas variantes, sendo todas elas incluídas sob a designação genérica de «método dos momentos». Por exemplo, em vez de se definir o sistema de equações com base nos momentos ordinários, pode recorrer-se aos momentos centrados. Ao fazê-lo, podem-se obter estimadores diferentes para os parâmetros.

3.3.2 Método da máxima verosimilhança

Em 1922, num célebre artigo, "On the mathematical foundations of theoretical statistics", em que introduziu os conceitos fundamentais de consistência, eficiência, verosimilhança e informação, R.A.Fischer criticou asperamente a ineficiência do método dos momentos e introduziu o método da máxima verosimilhança cujo desenvolvimento apresentou em 1925 na não menos célebre "Theory of statistical estimation".

Seja $\boldsymbol{Y}$ um vector aleatório com f.d.p. $g(\boldsymbol{y} \mid \boldsymbol{\theta})$, onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ é o vector que contém os parâmetros desconhecidos e Ω o espaço dos parâmetros.

Se, por exemplo, $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ representar uma amostra aleatória observada de tamanho n de um vector aleatório com f.d.p. $f(\cdot \mid \boldsymbol{\theta})$, então $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ é o valor observado de um vector aleatório $\boldsymbol{Y}$ com f.d.p.

$$g(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{y}_j \mid \boldsymbol{\theta}). \tag{3.9}$$

A função de verosimilhança para $\boldsymbol{\theta}$, formada a partir das observações , é dada por

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = g(\boldsymbol{y} \mid \boldsymbol{\theta}). \tag{3.10}$$

Pretende-se maximizar a função (3.10), sendo muitas vezes, esse valor um ponto estacionário de $L(\boldsymbol{\theta} \mid \boldsymbol{y})$. Nesse caso, pelo método de MV, um estimador $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ pode ser obtido como uma solução, em ordem a cada um do parâmetros, do sistema de equações

$$\frac{\partial L(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}} = 0. \tag{3.11}$$

O objectivo da estimação por MV (Lehmann, 1983) é determinar um estimador para cada um dos θ_i, de tal forma que defina uma sequência de soluções do sistema de equações de verosimilhança (3.11) que seja consistente e assintoticamente eficiente. Tal sequência existe sob determinadas condições de regularidade (Crámer, 1946). Essa sequência de soluções, com as desejadas propriedades assintóticas, é dada por $\hat{\boldsymbol{\theta}}$ para cada $\boldsymbol{\theta}$ como sendo a solução que globalmente maximiza a verosimilhança; ou seja, é o EMV.

Apesar disso, o sistema (3.11) pode não ter solução (por exemplo, quando o máximo global está na fronteira de Ω e não é ponto de estacionaridade da função de verosimilhança) ou admitir vários pontos estacionários (candidatos a extremos). Nos métodos iterativos, os pontos de sela não representam problema pois qualquer erro de arredondamento fará com que as iterações tendam a divergir deles. Em geral, para modelos de estimação, a verosimilhança tem um máximo global no interior do espaço dos parâmetros. Podem, porém, os pontos estacionários que se obtêm corresponderem a mínimos ou a máximos locais que não sejam globais.

Na prática, em vez de se trabalhar com a expressão (3.10) costuma-se trabalhar com o respectivo logaritmo, isto é, com

$$V(\boldsymbol{\theta} \mid \boldsymbol{y}) = \ln L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{j=1}^{n} \ln f(y_j \mid \boldsymbol{\theta}), \tag{3.12}$$

denominada por função de log-verosimilhança. Assim, equivalentemente a (3.11), para se obter o EMV de $\boldsymbol{\theta}$ pode-se resolver

$$\frac{\partial V(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}} = 0. \tag{3.13}$$

Observe-se que, enquanto a função de verosimilhança representa uma probabilidade, a função de log-verosimilhança representa uma soma de logaritmos de probabilidades não sendo, pois, uma probabilidade nem uma f.d.p..

Em termos gerais, os estimadores MV têm propriedades que os colocam entre os mais adequados. Salienta-se, em particular as seguintes:

- em geral, são consistentes;
- embora nem sempre sejam não enviesados (gozam até de uma propriedade de invariância que não é possuída pelos estimadores não enviesados) e eficientes, tendem a possuir estas propriedades à medida que as dimensões das amostras aumentam;
- se existir estimador suficiente, o EMV é, em geral, função desse estimador, verificando-se casos patológicos quando o EMV não é único (a correspondência não é biunívoca);
- se existe estimador mais eficiente, o EMV é único e coincide com esse estimador;
- frequentemente, as suas distribuições são assintoticamente normais.

3.3.2.1 Matriz de variâncias-covariâncias assintótica

O propósito das estatísticas suficientes é reduzir os dados sem perder informação; parece, portanto, lícito inquirir o que é informação? A informação – termo chave da estatística – está entre os conceitos menos consensuais e de mais difícil apreensão.

Tendo consciência das dificuldades limita-se a apresentação ao conceito de informação de Fisher e à posição das estatísticas suficientes no quadro do mesmo conceito.

Para introduzir o conceito de informação de Fisher e algumas proposições que se lhe referem, considere-se primeiro a família $\Im = \{L(\boldsymbol{\theta} \mid \boldsymbol{y}) : \boldsymbol{\theta} \in \Omega\}$. Suponha-se que toda e qualquer f.d.p. da família satisfaz as seguintes condições de regularidade:

[1] - Ω, é um subconjunto de IR^p;

[2] - Os conjuntos $\{\boldsymbol{y} : L(\boldsymbol{\theta} \mid \boldsymbol{y}) > 0\}$ são independentes dos parâmetros;

[3] - Para todo o $\boldsymbol{y}$ e para todo o $\boldsymbol{\theta} \in \Omega$, existem e são finitas as primeiras

derivadas de $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ em ordem a $\boldsymbol{\theta}$;

[4] - Tem-se

$$0 < E_{\theta}\left[\left(\frac{\partial V(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}\right)^2\right] < \infty,$$

para todo o $\boldsymbol{\theta} \in \Omega$.

O vector gradiente da função de log-verosimilhança isto é a função ou estatística "*score*", que exprime, dadas as observações, a taxa relativa de variação da verosimilhança em função de $\boldsymbol{\theta}$, é

$$S(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\frac{\partial L(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}}{L(\boldsymbol{\theta} \mid \boldsymbol{y})} = \frac{\partial V(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}. \tag{3.14}$$

Tem-se, ainda, da definição de função de verosimilhança, que

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \prod_{i=1}^{n} f(y_j \mid \boldsymbol{\theta}),$$

e, uma vez que $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ é a f.d.p. conjunta das observações,

$$\int \ldots \int L(\boldsymbol{\theta} \mid \boldsymbol{y}) dy_1 \ldots dy_n = 1. \tag{3.15}$$

Teorema 3.1 - Se as condições de regularidade [1]-[3] se verificarem e se a derivada do primeiro membro de (3.15) se poder trocar com o integral, então

$$E_{\theta}[S(\boldsymbol{\theta} \mid \boldsymbol{Y})] = 0, \tag{3.16}$$

onde E_{θ} representa a esperança matemática quando o vector dos parâmetros é $\boldsymbol{\theta}$.

Dem. Derivando ambos os membros de (3.15) em ordem a $\boldsymbol{\theta}$, e trocando a derivada com o integral no primeiro membro, obtém-se

$$\int\cdots\int \frac{\partial L(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}} dy_1 \ldots dy_n = 0.$$

O primeiro membro desta expressão é igual a

$$\int\cdots\int \left(\frac{1}{L(\boldsymbol{\theta}\mid \boldsymbol{y})}\frac{\partial L(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}\right) L(\boldsymbol{\theta}\mid \boldsymbol{y}) dy_1 \ldots dy_n = \int\cdots\int \left(\frac{\partial V(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}\right) L(\boldsymbol{\theta}\mid \boldsymbol{y}) dy_1 \ldots dy_n =$$

$$= E_{\boldsymbol{\theta}}\left[\frac{\partial V(\boldsymbol{\theta}\mid \boldsymbol{Y})}{\partial \boldsymbol{\theta}}\right] = E_{\boldsymbol{\theta}}[S(\boldsymbol{\theta}\mid \boldsymbol{Y})],$$

vindo $E_{\boldsymbol{\theta}}[S(\boldsymbol{\theta}\mid \boldsymbol{Y})] = 0$.#

A matriz de informação (esperada) de Fisher é definida por

$$Inf(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[S(\boldsymbol{\theta}\mid \boldsymbol{Y}) S^T(\boldsymbol{\theta}\mid \boldsymbol{Y})\right].$$

Teorema 3.2 - Verificadas as condições de regularidade [1]-[4] e se a segunda derivada do primeiro membro de (3.15) se pode obter derivando duas vezes sob a operação de integração, então

$$Inf(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{Y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]. \tag{3.17}$$

Dem. Derivando (3.15) duas vezes e se a troca entre derivação e integração for válida, então

$$\int\cdots\int \left[\left(\frac{1}{L(\boldsymbol{\theta}\mid \boldsymbol{y})}\frac{\partial L(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}}\right)\frac{\partial L(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}^T} + L(\boldsymbol{\theta}\mid \boldsymbol{y})\frac{\partial}{\partial \boldsymbol{\theta}}\left(\frac{1}{L(\boldsymbol{\theta}\mid \boldsymbol{y})}\frac{\partial L(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}^T}\right)\right] dy_1 \ldots dy_n = 0$$

ou, equivalentemente,

$$\int\cdots\int \left[S(\boldsymbol{\theta}\mid \boldsymbol{y}) S^T(\boldsymbol{\theta}\mid \boldsymbol{y}) + \frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] L(\boldsymbol{\theta}\mid \boldsymbol{y}) dy_1 \ldots dy_n = 0,$$

donde

$$\int\ldots\int S(\boldsymbol{\theta}\mid \boldsymbol{y})S^{T}(\boldsymbol{\theta}\mid \boldsymbol{y})L(\boldsymbol{\theta}\mid \boldsymbol{y})d\boldsymbol{y}_1\ldots d\boldsymbol{y}_n = -\int\ldots\int\frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}L(\boldsymbol{\theta}\mid \boldsymbol{y})d\boldsymbol{y}_1\ldots d\boldsymbol{y}_n \Leftrightarrow$$

$$\Leftrightarrow E_{\boldsymbol{\theta}}\left[S(\boldsymbol{\theta}\mid \boldsymbol{y})S^{T}(\boldsymbol{\theta}\mid \boldsymbol{y})\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}\right],$$

o que permite concluir que

$$Inf(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[S(\boldsymbol{Y}\mid \boldsymbol{\theta})S^{T}(\boldsymbol{Y}\mid \boldsymbol{\theta})\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{Y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}\right].\#$$

Considerando a matriz

$$I(\boldsymbol{\theta}\mid \boldsymbol{y}) = -\frac{\partial^2 V(\boldsymbol{\theta}\mid \boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{T}}, \tag{3.18}$$

também designada por matriz de informação observada de Fisher, então a matriz de informação (esperada) de Fisher, $Inf(\boldsymbol{\theta})$, é dada por

$$Inf(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[S(\boldsymbol{Y}\mid \boldsymbol{\theta})S^{T}(\boldsymbol{Y}\mid \boldsymbol{\theta})\right] = E_{\boldsymbol{\theta}}\left[I(\boldsymbol{\theta}\mid \boldsymbol{Y})\right].\# \tag{3.19}$$

Para o EMV, $\hat{\boldsymbol{\theta}}$, obtém-se a matriz de variâncias-covariâncias assintótica através da inversa da matriz de informação esperada $Inf(\boldsymbol{\theta})$ que, como $\boldsymbol{\theta}$ é desconhecido, pode ser aproximada por $Inf(\hat{\boldsymbol{\theta}})$. O erro padrão assintótico de $\hat{\theta}_i = (\hat{\boldsymbol{\theta}})_i$ é dado por $SE(\hat{\theta}_i) = \left(Inf^{-1}(\boldsymbol{\theta})\right)_{ii}^{1/2}$ (elementos da diagonal principal), podendo ser aproximado por

$$SE(\hat{\theta}_i) \approx \left(Inf^{-1}(\hat{\boldsymbol{\theta}})\right)_{ii}^{1/2} \quad (\mathrm{i} = 1,\ldots,p), \tag{3.20}$$

enquanto os restantes elementos da matriz $Inf^{-1}(\boldsymbol{\theta})$ dão as covariâncias assintóticas dos estimadores.

Uma vez que a matriz de informação esperada $Inf(\boldsymbol{\theta})$ envolve esperanças matemáticas, para se estimar a matriz de variâncias-covariâncias (de uma solução máxima) é mais conveniente, na prática, utilizar a matriz de informação observada $I(\boldsymbol{\theta}\mid \boldsymbol{y})$ dada por

(3.18), em vez da matriz $Inf(\theta)$, avaliada para $\theta = \hat{\theta}$. Assim, é frequente utilizar a aproximação

$$SE(\hat{\theta}_i) \approx \left(I^{-1}(\hat{\theta} \mid y)\right)_{ii}^{1/2} \quad (i = 1,\ldots,p). \tag{3.21}$$

O $SE(\hat{\theta})$ deve incluir, se necessário, algum factor de inflação da variância como se verá na secção 3.3.4.3.

Conhecido o erro padrão é possível construir intervalos de confiança para os diferentes parâmetros. O procedimento standard será considerar para extremos destes intervalos, com nível de confiança $1-\alpha$, os valores saídos da expressão $\hat{\theta}_i \pm V_{\frac{\alpha}{2}} SE(\hat{\theta}_i)$, onde $V_{\frac{\alpha}{2}}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da $N(0,1)$, designado por valor crítico, isto é, $V_{\frac{\alpha}{2}}$ é a solução da equação $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt = 1-\frac{\alpha}{2}$. Tal resulta dos EMV terem, sob certas condições de regularidade, distribuição assintoticamente normal.

Convém notar, por último, que este procedimento standard para a obtenção dos intervalos de confiança é, muitas vezes, "pobre" para amostras pequenas, podendo ser substituído por um outro qualquer procedimento que tenha maior "cobertura".

3.3.3. Aplicação às distribuições teóricas utilizadas

3.3.3.1 Distribuição normal

Seja $Y \sim N(\mu, \sigma^2)$, com f.d.p. dada por (2.7), e $(Y_1,\ldots,Y_n)$ uma amostra aleatória proveniente desta distribuição. Como $E[Y] = \mu$ e $V[Y] = \sigma^2$, do método dos momentos resultam os estimadores $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ e $\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Por outro lado, a função de log-verosimilhança é

$$V(\theta \mid y) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2,$$

e, portanto, os EMV são

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \quad \text{e} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 .$$

A matriz de informação esperada é dada por

$$Inf(\theta) = -E_\theta \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

e a matriz de variâncias-covariâncias assintótica dos estimadores de MV é $Inf^{-1}(\hat{\theta})$, que pode ser aproximada por

$$I^{-1}(\hat{\theta} \mid y) = \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & \frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - \hat{\mu}) \\ \frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(y_i - \hat{\mu}) & -\frac{n}{2\hat{\sigma}^4} + \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 \end{bmatrix}^{-1}$$

ou por

$$Inf^{-1}(\hat{\theta}) \cong \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}.$$

Uma vez que $\hat{\mu} = \bar{y}$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = n\hat{\sigma}^2$ e tendo em conta que o valor esperado das derivadas cruzadas em $I^{-1}(\hat{\theta} \mid y)$ é zero.

3.3.3.2 Distribuição lognormal

Como se viu em na secção 2.2.2, se $Y \sim LN(\lambda, \delta^2)$, vem, com $a = \exp(\delta^2)$,

$$E[Y] = \sqrt{a}e^{\lambda},\ V[Y] = a(a-1)e^{2\lambda},\ \gamma_1 = (a-1)^{\frac{1}{2}}(a+2) \text{ e } \gamma_2 = a^4 + 2a^3 + 3a^2 - 6.$$

Igualando os membros direitos destas equações aos correspondentes momentos amostrais, podem-se obter os estimadores dos parâmetros pelo método dos momentos.

A função de log-verosimilhança é

$$V(\boldsymbol{\theta} \mid \boldsymbol{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln y_i - \mu)^2 - \sum_{i=1}^{n}\ln y_i,$$

pelo que os EMV são

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\ln y_i \text{ e } \hat{\delta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln y_i - \hat{\lambda})^2}$$

Por raciocínio análogo ao utilizado para a distribuição normal é possível aproximar a matriz de informação esperada.

3.3.3.3 Distribuição Gama

Tendo em conta o que foi referido na secção 2.2.3, se $Y \sim G(\lambda, \delta, \rho)$, então

$$E[Y] = \lambda + \delta\rho,\ V[Y] = \delta^2\rho \text{ e } \gamma_1 = \frac{2}{\sqrt{\rho}} > 0,$$

donde, pelo método dos momentos, resulta

$$\hat{\lambda} = \bar{y} - \hat{\delta}\hat{\rho},\ \hat{\delta} = \frac{s}{\sqrt{p}} \text{ e } \hat{\rho} = \frac{4}{\hat{\gamma}_1^2},$$

onde $\bar{y}$, s e $\hat{\gamma}_1$ são, respectivamente, a média, o desvio padrão e a assimetria amostral.

A função de log-verosimilhança é

$$V(\boldsymbol{\theta} \mid \boldsymbol{y}) = -n\rho \ln \delta - n \ln \Gamma(\rho) + (\rho - 1)\sum_{i=1}^{n} \ln(y_i - \lambda) - \frac{1}{\delta}\sum_{i=1}^{n}(y_i - \lambda).$$

Resolvendo as equações de verosimilhança (3.13), facilmente se vê que não existem expressões explícitas para os EMV de $\boldsymbol{\theta}$; para se obter estimativas destes parâmetros deve-se utilizar um processo iterativo. Para isso, deve-se ter em conta os seguintes resultados.

Defina-se a derivada do logaritmo da função gama por

$$\Psi(z) = \frac{d \ln \Gamma(z)}{dz},$$

conhecida como função digama que pode ser definida pela série

$$\Psi(z) = -C + \sum_{n=0}^{+\infty}\left(\frac{1}{n+1} - \frac{1}{n+z}\right).$$

A constante C que muitas vezes aparece em integrais definidos é conhecida como constante de Euler e é definida pelo limite

$$C = \lim_{s \to \infty}\left(\sum_{m=1}^{s}\frac{1}{m} - \ln s\right) = 0{,}57721566490....$$

Note-se que $C = -\Psi(1)$.

O desenvolvimento em série para a derivada de ordem n de $\Psi(z)$ é dado por

$$\Psi^{(n)}(z) = (-1)^{n+1} n! \sum_{k=0}^{+\infty}\frac{1}{(z+k)^{n+1}}.$$

3.3.3.4 Distribuição Burr1

Da secção 2.2.4, se $Y \sim Bl(\lambda, \delta, \rho)$ então

$$E[Y] = \lambda + \delta(C + \Psi(\rho)),\ V[Y] = \delta^2(\Psi'(1) + \Psi'(\rho)),\ \gamma_1 = \frac{\Psi''(\rho) - \Psi''(1))}{\sqrt{(\Psi'(1) + \Psi'(\rho))^3}}$$

Estas equações permitem obter estimativas dos parâmetros pelo método dos momentos. Através das última equação (que só depende de ρ) é possível estimar o valor de ρ, através da assimetria da amostra. Conhecido o valor de ρ, utilizando as duas primeiras equações pode-se estimar os valores para λ e δ, uma vez que

$$\hat{\lambda} = \bar{y} - \hat{\delta}(C + \Psi(\hat{\rho})) \text{ e } \hat{\delta} = \frac{s}{\sqrt{\Psi'(1) + \Psi'(\hat{\rho})}},$$

onde $\bar{y}$ e s são, respectivamente, a média e o desvio padrão da amostra.

Note-se que o coeficiente de assimetria permite estimar ρ se a assimetria amostral estiver compreendida entre –2 e 1,13955. Por exemplo, quando $r \to +\infty$, vem $\gamma_1 \to 1{,}13955$, valor máximo que o coeficiente de assimetria da distribuição pode tomar. Este coeficiente pode tomar valores positivos $(\rho > 1)$ e valores negativos $(\rho < 1)$, anulando-se para $\Psi''(\rho) = -2Z(3) \Leftrightarrow \rho = 1$.

A função de log-verosimilhança é

$$V(\boldsymbol{\theta} \mid \boldsymbol{y}) = n \ln \rho - n \ln \delta - \frac{1}{\delta} \sum_{i=1}^{n} (y_i - \lambda) - (\rho + 1) \sum_{i=1}^{n} \ln\left(1 + e^{-\frac{y_i - \lambda}{\delta}}\right),$$

donde se conclui que não existem expressões explícitas para os EMV dos diferentes parâmetros. Devendo-se recorrer a uma qualquer método iterativo.

3.3.3.5 Distribuição Burr2

Da secção 2.2.5, se $Y \sim B2(\lambda,\delta,\kappa,\rho)$, não é possível obter estimadores pelo método dos momentos e como a função de log-verosimilhança é dada por

$$V(\theta \mid y) = n\ln\frac{2\kappa\rho}{\delta} + \frac{1}{\delta}\sum_{i=1}^{n}(y_i-\lambda) + (\rho-1)\sum_{i=1}^{n}\ln\left(1+e^{\frac{y_i-\lambda}{\delta}}\right) - 2\sum_{i=1}^{n}\ln\left(2-\kappa+\kappa\left(1+e^{\frac{y_i-\lambda}{\delta}}\right)^{\rho}\right)$$

também não é possível obter expressões explícitas para os parâmetros pelo método de MV. Como tal, para se obterem estimativas dos mesmos, deve-se recorrer a qualquer método iterativo.

3.3.3.6 Mistura de distribuições normais

O problema da estimação dos parâmetros de uma mistura pode-se considerar, ainda hoje, em aberto. Este facto deve-se, sem dúvida, à complexidade do problema em si, que, no caso do método dos momentos, conduz a um sistema de equações não só de difícil resolução como, muitas vezes, à não existência de solução real ou a uma solução para os parâmetros $\varphi_i \ (i=1,\ldots,d)$ não compatível com o domínio de existência destes, Rider (1961).

Apresenta-se o desenvolvimento do método dos momentos para uma mistura em localização de duas distribuições normais, visando, principalmente, a obtenção de valores iniciais para os algoritmo iterativos, apresentados nas secções seguintes.

Considerando, uma v.a. Y com f.d.p. dada por

$$f(y \mid \Pi) = \frac{\varphi}{\delta\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\lambda_1}{\delta}\right)^2} + \frac{1-\varphi}{\delta\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\lambda_2}{\delta}\right)^2}, \ y>0,$$

$\lambda_1,\lambda_2 \in IR$ e $\delta>0$.

Fazendo, sem perda de generalidade, $\lambda_2 = \lambda$ e $\lambda_1 = \lambda + d$, tem-se

(1) $\mu_1 = \lambda + \varphi d$

(2) $\mu_2 = \delta^2 + \varphi(1-\varphi)d^2$

(3) $\mu_3 = \varphi d^3(1-\varphi)(1-2\varphi)$

(4) $\mu_4 = 3\delta^4 + 6d^2\delta^2\varphi(1-\varphi) + d^4\varphi(1-\varphi)(1-3\varphi+3\varphi^2)$

Resolvendo as equações (1), (2) e (3) em ordem a λ, δ^2 e d respectivamente, sai,

(5) $\lambda = \mu_1 - \varphi d$

(6) $\delta^2 = \mu_2 - \varphi(1-\varphi)d^2$

(7) $d = \left(\frac{\gamma_1 \mu_2^{3/2}}{\varphi(1-\varphi)(1-2\varphi)} \right)^{\frac{1}{3}}$

Substituindo (6) em (4), resulta

$$d = \left(\frac{\mu_4 - 3\mu_2^2}{\varphi(1-\varphi)(1-6\varphi+6\varphi^2)} \right)^{\frac{1}{4}},$$

que, substituindo em (3), dá

$$\frac{\mu_3^4}{(\mu_4 - 3\mu_2^2)^3} = \frac{\varphi(1-\varphi)(1-2\varphi)^4}{(1-6\varphi+6\varphi^2)^3},$$

ou seja

(8) $$\frac{\gamma_1^4}{\gamma_2^3} = \frac{\varphi(1-\varphi)(1-2\varphi)^4}{(1-6\varphi+6\varphi^2)^3},$$

donde se pode obter o valor de φ.

Assim através de $\bar{y}, s^2, \hat{\gamma}_1$ e $\hat{\gamma}_2$, a média, a variância, a simetria e a curtose da amostra, respectivamente, substituídos nas equações (5), (6), (7) e (8), pode-se obter estimativas para os parâmetros da distribuição. Repare-se que, para uma mistura de duas distribuições, obtêm-se duas estimativas para cada parâmetro, facto que resulta de uma troca nos valores das proporções da mistura.

O método dos momentos, continua ainda hoje a ser desenvolvido e aplicado devido ás dificuldades teóricas e numéricas apresentadas pelo método da MV. Por exemplo, uma vez que a função de log-verosimilhança é

$$V(\boldsymbol{\Pi} \mid \boldsymbol{y}) = n\ln\left(\frac{\varphi}{\delta\sqrt{2\pi}}\right) - \frac{1}{2\delta^2}\sum_{i=1}^{n}(y_i - \lambda_1)^2 + n\ln\left(\frac{1-\varphi}{\delta\sqrt{2\pi}}\right) - \frac{1}{2\delta^2}\sum_{i=1}^{n}(y_i - \lambda_2)^2,$$

não é possível resolver analiticamente o sistema constituído pelas equações de MV, devendo recorrer-se a métodos numéricos, entre os quais o algoritmo EM.

3.3.4 Métodos iterativos de estimação

Muitas vezes, na prática, a função de log-verosimilhança não pode ser maximizada analiticamente ou não se conseguem expressões explícitas para os EMV. Nestes casos, é possível a utilização de algoritmos iterativos para tentar determinar estimativas de MV para $\boldsymbol{\theta}$. As principais alternativas são os métodos de Newton e o algoritmo EM.

De entre os vários métodos numéricos, o algoritmo EM é o que oferece a "garantia" de se atingir o máximo global, embora seja um processo relativamente lento. O método de Newton-Rapshon é bastante mais rápido mas nem sempre converge para o máximo pretendido. A conjugação dos dois métodos poderá levar à construção de um processo iterativo onde se atinge "rapidamente" o máximo global. Paralelamente à "simplicidade" de cálculo que envolve a aplicação do método dos momentos, este, e alguns métodos gráficos, permitem a obtenção de uma solução inicial para os métodos iterativos que se seguem.

3.3.4.1 Métodos de Newton

Na análise numérica existem várias técnicas para determinar os zeros de uma função específica, incluindo o método de Newton-Raphson, e os métodos de Newton modificados. Estes últimos, incluem o algoritmo "scoring" de Fisher e a sua versão modificada utilizando a matriz de informação empírica no lugar da matriz de informação esperada.

Como qualquer processo iterativo precisam de uma solução inicial (valores de "arranque") e uma vez que estes algoritmos poderão ser muito sensíveis a essa solução, convém procurá-la através de um método que permita uma solução rápida e pouco laboriosa. Uma maneira de a conseguir, quando possível, é através do método dos momentos.

3.3.4.1.1 Método de Newton-Raphson.

O método de Newton-Rapshon é um algoritmo iterativo para a obtenção de estimativas de MV. Visando a resolução da equação de verosimilhança

$$S(\boldsymbol{\theta} \mid \boldsymbol{y}) = 0, \tag{3.22}$$

o método de Newton-Raphson aproxima, na iteração $k+1$ o vector gradiente $S(\boldsymbol{\theta} \mid \boldsymbol{y})$ da função de log-verosimilhança $V(\boldsymbol{\theta} \mid \boldsymbol{y})$ por um desenvolvimento em serie de Taylor em torno do ponto $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ (valor obtido na iteração k). Tem-se

$$S(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx S(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}) - I(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}). \tag{3.23}$$

Obtém-se uma nova actualização $\boldsymbol{\theta}^{(k+1)}$ considerando-a como sendo um zero do segundo membro de (3.23). Assim

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + I^{-1}(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}) S(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}) \tag{3.24}$$

A sequência de iterações $\{\boldsymbol{\theta}^{(k)}\}$, $k \in IN_0$, converge para o EMV de $\boldsymbol{\theta}$ se a função de log-verosimilhança for côncava e unimodal. Quando esta não for côncava, não se tem a garantia do método de Newton-Raphson convergir a partir de um valor inicial arbitrário. Sob condições razoáveis em $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ e um valor inicial suficientemente próximo da estimativa de MV, a sequência de iterações $\boldsymbol{\theta}^{(k)}$ produzida pelo método de Newton-Raphson goza de convergência local quadrática para uma solução $\boldsymbol{\theta}^*$ de (3.22). Isto é, dada uma norma $\| \cdot \|$ em Ω, existe uma constante h tal que, se $\boldsymbol{\theta}^{(0)}$ estiver suficientemente próximo de $\boldsymbol{\theta}^*$, então

$$\left\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{*}\right\| \leq h\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{*}\right\|^{2} \tag{3.25}$$

verifica-se para $k = 0,1,2,\ldots$ (ver McLachlan e Krishnan, 1997)

Uma vez que a convergência quadrática é muito rápida, é considerada como a mais valia do método de Newton-Raphson. Contudo, podem surgir inúmeros problemas com a aplicação deste método. Em primeiro lugar, requer em cada iteração o cálculo da matriz de informação $p \times p$, $I\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)$ e a solução de um sistema de p equações. Assim os cálculos computacionais requeridos para uma iteração do método de Newton-Raphson tornam-se "saturantes" á medida que p aumenta. Para além disso, há, nalguns casos, dificuldades neste método na sua forma básica (3.24) relativamente à escolha dos valores iniciais quando se pretende que a sequência de iterações $\left\{\boldsymbol{\theta}^{(k)}\right\}$ convirja para a solução de (3.22) correspondente ao máximo global da função de log-verosmilhança. Há tendência para se obter tanto pontos de sela (de onde depois se acaba por sair ou que não constituem problema) como mínimos locais e máximos locais. Para alguns casos, Böhning and Lindsay (1988) mostraram como este método pode ser modificado, tornando-se monótono.

Uma vez que o método de Newton-Raphson requer actualização de $I\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)$ em cada iteração k, providencia de imediato uma estimativa da matriz de variâncias-covariâncias no seu valor limite $\hat{\boldsymbol{\theta}}$ (assumindo que se trata do EMV), através da inversa da matriz informação observada $I^{-1}\left(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}\right)$. Se o valor inicial é um estimador $\sqrt{n}-$consistente de $\boldsymbol{\theta}$, então a iteração $\boldsymbol{\theta}^{(1)}$ é um estimador assintoticamente eficiente de $\boldsymbol{\theta}$ (McLachlan e Krishnan, 1997).

3.3.4.1.2 Métodos de Newton modificados

De entre os métodos de Newton modificados encontra-se o método "scoring" de Fisher. Este segue um procedimento idêntico ao método de Newton-Raphson, contudo, como alternativa ao cálculo das derivadas parciais de 2ª ordem da função log-verosimilhança, necessárias para a formação de matriz de informação observada $I(\boldsymbol{\theta} \mid \boldsymbol{y})$ para dados i.i.d., em cada actualização $\boldsymbol{\theta}^{(k)}$ de $\boldsymbol{\theta}$, a matriz de informação esperada

$Inf\left(\theta^{(k)}\right)$ é aproximada pela matriz de informação empírica $I_e\left(\theta^{(k)} \mid y\right)$ que se deduz de seguida.

Como os dados são i.i.d., assume-se que as observações y consistem em observações $y_1,\ldots,y_n$ de n v.a.'s i.i.d. com f.d.p. comum, $f(y \mid \theta)$. A função de log-verosimilhança pode ser expressa da seguinte forma

$$V(\theta \mid y)=\sum_{j=1}^{n}\ln L_j(\theta \mid y), \tag{3.26}$$

onde

$$L_j(\theta \mid y)=f(y_j \mid \theta) \tag{3.27}$$

é a função de verosimilhança para θ formada a partir da observação y_j $(j=1,\ldots,n)$.

Pode-se, então, escrever o estatística "*score*"

$$S(\theta \mid y)=\sum_{j=1}^{n}s(\theta \mid y_j), \tag{3.28}$$

com

$$s(\theta \mid y_j)=\frac{\partial \ln L_j(\theta \mid y)}{\partial \theta}, \tag{3.29}$$

a estatística "*score*" para a observação y_j.

Então a matriz de informação esperada pode ser aproximada pela matriz

$$I_e(\theta \mid y)=\sum_{j=1}^{n}s(\theta \mid y_j)s^T(\theta \mid y_j)-\frac{1}{n}S(\theta \mid y)S^T(\theta \mid y). \tag{3.30}$$

Para $\theta=\hat{\theta}$

$$I_e(\hat{\theta} \mid y)=\sum_{j=1}^{n}s(\hat{\theta} \mid y_j)s^T(\hat{\theta} \mid y_j), \tag{3.31}$$

uma vez que $S(\hat{\theta} \mid y)=0$.

Meilijson (1989) apelidou $I_e(\hat{\theta} \mid y)$ de matriz de informação observada empírica. É utilizada na prática para aproximar a matriz de informação observada, $I(\hat{\theta} \mid y)$.

A utilização de (3.31) pode ser justificada da seguinte maneira

$$I(\theta \mid y) = -\frac{\partial^2 V(\theta \mid y)}{\partial\theta\partial\theta^T} = -\sum_{j=1}^{n} \frac{\partial^2 \ln L_j(\theta \mid y)}{\partial\theta\partial\theta^T} =$$

$$= \sum_{j=1}^{n} \left\{ \frac{\partial \ln L_j(\theta \mid y)}{\partial\theta} \right\} \left\{ \frac{\partial \ln L_j(\theta \mid y)}{\partial\theta} \right\}^T - \sum_{j=1}^{n} \frac{1}{L_j(\theta \mid y)} \left\{ \frac{\partial^2 L_j(\theta \mid y)}{\partial\theta\partial\theta^T} \right\}$$

e, como o segundo termo do membro direito desta última expressão tem esperança nula,

$$I(\hat{\theta} \mid y) \approx \sum_{j=1}^{n} s(\hat{\theta} \mid y_j) s^T(\hat{\theta} \mid y_j) = I_e(\hat{\theta} \mid y). \tag{3.32}$$

Esta aproximação será tanto melhor quanto mais próximo estiver $\hat{\theta}$ de θ.

Como é evidente, uma vez que $I_e(\theta^{(k)} \mid y)$ não envolve as derivadas parciais de 2ª ordem da função de log-verosimilhança, este método torna-se menos exigente em termos computacionais que o método de Newton-Raphson.

Se uma solução θ^* de (3.22) é suficientemente próxima de θ e se n for suficientemente grande, prova-se (veja-se McLachlan e Krishnan, 1997), com probabilidade um, sob condições razoáveis para a função de verosimilhança $L(\theta \mid y)$, que a sequência de iterações $\{\theta^{(k)}\}$, gerada pelo método "scoring", exibe convergência linear local para θ^*. Isto é, dada uma norma $\| \cdot \|$ em Ω, existe uma constante $h < 1$ tal que,

$$\left\| \theta^{(k+1)} - \theta^* \right\| \leq h \left\| \theta^{(k)} - \theta^* \right\|, \tag{3.33}$$

para $k = 0,1,2,\ldots$ desde que $\theta^{(0)}$ seja suficientemente próximo de θ^*.

O método de Newton modificado utilizando a matriz de informação empírica (3.32) em (3.23) é análogo ao método de Gauss-Newton para estimação não linear dos mínimos quadrados.

3.3.4.2 Algoritmo EM

Como muitos outros métodos para o cálculo do EMV, o algoritmo EM é um método para determinar os zeros de uma função, $S(\theta \mid y)$. Apesar do algoritmo EM ter sido formulado e aplicado numa variedade de problemas este nome foi dado por Dempster, Laird e Rubin (1977) no seu artigo fundamental (DLR). Foi neste artigo que as ideias foram sintetizadas, estabelecida uma formulação geral do algoritmo, as suas propriedades estudadas e indicadas numa série de aplicações.

O algoritmo EM providencia um processo iterativo para calcular o EMV onde o vector das observações é considerado como sendo incompleto e como uma função, observada, dos chamados dados completos para o qual a estimação por MV será do ponto de vista computacional mais fácil de tratar, de maneira a se obter, se possível, uma expressão explícita para o EMV. A noção de "dados incompletos" inclui o sentido convencional de dados perdidos, mas também se aplica a situações onde os dados completos representam o que pode estar disponível de uma experiência. Neste último caso, os dados completos podem conter algumas variáveis que nunca serão observadas.

Em cada iteração do algoritmo EM há dois passos – o passo da determinação da esperança matemática, o passo-E e o passo de maximização, o passo-M. O passo-E consiste em considerar as observações (dados incompletos) como dados completos (observações mais dados "perdidos"), permitindo a utilização do passo-M (mais simples) a este conjunto de dados "completados". Ou seja, é com a função de log-verosimilhança dos dados completos que se trabalha no passo-E, e, uma vez que parte dos dados é desconhecida, esta é substituída pela sua esperança condicional dadas as observações. Começando com valores iniciais razoáveis para os parâmetros, os passos E e M são repetidos iterativamente até à convergência. A escolha da solução inicial requer uma particular atenção na medida em que a velocidade de convergência deste algoritmo se pode tornar extremamente lenta devido a uma má escolha desta. Na verdade, em alguns casos em que a função de verosimilhança não é limitada superiormente no espaço paramétrico, a

sucessão de estimativas geradas pelo algoritmo EM pode divergir se a solução inicial for escolhida demasiado próximo da fronteira. Como já foi referido o método dos momentos, quando possível, fornece valores iniciais razoáveis.

Outro aspecto a ter em conta é o facto do algoritmo EM não produzir estimativas da matriz de variâncias-covariâncias para os EMV. Contudo, desenvolvimentos subsequentes do artigo DLR providenciam métodos para essa estimação, e que poderão ser integrados no esquema computacional do algoritmo EM (nomeadamente o método de Newton-Raphson).

Seja $\boldsymbol{Y}$ o vector aleatório correspondente às observações $\boldsymbol{y}$, com f.d.p. $g(\boldsymbol{y};\boldsymbol{\theta})$, onde $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^T$ é o vector que contém os parâmetros desconhecidos, sendo o espaço dos parâmetros Ω. Seja então $\boldsymbol{x}=(\boldsymbol{y},\boldsymbol{z})$ o vector que contém, os chamados dados completos com $\boldsymbol{z}$ o vector que contém os dados adicionais, referidos como os dados não observados ou perdidos. Considerando o vector aleatório $\boldsymbol{X}$ correspondente ao vector dos dados completos $\boldsymbol{x}$ com f.d.p. $g_C(\boldsymbol{x}\mid\boldsymbol{\theta})$, então a função de log-verosimilhança dos dados completos (que pode ser formada para $\boldsymbol{\theta}$ se $\boldsymbol{x}$ for totalmente observado) será dada por

$$V_C(\boldsymbol{\theta}\mid\boldsymbol{x})=\ln L_C(\boldsymbol{\theta}\mid\boldsymbol{x})=\ln g_C(\boldsymbol{x}\mid\boldsymbol{\theta}). \tag{3.34}$$

Formalmente, existem dois espaços amostrais X e Y. Em vez de se observar o vector dos dados completos $\boldsymbol{x}$ em X, observa-se o vector dos dados incompletos $\boldsymbol{y}=\boldsymbol{y}(\boldsymbol{x})$ em Y. Resulta que a f.d.p. dos dados incompletos é

$$g(\boldsymbol{y}\mid\boldsymbol{\theta})=\int_{X(\boldsymbol{y})} g_C(\boldsymbol{x}\mid\boldsymbol{\theta})d\boldsymbol{x}, \tag{3.35}$$

onde $X(\boldsymbol{y})$ é o subconjunto de X constituído pelos valores de $\boldsymbol{x}$ para os quais $\boldsymbol{y}=\boldsymbol{y}(\boldsymbol{x})$.

O algoritmo EM aproxima o problema de resolver a equação de log-verosimilhança dos dados incompletos (3.13) indirectamente procedendo iterativamente em termos da função de log-verosimilhança dos dados completos, $V_C(\boldsymbol{\theta}\mid\boldsymbol{x})$. Esta última, como não é observável (não está disponível), substitui-se pela sua esperança condicionada dado $\boldsymbol{y}$, utilizando a actualização corrente para $\boldsymbol{\theta}$.

Mais especificamente, seja $\theta^{(0)}$ um valor inicial, qualquer, para θ. Então na primeira iteração, o passo-E necessita do cálculo de

$$Q(\theta;\theta^{(0)}) = E_{\theta^{(0)}}[V_C(\theta \mid x) \mid y]. \tag{3.36}$$

O passo-M requer a maximização (ou pelo menos, a melhoria) de $Q(\theta;\theta^{(0)})$ com respeito a θ ao longo do espaço de parâmetros Ω. Isto é, escolhe-se $\theta^{(1)}$ de maneira a que $Q(\theta^{(1)};\theta^{(0)}) = \max_{\theta \in \Omega} Q(\theta;\theta^{(0)})$ ou, pelo menos, de modo a que $Q(\theta^{(1)};\theta^{(0)}) \geq Q(\theta^{(0)};\theta^{(0)})$, $\forall_{\theta^{(0)} \in \Omega}$ (algoritmo EM generalizado). Os passos-E e -M são repetidos novamente, contudo desta vez com $\theta^{(0)}$ substituído pela actualização $\theta^{(1)}$. Na iteração de ordem (k+1), estes dois passos são definidos da seguinte maneira:

Passo-E. Calcular $Q(\theta;\theta^{(k)})$, onde $Q(\theta;\theta^{(k)}) = E_{\theta^{(k)}}[V_C(\theta|x) \mid y]$. (3.37)

Passo-M. Escolher $\theta^{(k+1)}$ como sendo um valor qualquer de $\theta \in \Omega$ que maximiza ou majora (conforme seja EM ou EM generalizado) $Q(\theta;\theta^{(k)})$; isto é, garantindo-se que

$$Q(\theta^{(k+1)};\theta^{(k)}) \geq Q(\theta^{(k)};\theta^{(k)}). \tag{3.38}$$

Os passos-E e -M são alternadamente repetidos até que a diferença

$$L(\theta^{(k+1)} \mid y) - L(\theta^{(k)} \mid y) < \varepsilon, \tag{3.39}$$

valor fixado arbitrariamente pequeno. Admite-se a convergência da sequência de valores de verosimilhança $\{L(\theta^{(k)} \mid y)\}$. DLR mostrou que a função de verosimilhança $L(\theta \mid y)$ é não decrescente após uma iteração quer do algoritmo EM quer do algoritmo EM generalizado; isto é,

$$L(\theta^{(k+1)} \mid y) \geq L(\theta^{(k)} \mid y) \tag{3.40}$$

para $k = 0,1,2,\ldots$.

Para se provar esta última desigualdade, seja a f.d.p. condicional de X dado $Y = y$

$$k(x \mid y;\theta) = \frac{g_C(x \mid \theta)}{g(y \mid \theta)} = \frac{L_C(\theta \mid x)}{L(\theta \mid y)}. \tag{3.41}$$

Então a função de log-verosimilhança, $V(\theta \mid y) = \ln L(\theta \mid y)$, é equivalente a

$$V(\theta \mid y) = V_C(\theta \mid x) - \ln k(x \mid y;\theta), \tag{3.42}$$

com $V_C(\theta \mid x) = \ln L_C(\theta \mid x)$, isto é, a função de log-verosimilhança dos dados incompletos é igual à diferença entre a função de log-verosimilhança dos dados completos e o logaritmo da f.d.p. condicional de X dado $Y = y$.

Considerando a esperança matemática, em ambos os lados de (3.42), com respeito à distribuição condicional de X dado $Y = y$, para $\theta = \theta^{(k)}$, vem

$$V(\theta \mid y) = Q(\theta;\theta^{(k)}) - H(\theta;\theta^{(k)}) \tag{3.43}$$

com

$$H(\theta;\theta^{(k)}) = E_{\theta^{(k)}}[\ln k(X \mid y;\theta) \mid y]. \tag{3.44}$$

De (3.43) obtém-se

$$V(\theta^{(k+1)} \mid y) = Q(\theta^{(k+1)};\theta^{(k)}) - H(\theta^{(k+1)};\theta^{(k)}) \text{ e } V(\theta^{(k)} \mid y) = Q(\theta^{(k)};\theta^{(k)}) - H(\theta^{(k)};\theta^{(k)}).$$

Então

$$V(\theta^{(k+1)} \mid y) - V(\theta^{(k)} \mid y) = Q(\theta^{(k+1)};\theta^{(k)}) - Q(\theta^{(k)};\theta^{(k)}) - \left(H(\theta^{(k+1)};\theta^{(k)}) - H(\theta^{(k)};\theta^{(k)})\right), \tag{3.45}$$

concluindo-se que se verifica (3.40) se

$$H(\theta^{(k+1)};\theta^{(k)}) - H(\theta^{(k)};\theta^{(k)}) \leq 0, \tag{3.46}$$

uma vez que $\theta^{(k+1)}$ é escolhido por forma a

$$Q\left(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}\right) \geq Q\left(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}\right). \tag{3.47}$$

Apresenta-se de seguida uma definição e um teorema (bem conhecidos) para se provar (3.46).

Definição 3.4 *Funções convexas.* Uma função $f(x)$ diz-se convexa no intervalo $[a,b]$, se para quaisquer dois pontos $x_1, x_2 \in [a,b]$,

$$f\left(\frac{x_1+x_2}{2}\right) \leq f\left(\frac{x_1}{2}\right) + f\left(\frac{x_2}{2}\right). \tag{3.48}$$

Analogamente, uma função $f(x)$ diz-se côncava no intervalo $[a,b]$, se para quaisquer dois pontos $x_1, x_2 \in [a,b]$ se a função $-f(x)$ for convexa nesse intervalo.

Teorema 3.3 *Desigualdade de Jensen.* Seja $f(x)$ uma função convexa no intervalo $[a,b]$, e seja X uma v.a. de maneira que $P\{X \in (a,b)\} = 1$ e que existam as esperanças $E[X]$ e $E[f(X)]$. Então

$$E[f(X)] \geq f(E[X]). \tag{3.49}$$

Então, para qualquer $\boldsymbol{\theta}$, resulta de (3.44) que

$$\begin{aligned} H\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) - H\left(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}\right) &= E_{\boldsymbol{\theta}^{(k)}}\left[\ln k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y}\right] - E_{\boldsymbol{\theta}^{(k)}}\left[\ln k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) \mid \boldsymbol{y}\right] = \\ &= E_{\boldsymbol{\theta}^{(k)}}\left[\ln \frac{k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})}{k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right)} \mid \boldsymbol{y}\right] \leq \ln E_{\boldsymbol{\theta}^{(k)}}\left[\frac{k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})}{k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right)} \mid \boldsymbol{y}\right] = \\ &= \ln \int_{X(y)} \frac{k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta})}{k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right)} k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) d\boldsymbol{x} = \ln \int_{X(y)} k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}) d\boldsymbol{x} = 0. \end{aligned}$$

A desigualdade nesta última expressão é uma consequência da desigualdade de Jensen e do facto da função logarítmica ser côncava. Ficou provado que a função de verosimilhança é não decrescente depois de uma qualquer iteração do algoritmo EM ou do algoritmo EM generalizado.

Por outro lado a função de verosimilhança é estritamente crescente se

$$Q\left(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}\right) > Q\left(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}\right),\ \forall_{\boldsymbol{\theta}^{(k)}\in\Omega}. \tag{3.50}$$

Nesse caso, para uma sequência de valores da função de verosimilhança $\left\{L\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)\right\}$ limitada superiormente, $L\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)$ converge monotonamente para algum $L\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{y}\right)$. Na maior parte das aplicações, $L\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{y}\right)$ é um ponto de estacionaridade.

Com efeito, derivando ambos os membros de (3.43),

$$\frac{\partial V(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} - \frac{\partial H\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}, \tag{3.51}$$

e como

$$H\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) = E_{\boldsymbol{\theta}^{(k)}}\left[\ln k(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y}\right], \tag{3.52}$$

vem

$$\left[\frac{\partial H\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = E_{\boldsymbol{\theta}^{(k)}}\left[\frac{\partial \ln k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) \mid \boldsymbol{y}}{\partial \boldsymbol{\theta}}\right] = E_{\boldsymbol{\theta}^{(k)}}\left[\frac{\frac{\partial k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) \mid \boldsymbol{y}}{\partial \boldsymbol{\theta}}}{k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) \mid \boldsymbol{y}}\right] =$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{X(y)} \ln k\left(\boldsymbol{X} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}\right) dx = 0$$

(admitindo a possibilidade de troca dos operadores de derivação e de esperança matemática) e, assim,

$$\frac{\partial V\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{y}\right)}{\partial \boldsymbol{\theta}} = \left[\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{*}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*}} = 0, \tag{3.53}$$

pelo que $\boldsymbol{\theta}^{*}$ é um ponto estacionário de $L(\boldsymbol{\theta} \mid \boldsymbol{y})$.

O que quer dizer que, no caso de haver vários pontos de estacionaridade (mínimos ou máximos locais e pontos de sela), o algoritmo pode não convergir para o máximo global (caso este exista). A convergência depende do valor inicial $\boldsymbol{\theta}^{(0)}$. Dado que o algoritmo faz

crescer a verosimilhança, não se atingirá um mínimo local. Também os pontos de sela não serão atingidos pois uma pequena perturbação (nem que seja de arredondamento) no valor dos parâmetros relativamente a este ponto faz com que o algoritmo não estabilize, isto é não convirja. Como o algoritmo estabiliza nos máximos locais, devem-se utilizar vários valores iniciais, tendo em conta o maior valor da função de log-verosimilhança.

A convergência da sequência de valores da função de verosimilhança $\{L(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\}$ para algum $L(\boldsymbol{\theta}^{*} \mid \boldsymbol{y})$ não implica automaticamente a convergência da correspondente sequência de iterações $\{\boldsymbol{\theta}^{(k)}\}$ para $\boldsymbol{\theta}^{*}$. Do ponto de vista numérico, a convergência de $\{\boldsymbol{\theta}^{(k)}\}$ não é tão importante como a convergência de $\{L(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\}$ para um ponto estacionário, Wu (1983).

Caso $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ seja unimodal em Ω e se se verificarem certas condições de diferenciabilidade, qualquer sequência do algoritmo EM converge para o único EMV, qualquer que seja o ponto $\boldsymbol{\theta}^{(0)}$.

Uma consequência de (3.40) é a consistência do algoritmo EM. Se o EMV $\hat{\boldsymbol{\theta}}$ for um máximo global de $L(\boldsymbol{\theta} \mid \boldsymbol{y})$, então deve satisfazer

$$Q(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}) \geq Q(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}) \tag{3.54}$$

para todo o $\boldsymbol{\theta}$. Caso contrário, isto é, caso

$$Q(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}) < Q(\boldsymbol{\theta}_0;\hat{\boldsymbol{\theta}}) \tag{3.55}$$

para algum $\boldsymbol{\theta}_0$, viria que

$$L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) < L(\boldsymbol{\theta}_0 \mid \boldsymbol{y}), \tag{3.56}$$

o que contraria o facto de $\hat{\boldsymbol{\theta}}$ ser o máximo global de $L(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Um estudo mais detalhado sobre convergência do algoritmo EM pode-se encontrar, por exemplo, em McLachlan e Krishnan (1997)

3.3.4.2.1 Matriz de variâncias-covariâncias assintótica

A rapidez da convergência do algoritmo EM depende da fracção de informação perdida. Quando a porção de dados "perdidos" é grande, a convergência do algoritmo pode ser muito lenta. Por outro lado, quando a fracção de dados "perdidos" é pequena, o aumento na variância dos parâmetros ajustados é pequena. A convergência pode ser acelerada combinando o algoritmo EM com o método de Newton-Raphson, visando a convergência quadrática na vizinhança de $\hat{\boldsymbol{\theta}}$. Esta combinação para além de garantir que o máximo global é atingido, permite a obtenção da matriz de variâncias-covariâncias $I^{-1}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})$, necessária à construção dos intervalos de confiança, uma vez que o algoritmo EM não providencia automaticamente uma estimativa desta matriz para o EMV do vector dos parâmetros $\boldsymbol{\theta}$. Comece-se, então, por deduzir $I(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Derivando duas vezes ambos os membros de (3.42) obtém-se

$$I(\boldsymbol{\theta} \mid \boldsymbol{y}) = I_C(\boldsymbol{\theta} \mid \boldsymbol{x}) + \frac{\partial^2 \ln k(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}, \tag{3.57}$$

com

$$I(\boldsymbol{\theta} \mid \boldsymbol{y}) = -\frac{\partial^2 V(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.58}$$

e

$$I_C(\boldsymbol{\theta} \mid \boldsymbol{x}) = -\frac{\partial^2 V_C(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{3.59}$$

Considerando a esperança matemática com respeito à distribuição condicional de $\boldsymbol{x}$ dado $\boldsymbol{y}$ em (3.57), vem

$$I(\boldsymbol{\theta} \mid \boldsymbol{y}) = Inf_C(\boldsymbol{\theta} \mid \boldsymbol{y}) - Inf_m(\boldsymbol{\theta} \mid \boldsymbol{y}), \tag{3.60}$$

ou seja; a informação observada é igual à diferença entre a informação completa (esperança condicionada) e a informação perdida. A igualdade (3.60) é conhecida por "*principio de informação perdida*" (Orchard e Woodbury, 1972), uma vez que

$$Inf_C(\theta \mid y) = E_\theta[I_C(\theta \mid X) \mid y] \tag{3.61}$$

é a esperança condicionada da matriz de informação esperada para os dados completos $I_C(\theta \mid x)$ dado y, e

$$Inf_m(\theta \mid y) = -E_\theta\left[\frac{\partial^2 \ln k(X \mid y;\theta)}{\partial\theta\partial\theta^T} \mid y\right] \tag{3.62}$$

é a matriz de informação esperada para θ baseada em x (ou equivalentemente, nos dado não observados z) condicionada a saber y. Esta pode ser interpretada como a informação "perdida" em consequência de se observar apenas y e não z.

Esta matriz pode ser apresentada como uma diferença envolvendo as funções "score" para dados completos e para dados incompletos. Como

$$I(\theta \mid y) = -\frac{\partial^2 \ln g(y \mid \theta)}{\partial\theta\partial\theta^T} = -\frac{\partial S(\theta \mid y)}{\partial\theta}, \tag{3.63}$$

e

$$\begin{aligned}\frac{\partial^2 \ln g(y \mid \theta)}{\partial\theta\partial\theta^T} &= \frac{\dfrac{\partial^2 g(y \mid \theta)}{\partial\theta\partial\theta^T}}{g(y \mid \theta)} - \frac{\dfrac{\partial g(y \mid \theta)}{\partial\theta}\dfrac{\partial g(y \mid \theta)}{\partial\theta^T}}{g^2(y \mid \theta)} = \\ &= \frac{\dfrac{\partial^2 g(y \mid \theta)}{\partial\theta\partial\theta^T}}{g(y \mid \theta)} + S(\theta \mid y)S^T(\theta \mid y),\end{aligned} \tag{3.64}$$

do segundo membro de (3.64), considerando que se verificam as condições de regularidade de maneira a se poder trocar o integral pela derivada e que $g(y \mid \theta) = \int_{X(y)} g_c(x \mid \theta)dx$, resulta

$$\frac{\partial^2 g(y|\theta)}{\partial\theta\partial\theta^T}\Big/ g(y|\theta) = \int_{X(y)} \frac{\partial^2 g_C(x|\theta)}{\partial\theta\partial\theta^T} dx \Big/ g(y|\theta) =$$

$$= \int_{X(y)} \frac{\partial^2 V_C(\theta|x)}{\partial\theta\partial\theta^T} \frac{g_C(x|\theta)}{g(y|\theta)} dx + \int_{X(y)} \frac{\partial V_C(\theta|x)}{\partial\theta} \frac{\partial V_C(\theta|x)}{\partial\theta^T} \frac{g_C(x|\theta)}{g(y|\theta)} dx = \tag{3.65}$$

$$= -\int_{X(y)} I_C(\theta|x) k(x|y;\theta) dx + \int_{X(y)} S_C(\theta|x) S_C^T(\theta|x) k(x|y;\theta) dx =$$

$$= -E_\theta[I_C(\theta|X)|y] + E_\theta[S_C(\theta|X) S_C^T(\theta|X)|y] = -Inf_C(\theta|y) + E_\theta[S_C(\theta|X) S_C^T(\theta|X)|y]$$

Então, de (3.60) vem

$$I(\theta|y) = Inf_C(\theta|y) - E_\theta[S_C(\theta|X) S_C^T(\theta|X)|y] + S(\theta|y) S^T(\theta|y) \tag{3.66}$$

e, de (3.60) e (3.66),

$$Inf_m(\theta|y) = E_\theta[S_C(\theta|X) S_C^T(\theta|X)|y] - S(\theta|y) S^T(\theta|y).\# \tag{3.67}$$

Louis (1982) mostrou que a matriz de informação perdida (3.67) pode ser escrita da seguinte forma

$$Inf_m(\theta|y) = \mathrm{cov}_\theta[S_C(\theta|X)|y], \tag{3.68}$$

uma vez que

$$S(\theta|y) = E_\theta[S_C(\theta|X)|y]. \tag{3.69}$$

Da expressão (3.42) vem

$$S(\theta|y) = S_C(\theta|x) - \frac{\partial \ln k(x|y;\theta)}{\partial\theta}, \tag{3.70}$$

sendo

$$S(\theta|y) = \frac{\partial V(\theta|x)}{\partial\theta} \tag{3.71}$$

e

$$S_C(\theta \mid y) = \frac{\partial V_C(\theta \mid x)}{\partial \theta}. \tag{3.72}$$

Considerando a esperança matemática com respeito à distribuição condicional de X dado $Y = y$ para θ vem

$$S(\theta \mid y) = E_\theta[S_C(\theta \mid X) \mid y] - E_\theta\left[\frac{\partial \ln k(x \mid y;\theta)}{\partial \theta} \mid y\right] \tag{3.73}$$

e, admitindo que se verificam as condições regulares que permitem trocar as operações de integração e derivação, e que

$$\begin{aligned} E_\theta\left[\frac{\partial \ln k(x \mid y;\theta)}{\partial \theta} \mid y\right] &= \int_{X(y)} \frac{\partial \ln k(x \mid y;\theta)}{\partial \theta} k(x \mid y;\theta)dx = \int_{X(y)} \frac{\partial k(x \mid y;\theta)}{\partial \theta} dx = \\ &= \frac{\partial}{\partial \theta} \int_{X(y)} k(x \mid y;\theta)dx = 0, \end{aligned} \tag{3.74}$$

vem

$$S(\theta \mid y) = E_\theta[S_C(\theta \mid X) \mid y].\# \tag{3.75}$$

De (3.66) resulta que a matriz de informação observada para $\theta = \hat{\theta}$ é

$$I(\hat{\theta} \mid y) = Inf_C(\hat{\theta} \mid y) - E_\theta[S_C(\theta \mid X) S_C^T(\theta \mid X) \mid y]_{\theta=\hat{\theta}}, \tag{3.76}$$

uma vez que

$$S(\hat{\theta} \mid y) = 0.$$

O cálculo desta última expressão para obtenção de $I(\hat{\theta} \mid y)$ envolve esperanças matemáticas que poderão ser difíceis ou até mesmo impossíveis de obter. Assim, considera-se uma aproximação a esta matriz através de

$$I_e(\theta \mid y) = \sum_{j=1}^{n} s(\theta \mid y_j) s^T(\theta \mid y_j) - n^{-1} S(\theta \mid y) S^T(\theta \mid y), \tag{3.77}$$

com $S(\boldsymbol{\theta} \mid \boldsymbol{y})$ e $s(\boldsymbol{\theta} \mid \boldsymbol{y}_j)$ dados por (3.28) e (3.29) respectivamente. A justificação da utilização de $I_e(\boldsymbol{\theta} \mid \boldsymbol{y})$ como aproximação a $I(\boldsymbol{\theta} \mid \boldsymbol{y})$ foi feita na secção 3.3.4.1.2.

Como se pode reparar, esta aproximação evita o cálculo das derivadas parciais de 2ª ordem. A inversa de $I_e(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) = \sum_{j=1}^{n} s(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}_j) s^T(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}_j)$ dá uma aproximação à matriz de variâncias-covariâncias do EMV de $\boldsymbol{\theta}$, sendo a variância dos componentes de $\boldsymbol{\theta}$ aproximada pelos elementos da diagonal principal. $I_e(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})$ providencia uma aproximação de $I(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})$, que permite utilizar, por exemplo, o método de Newton-Rapshon para calcular o EMV, isto é

$$\boldsymbol{\theta}^{(k+1)} \approx \boldsymbol{\theta}^{(k)} + I_e^{-1}(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}) S(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)}), \tag{3.78}$$

Prova-se que $\frac{I_e(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})}{n}$ é um estimador consistente da matriz de informação esperada (Mclachlan e Krishnan, 1997)..

3.3.4.2.2 Aplicação a um modelo de misturas de distribuições normais

Como foi referido, na secção 3.3.3.6, os EMV para modelos de mistura não têm expressões explícitas, devendo-se, por isso, aplicar um método iterativo para a obtenção das suas estimativas. A publicação do artigo DLR estimulou o interesse para a utilização de distribuições de misturas finitas, uma vez que, quando considerado como um problema de dados incompletos, o ajuste deste tipo de modelos é bastante simplificado pelo algoritmo EM.

Uma vasta variedade de aplicações sobre modelos de misturas finitas são dados em McLachlan e Basford (1988) e McLachlan (1997).

Supondo que a f.d.p. de uma v.a. unidimensional Y é uma mistura-φ em localização de distribuições normais com o mesmo desvio padrão $\delta > 0$ dada por

$$f(y \mid \mathbf{\Pi}) = \sum_{i=1}^{d} \frac{\varphi_i}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\lambda_i}{\delta}\right)^2}, \tag{3.79}$$

onde d é o número de componentes da mistura, $\varphi_d = 1 - \sum_{i=1}^{d-1} \varphi_i$, $\lambda_i \in IR$ e $\mathbf{\Pi} = (\varphi_1, \ldots, \varphi_{d-1}, \lambda_1, \ldots, \lambda_d, \delta)$. Seja

$$y = (y_1, \ldots, y_n)^T$$

uma amostra aleatória observada da v.a. $Y = (Y_1, \ldots, Y_n)$, cada com f.d.p. $f(y \mid \mathbf{\Pi})$.

Note-se que a estimação de $\mathbf{\Pi}$ com base em y só faz sentido se a mistura for identificável; isto é, se diferentes valores de $\mathbf{\Pi}$ determinarem membros distintos da família

$$\{f(y \mid \mathbf{\Pi}) : \mathbf{\Pi} \in \Omega\}, \tag{3.80}$$

onde Ω é o espaço dos parâmetros, o que é verdade para misturas de normais uma vez que se pode determinar $\mathbf{\Pi}$ através de uma permutação das componentes. Por exemplo, para o caso particular de uma mistura de duas normais, não se distingue $(\varphi_1, \lambda_1, \lambda_2, \delta^2)^T$ de $(\varphi_2, \lambda_2, \lambda_1, \delta^2)^T$; contudo esta falta de identificabilidade não tem reflexos em termos práticos, pois pode ser facilmente ultrapassado pela restrição $\varphi_1 \leq \varphi_2$ (ver McLachlan e Basford, 1988, secção 1.5).

A função de log-verosimilhança para $\mathbf{\Pi}$ que pode ser formada a partir das observações y é dada por

$$V(\mathbf{\Pi} \mid y) = \sum_{j=1}^{n} \ln\left\{\sum_{i=1}^{d}\left(\frac{\varphi_i}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j-\lambda_i}{\delta}\right)^2}\right)\right\}. \tag{3.81}$$

Tratando este problema como um problema de dados incompletos, introduz-se o vector dos dados perdidos (não observados)

$$z = \left(z_1^T, \ldots, z_n^T\right)^T \tag{3.82}$$

onde z_j é um vector d-dimensional, $z_j = \left(z_{1j}, \ldots, z_{dj}\right)^T$ com

$$z_{ij} = \left(z_j\right)_i = \begin{cases} 1 & \text{se } y_j \text{ provém da i - ésima componente da mistura} \\ 0 & \text{caso contrário} \end{cases}$$

$\left(i = 1, \ldots d; j = 1, \ldots, n\right)$.

Se os z_{ij} forem observados, então os EMV para φ_i são

$$\sum_{j=1}^{n} \frac{z_{ij}}{n}, \tag{3.83}$$

o que não é mais do que a proporção da amostra que provém da i-ésima componente da mistura.

Seja o vector dos dados completos dados por

$$x = \left(y^T, z^T\right)^T. \tag{3.84}$$

A função de verosimilhança dos dados completos para θ tem a forma multinomial

$$L_C\left(\Pi \mid x\right) = \prod_{j=1}^{n} \prod_{i=1}^{d} \left(\frac{\varphi_i}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j - \lambda_i}{\delta}\right)^2} \right)^{z_{ij}} \tag{3.85}$$

a correspondente função de log-verosimilhança será

$$V_C\left(\Pi \mid x\right) = \sum_{j=1}^{n} \sum_{i=1}^{d} z_{ij} \ln \varphi_i + \sum_{j=1}^{n} \sum_{i=1}^{d} z_{ij} \ln \left\{ \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_j - \lambda_i}{\delta}\right)^2} \right\} \tag{3.86}$$

ou

$$V_C(\boldsymbol{\Pi} \mid \boldsymbol{x}) = \sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij} \ln \varphi_i - \frac{1}{2} n \ln\left(2\pi\delta^2\right) - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij}\left(\frac{y_j - \lambda_i}{\delta}\right)^2, \quad (3.87)$$

uma vez que

$$\sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij} = n. \quad (3.88)$$

Passo-E. Calcular $Q\left(\boldsymbol{\Pi}; \boldsymbol{\Pi}^{(k)}\right) = E_{\boldsymbol{\theta}^{(k)}}\left[V_C(\boldsymbol{\Pi} \mid \boldsymbol{x}) \mid \boldsymbol{y}\right]$.

Como (3.87) é linear em relação aos dados não observados z_{ij}, o passo-E (na iteração $k+1$) requer, apenas, o cálculo da esperança condicional de Z_{ij} dadas as observações $\boldsymbol{y}$, onde Z_{ij} é a v.a. correspondente a z_{ij}. Vem

$$Q\left(\boldsymbol{\Pi}; \boldsymbol{\Pi}^{(k)}\right) = -\frac{1}{2} n \ln\left(2\pi\delta^2\right) + \sum_{j=1}^{n}\sum_{i=1}^{d} \ln \varphi_i E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij} \mid \boldsymbol{y}\right] - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{d}\left(\frac{y_j - \lambda_i}{\delta}\right)^2 E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij} \mid \boldsymbol{y}\right].$$

Seja

$$z_{ij}^{(k)} = E_{\boldsymbol{\theta}^{(k)}}\left[Z_{ij} \mid \boldsymbol{y}\right] = P_{\boldsymbol{\theta}^{(k)}}\left\{Z_{ij} = 1 \mid \boldsymbol{y}\right\} = \frac{\varphi_i^{(k)} e^{-\frac{1}{2}\left(\frac{y_j - \lambda_i^{(k)}}{\delta^{(k)}}\right)^2}}{\sum_{i=1}^{d} \varphi_i^{(k)} e^{-\frac{1}{2}\left(\frac{y_j - \lambda_i^{(k)}}{\delta^{(k)}}\right)^2}} \quad (3.89)$$

para $(j = 1, \ldots, n)$. Esta quantidade é a probabilidade de que o j-ésimo elemento da amostra com valor observado y_j pertença à i-ésima componente da mistura (com estimativas actualizadas dos membros das componentes das misturas).

Resulta

$$Q\left(\boldsymbol{\Pi}; \boldsymbol{\Pi}^{(k)}\right) = -\frac{n}{2} \ln\left(2\pi\delta^2\right) + \sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij}^{(k)} \ln \varphi_i - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij}^{(k)}\left(\frac{y_j - \lambda_i}{\delta}\right)^2. \quad (3.90)$$

Passo-M. Na iteração ordem $(k+1)$ substitui-se cada z_{ij} pela sua esperança condicional actualizada $z_{ij}^{(k)}$ em (3.83), obtendo-se

$$\varphi_i^{(k+1)} = \sum_{j=1}^{n} \frac{z_{ij}^{(k)}}{n} \tag{3.91}$$

para $i = 1,\ldots,d$. Como é lógico, maximizando $Q\left(\boldsymbol{\Pi};\boldsymbol{\Pi}^{(k)}\right)$ em ordem aos φ_i (com a condição $\varphi_1 + \ldots + \varphi_d = 1$) obtém-se (3.91).

Derivando $Q\left(\boldsymbol{\Pi};\boldsymbol{\Pi}^{(k)}\right)$ em ordem aos restantes parâmetros obtém-se

$$\frac{\partial Q\left(\boldsymbol{\Pi};\boldsymbol{\Pi}^{(k)}\right)}{\partial \lambda_i} = \frac{1}{\delta^2}\sum_{j=1}^{n} z_{ij}^{(k)}\left(y_j - \lambda_i\right) \tag{3.92}$$

e

$$\frac{\partial Q\left(\boldsymbol{\Pi};\boldsymbol{\Pi}^{(k)}\right)}{\partial \delta^2} = -\frac{n}{2\delta^2} + \frac{1}{2\delta^4}\sum_{j=1}^{n}\sum_{i=1}^{d} z_{ij}^{(k)}\left(y_j - \lambda_i\right)^2, \tag{3.93}$$

donde

$$\lambda_i^{(k+1)} = \frac{\sum_{j=1}^{n} y_j z_{ij}^{(k)}}{\sum_{i=1}^{n} z_{ij}^{(k)}} \tag{3.94}$$

e

$$\delta^{2(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{d}\left(y_j - \lambda_i^{(k+1)}\right)^2 z_{ij}^{(k)} \ . \tag{3.95}$$

Para a estimação da matriz de variâncias-covariâncias, $I\left(\boldsymbol{\Pi} \mid \boldsymbol{y}\right)$ pode ser aproximada por $I_e\left(\boldsymbol{\Pi} \mid \boldsymbol{y}\right)$.

Na procura do máximo global, como as componentes da mistura pertencem à mesma família paramétrica, a função de log-verosimilhança terá *d* máximos locais do mesmo valor correspondendo a uma troca dos valores das componentes. Este problema de falta de identificabilidade pode ser contornado através de restrições sobre os parâmetros $\hat{\boldsymbol{\Pi}}$ (no

caso de $d = 2$ foram vistas em cima). O algoritmo EM deve ser aplicado com base numa cuidada escolha de soluções iniciais na busca de todos os máximos locais. Uma maneira de se contornar esse problema e garantir que o máximo global é atingindo será repetir o processo iterativo para várias soluções iniciais ou aplicar o algoritmo EM com Newton-Raphson (substituindo $I(\boldsymbol{\Pi} \mid \boldsymbol{y})$ por $I_e(\boldsymbol{\Pi} \mid \boldsymbol{y})$). Em ambas as situações tendo como solução inicial, quando possível, as estimativas obtidas pelo método dos momentos.

No caso de misturas de distribuições normais univariadas, o estimador dos parâmetros $\hat{\boldsymbol{\Pi}}$ correspondente ao máximo local é consistente e eficiente (McLachlan, 1988).

3.3.4.2.3 Algoritmo EM para dados agrupados

Seja W uma v.a. com f.d.p. $f(w \mid \boldsymbol{\theta})$, onde $\boldsymbol{\theta}$ é o vector dos parâmetros desconhecidos. Suponha-se que o espaço amostral W de W esta dividido em v classes mutuamente exclusivas $\mathsf{W}_j = [w_{j-1}, w_j[\quad (j = 1,\ldots,v)$, com $w_o = -\infty$ e $w_v = +\infty$. As observações, independentes, são registadas nas diferentes classes, isto é, as observações individuais não são registadas, mas sim as classes a que estas pertencem, as classes W_j.

Se $n = \sum_{j=1}^{v} n_j$, for o tamanho da amostra, com n_j o número de observações na classe W_j, o vector das observações $\boldsymbol{y} = (n_1,\ldots,n_v)^T$ (dados incompletos) tem distribuição multinomial, consistindo em n extracções de v classes com probabilidades $P_j(\boldsymbol{\theta})$ $(j = 1,\ldots,v)$, a probabilidade de um elemento da amostra pertencer a W_j, é dada por

$$P_j(\boldsymbol{\theta}) = P[W \in \mathsf{W}_j] = \int_{\mathsf{W}_j} f(w \mid \boldsymbol{\theta}) dw. \tag{3.96}$$

Claro que

$$\sum_{j=1}^{v} P_j(\boldsymbol{\theta}) = 1.$$

Assim, a função de verosimilhança para as observações (densidade de probabilidade de $\boldsymbol{y}$) é

$$L(\theta \mid y) = \binom{n}{n_1 \ldots n_v} P_1^{n_1}(\theta) P_2^{n_2}(\theta) \ldots P_v^{n_v}(\theta), \tag{3.97}$$

onde $\binom{n}{n_1 \ldots n_v} = \frac{n!}{n_1! n_2! \ldots n_v!}$, e a função de log-verosimilhança é

$$V(\theta \mid y) = \sum_{j=1}^{v} n_j \ln P_j(\theta) + C_1, \tag{3.98}$$

com $C_1 = \ln\left(\frac{n!}{\prod_{j=1}^{v} n_j!}\right)$.

Das n_j observações que caíram na classe W_j, sejam $w_{j1}, \ldots, w_{jn_j}$ os seus verdadeiros valores. Então se $w_j = (w_{j1}, \ldots, w_{jn_j})^T$ $(j = 1, \ldots, v)$ for considerado como o vector que contém os dados "perdidos" da classe W_j, $x = (y^T, w_1^T, \ldots, w_v^T)^T$, é o vector dos dados completos (dados conhecidos e desconhecidos). Aqui $z = (w_1^T, \ldots, w_v^T)^T$.

A função de verosimilhança dos dados completos (densidade de probabilidade de x) é

$$L_C(\theta \mid x) = \prod_{j=1}^{v} \prod_{l=1}^{n_j} f(w_{jl} \mid \theta), \tag{3.99}$$

donde a função de log-verosimilhança dos dados completos é

$$V_C(\theta \mid x) = \sum_{j=1}^{v} \sum_{l=1}^{n_j} \ln f(w_{jl} \mid \theta). \tag{3.100}$$

As observações w_{ji} $(i = 1, \ldots, n_j)$ são n_j observações individuais com f.d.p.

$$h_j(w \mid \theta) = \frac{f(w \mid \theta)}{P_j(\theta)}, \ (j = 1, \ldots, v), \tag{3.101}$$

pelo que a função de log-verosimilhança dos dados completos é

$$V_C(\boldsymbol{\theta} \mid \boldsymbol{x}) = V(\boldsymbol{\theta} \mid \boldsymbol{y}) + \sum_{j=1}^{v} \sum_{l=1}^{n_j} \ln h_j(w_{jl} \mid \boldsymbol{\theta}) - C_1 . \tag{3.102}$$

Logo, à parte a constante C_1 (que não tem influência para efeitos de maximização), a função de log-verosimilhança completa pode considerar-se como a soma da função de log-verosimilhança dos dados observados com a função de log-verosimilhança dos dados "perdidos".

Passo-E. Calcular

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}[V_C(\boldsymbol{\theta} \mid \boldsymbol{x}) \mid \boldsymbol{y}]. \tag{3.103}$$

Como

$$E_{\boldsymbol{\theta}^{(k)}}[V_C(\boldsymbol{\theta} \mid \boldsymbol{x}) \mid \boldsymbol{y}] = E_{\boldsymbol{\theta}^{(k)}}\left[\sum_{j=1}^{v} \sum_{l=1}^{n_j} \ln f(w_{jl} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}\right] = \sum_{j=1}^{v} \sum_{l=1}^{n_j} E_{\boldsymbol{\theta}^{(k)}}[\ln f(w_{jl} \mid \boldsymbol{\theta}) \mid \boldsymbol{y}], \tag{3.104}$$

e uma vez que esta esperança não varia dentro da classe W_j e, nessa classe, condicionar em $\boldsymbol{y}$ equivale a condicionar $\boldsymbol{w}$ a estar em W_j, vem

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \sum_{j=1}^{v} n_j E_{\boldsymbol{\theta}^{(k)}}[\ln f(W \mid \boldsymbol{\theta}) \mid W \in \mathsf{W}_j] = \sum_{j=1}^{v} n_j Q_j(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \tag{3.105}$$

com

$$Q_j(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}[\ln f(W \mid \boldsymbol{\theta}) \mid W \in \mathsf{W}_j] = \frac{\int_{\mathsf{W}_j} \ln f(w \mid \boldsymbol{\theta}) f(w \mid \boldsymbol{\theta}^{(k)}) dw}{\int_{\mathsf{W}_j} f(w \mid \boldsymbol{\theta}^{(k)}) dw} = \frac{\int_{\mathsf{W}_j} \ln f(w \mid \boldsymbol{\theta}) f(w \mid \boldsymbol{\theta}^{(k)}) dw}{P_j(\boldsymbol{\theta}^{(k)})}$$

A aproximação habitual de considerar o ponto médio $\overline{w}$ de cada classe W_j e utilizar $n_j \ln f(\overline{w} \mid \boldsymbol{\theta})$ é aqui evitada utilizando-se antes o valor esperado corrente de $\ln f(w \mid \boldsymbol{\theta})$ na condição de $w \in \mathsf{W}_j$.

Passo-M. $\boldsymbol{\theta}^{(k+1)}$ é o valor que maximiza $Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)$, ou seja, em condições de regularidade, é o valor de $\boldsymbol{\theta}$ que verifica o sistema de equações

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{v} n_j \frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} = 0. \tag{3.106}$$

Note-se que, admitindo poder trocar o operador derivação com o de esperança matemática,

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{v} n_j E_{\boldsymbol{\theta}^{(k)}}\left[\frac{\partial \ln\left(f\left(W \mid \boldsymbol{\theta}\right)\right)}{\partial \boldsymbol{\theta}} \mid W \in \mathsf{W}_j\right]. \tag{3.107}$$

Uma vez que

$$\frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} = E_{\boldsymbol{\theta}^{(k)}}\left[\frac{\partial \ln\left(f\left(W \mid \boldsymbol{\theta}\right)\right)}{\partial \boldsymbol{\theta}} \mid W \in \mathsf{W}_j\right], \tag{3.108}$$

resulta

$$\left[\frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = \frac{\int_{\mathsf{W}_j} \frac{\partial f\left(w \mid \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} dw}{P_j\left(\boldsymbol{\theta}^{(k)}\right)} = \frac{\frac{\partial P_j\left(\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}}{P_j\left(\boldsymbol{\theta}^{(k)}\right)} = \left[\frac{\partial \ln P_j\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}. \tag{3.109}$$

De (3.105) e (3.109), vem

$$\left[\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = \sum_{j=1}^{v} n_j \frac{\partial \ln P_j\left(\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}, \tag{3.110}$$

que coincide com a expressão de $\mathrm{S}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$ obtida derivando $V\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$ (dada por (3.98)) em ordem a $\boldsymbol{\theta}$ no ponto $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$.

Logo a estatística "*score*" dos dados incompletos é dada por

$$S\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right) = \left[\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}. \#$$

3.3.4.2.3.1 Matriz de variâncias-covariâncias assintótica

Correspondendo a (3.77), pode-se aproximar a matriz de informação observada para dados i.i.d. agrupados, $I\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)$, por

$$I_{e,g}\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)=\sum_{j=1}^{v} n_j s_j\left(\boldsymbol{\theta}^{(k)}\right) s_j^T\left(\boldsymbol{\theta}^{(k)}\right)-n\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)\bar{s}^T\left(\boldsymbol{\theta}^{(k)}\right) \tag{3.111}$$

com

$$s_j\left(\boldsymbol{\theta}^{(k)}\right)=\frac{\partial \ln P_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}=\left.\frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}=\left[\begin{array}{c}\dfrac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \theta_p}\end{array}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} \tag{3.112}$$

a estatística "*score*" dos dados incompletos para a *j-ésima* classe, W_j (que corresponde a $s\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{w}_j\right)$, a estatística "*score*" para a observação $\boldsymbol{w}_j$, no caso dos dados não agrupados) e

$$\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)=\frac{1}{n}\sum_{j=1}^{v} n_j s_j\left(\boldsymbol{\theta}^{(k)}\right). \tag{3.113}$$

Ficando

$$I_{e,g}\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)=\sum_{j=1}^{v} n_j\left[s_j\left(\boldsymbol{\theta}^{(k)}\right)-\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)\right]\left[s_j\left(\boldsymbol{\theta}^{(k)}\right)-\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)\right]^T \tag{3.114}$$

matriz definida positiva.

A utilização desta última expressão como uma aproximação à matriz de informação observada está desenvolvida, por exemplo, em Mclachlan e Krishnan (1997 Pag. 124). Tal como na secção anterior, enquanto $I_{e,g}^{-1}\left(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}\right)$ dá uma aproximação à matriz de variâncias-covariâncias do EMV $\hat{\boldsymbol{\theta}}$, $I_{e,g}^{-1}\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)$ pode ser utilizada para a aplicação do método de Newton-Rapshon,

$$\boldsymbol{\theta}^{(k+1)} \approx \boldsymbol{\theta}^{(k)}+I_{e,g}^{-1}\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right) S\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right), \tag{3.115}$$

onde

$$S(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{j=1}^{v} n_j s_j(\boldsymbol{\theta}), \tag{3.116}$$

é a estatística "score" dos dados incompletos.

Prova-se que $\frac{I_{e,g}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})}{n}$ é um estimador consistente da matriz de informação esperada para uma observação (Mclachlan e Krishnan, 1997).

3.3.4.2.3.2 Aplicação à distribuição normal

Seja W uma v.a. que segue uma distribuição normal com média μ e variância σ^2. Fazendo $\boldsymbol{\theta} = \begin{bmatrix} \mu \\ S \end{bmatrix}$, com $S = \sigma^2$, a f.d.p. é,

$$f(w \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi S}} e^{-\frac{(w-\mu)^2}{2S}}.$$

O espaço amostral W de W, está dividido em v classes mutuamente exclusivas, W_j,

$$\mathsf{W}_j = [w_{j-1}, w_j[\ , \ j = 1,\ldots,v$$

onde $w_0 = -\infty$ e $w_v = \infty$.

O objectivo será estimar μ e S (consequentemente σ^2) com base nas frequências observadas,

$$\boldsymbol{y} = (n_1,\ldots,n_v)^T,$$

onde n_j representa o número de observações de W pertencente à classe W_j $(j = 1,\ldots,v)$ e $n = \sum_{j=1}^{v} n_j$. O vector dos dados completos $\boldsymbol{x}$ é dados por

$$x = \left(y^T, w_1^T, \ldots, w_v^T\right)^T$$

onde

$$w_j = \left(w_{j1}, \ldots, w_{jn_j}\right)^T$$

contém as n_j observações individuais não observadas (dados desconhecidos) de W que pertencem à classe W_j $(j = 1, \ldots, v)$.

A função de log-verosimilhança dos dados completos é dada por

$$V_C(\theta \mid x) = \sum_{j=1}^{v}\sum_{l=1}^{n_j} \ln f\left(w_{jl} \mid \theta\right) = -\frac{n}{2}\ln 2\pi S - \sum_{j=1}^{v}\sum_{l=!}^{n_j} \frac{\left(w_{jl} - \mu\right)^2}{2S}$$

Passo-E. Calcular

$$Q\left(\theta; \theta^{(k)}\right) = \sum_{j=1}^{v} n_j Q_j\left(\theta; \theta^{(k)}\right) = \sum_{j=1}^{v} n_j E_{\theta^{(k)}}\left[\ln f(W \mid \theta) \mid W \in \mathsf{W}_j\right]$$

onde

$$E_{\theta^{(k)}}\left[\ln f(W \mid \theta) \mid W \in \mathsf{W}_j\right] = \frac{\int_{\mathsf{W}_j} \ln f\left(w \mid \theta\right) f\left(w \mid \theta^{(k)}\right) dw}{P_j\left(\theta^{(k)}\right)}.$$

Como

$$P_j\left(\theta^{(k)}\right) = \int_{\mathsf{W}_j} f\left(w \mid \theta^{(k)}\right) \mathrm{d}w = \Phi\left(z_j\left(\theta^{(k)}\right)\right) - \Phi\left(z_{j-1}\left(\theta^{(k)}\right)\right),$$

com

$$z_j\left(\theta^{(k)}\right) = \frac{w_j - \mu^{(k)}}{\sqrt{S^{(k)}}},$$

vem

$$\int_{\mathsf{W}_j} \left(\ln f\left(w \mid \theta\right)\right) f\left(w \mid \theta^{(k)}\right) \mathrm{d}w = \int_{\mathsf{W}_j} \left(-\frac{1}{2}\ln 2\pi S - \frac{1}{2S}\left(w - \mu\right)^2\right) \frac{1}{\sqrt{2\pi S^{(k)}}} e^{-\frac{\left(w-\mu^{(k)}\right)^2}{2S^{(k)}}} \mathrm{d}w =$$

$$= \left(-\frac{1}{2}\ln 2\pi S - \frac{\left(\mu^{(k)} - \mu\right)^2}{2S} \right) P_j\left(\theta^{(k)}\right) + \frac{\sqrt{S^{(k)}}}{S\sqrt{2\pi}}\left(\mu^{(k)} - \mu\right)\left(e^{-\frac{z_j^2\left(\theta^{(k)}\right)}{2}} - e^{-\frac{z_{j-1}^2\left(\theta^{(k)}\right)}{2}} \right) +$$

$$+ \frac{S^{(k)}}{2S\sqrt{2\pi}}\left(z_j\left(\theta^{(k)}\right) e^{-\frac{z_j^2\left(\theta^{(k)}\right)}{2}} - z_{j-1}\left(\theta^{(k)}\right) e^{-\frac{z_{j-1}^2\left(\theta^{(k)}\right)}{2}} \right) - \frac{S^{(k)}}{2S} P_j\left(\theta^{(k)}\right)$$

e assim

$$Q_j\left(\theta;\theta^{(k)}\right) = -\frac{1}{2}\ln 2\pi S - \frac{\left(\mu^{(k)} - \mu\right)^2}{2S} - \frac{S^{(k)}}{2S} + \frac{1}{S}\left[\left(\mu^{(k)} - \mu\right)\sqrt{S^{(k)}} A_j\left(\theta^{(k)}\right) + \frac{S^{(k)}}{2} B_j\left(\theta^{(k)}\right) \right],$$

donde

$$Q\left(\theta;\theta^{(k)}\right) = -\frac{n}{2}\left(\ln 2\pi S + \frac{1}{S}\left(\left(\mu^{(k)} - \mu\right)^2 + S^{(k)} \right) \right) + \frac{1}{S}\sum_{j=1}^{v} n_j \left[\left(\mu^{(k)} - \mu\right)\sqrt{S^{(k)}} A_j\left(\theta^{(k)}\right) + \frac{S^{(k)}}{2} B_j\left(\theta^{(k)}\right) \right]$$

com

$$A_j\left(\theta^{(k)}\right) = \frac{1}{\sqrt{2\pi} P_j\left(\theta^{(k)}\right)}\left(e^{-\frac{z_j^2\left(\theta^{(k)}\right)}{2}} - e^{-\frac{z_{j-1}^2\left(\theta^{(k)}\right)}{2}} \right)$$

e

$$B_j\left(\theta^{(k)}\right) = \frac{1}{\sqrt{2\pi} P_j\left(\theta^{(k)}\right)}\left(z_j\left(\theta^{(k)}\right) e^{-\frac{z_j^2\left(\theta^{(k)}\right)}{2}} - z_{j-1}\left(\theta^{(k)}\right) e^{-\frac{z_{j-1}^2\left(\theta^{(k)}\right)}{2}} \right).$$

Passo-M. Derivando em ordem a cada um dos parâmetros, obtém-se

$$\frac{\partial Q_j\left(\theta;\theta^{(k)}\right)}{\partial \mu} = \frac{\mu^{(k)} - \mu}{S} - \frac{\sqrt{S^{(k)}}}{S} A_j\left(\theta^{(k)}\right)$$

e

$$\frac{\partial Q_j\left(\theta;\theta^{(k)}\right)}{\partial S} = \frac{\left(\mu^{(k)} - \mu\right)^2}{2S^2} - \frac{1}{2S} + \frac{S^{(k)}}{2S^2} - \frac{1}{S^2}\left[\left(\mu^{(k)} - \mu\right)\sqrt{S^{(k)}} A_j\left(\theta^{(k)}\right) + \frac{S^{(k)}}{2} B_j\left(\theta^{(k)}\right) \right].$$

Resolvendo as equações que se seguem, obtêm-se as expressões que permitem fazer "correr" o algoritmo EM,

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \mu}=0 \Rightarrow \sum_{j=1}^{v} n_j \frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \mu}=0 \Leftrightarrow \mu^{(k+1)}=\mu^{(k)}-\sqrt{S^{(k)}}A\left(\boldsymbol{\theta}^{(k)}\right)$$

e

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial S}=0 \Rightarrow \sum_{j=1}^{v} n_j \frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial S}=0 \Leftrightarrow S^{(k+1)}=S^{(k)}-S^{(k)}A^2\left(\boldsymbol{\theta}^{(k)}\right)-S^{(k)}B\left(\boldsymbol{\theta}^{(k)}\right)$$

onde $A\left(\boldsymbol{\theta}^{(k)}\right)=\frac{1}{n}\sum_{j=1}^{v} n_j A_j\left(\boldsymbol{\theta}^{(k)}\right)$ e $B\left(\boldsymbol{\theta}^{(k)}\right)=\frac{1}{n}\sum_{j=1}^{v} n_j B_j\left(\boldsymbol{\theta}^{(k)}\right)$,

o que completa a implementação da iteração de ordem $(k+1)$ do algoritmo EM. Uma vez obtido o estimador $\hat{\boldsymbol{\theta}}$, pode-se aproximar a matriz de variâncias-covariâncias assintótica por $I_{e,g}^{-1}\left(\boldsymbol{\theta}^{(k)};\boldsymbol{y}\right)$, a matriz de informação empírica para dados agrupados.

Considera-se agora a aplicação do método Newton-Raphson a este exemplo. De (3.116), a estatística "score" dos dados incompletos para $\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}$ é dada por

$$S\left(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y}\right)=\sum_{j=1}^{v} n_j s_j\left(\boldsymbol{\theta}^{(k)}\right)$$

onde

$$s_j\left(\boldsymbol{\theta}^{(k)}\right)=\left[\begin{array}{c}\frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \mu} \\ \frac{\partial Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial s}\end{array}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}=\left[\begin{array}{c}-\frac{1}{\sqrt{S^{(k)}}}A_j\left(\boldsymbol{\theta}^{(k)}\right) \\ -\frac{1}{2S^{(k)}}B_j\left(\boldsymbol{\theta}^{(k)}\right)\end{array}\right]$$

e uma vez que a matriz de informação empírica é

$$I_{e,g}\left(\boldsymbol{\theta}^{(k)};\boldsymbol{y}\right)=\sum_{j=1}^{v} n_j\left[s_j\left(\boldsymbol{\theta}^{(k)}\right)-\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)\right]\left[s_j\left(\boldsymbol{\theta}^{(k)}\right)-\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)\right]^T,$$

onde

$$\bar{s}\left(\theta^{(k)}\right)=\frac{1}{n}\sum_{j=1}^{v}n_j s_j\left(\theta^{(k)}\right)=\begin{bmatrix}-\dfrac{1}{\sqrt{S^{(k)}}}A\left(\theta^{(k)}\right)\\ -\dfrac{1}{2S^{(k)}}B\left(\theta^{(k)}\right)\end{bmatrix}$$

segue

$$I_{e,g}\left(\theta^{(k)};y\right)=\begin{bmatrix}\dfrac{1}{S^{(k)}}\left[nA^2\left(\theta^{(k)}\right)-\sum_{j=1}^{v}n_j A_j^2\left(\theta^{(k)}\right)\right] & -\dfrac{1}{2\sqrt{\left(S^{(k)}\right)^3}}\left[nA\left(\theta^{(k)}\right)B\left(\theta^{(k)}\right)-\sum_{j=1}^{v}n_j A_j\left(\theta^{(k)}\right)B_j\left(\theta^{(k)}\right)\right]\\ -\dfrac{1}{2\sqrt{\left(S^{(k)}\right)^3}}\left[nA\left(\theta^{(k)}\right)B\left(\theta^{(k)}\right)-\sum_{j=1}^{v}n_j A_j\left(\theta^{(k)}\right)B_j\left(\theta^{(k)}\right)\right] & \dfrac{1}{4\left(S^{(k)}\right)^2}\left[nB^2\left(\theta^{(k)}\right)-\sum_{j=1}^{v}n_j B_j^2\left(\theta^{(k)}\right)\right]\end{bmatrix}$$

podendo-se assim aplicar o algoritmo iterativo,

$$\theta^{(k+1)}\approx\theta^{(k)}+I_{e,g}^{-1}\left(\theta^{(k)};y\right)S\left(\theta^{(k)}\mid y\right).$$

3.3.4.3 Algoritmos iterativos por simulação

Os problemas de dados incompletos conduzem, muitas vezes, a funções de verosimilhança complicadas envolvendo integrais difíceis, se não mesmo impossíveis de calcular. Por outro lado, as derivadas, necessárias para se chegar ao EMV, poderão ser intratáveis.

Como já foi referido, o algoritmo EM é um processo iterativo para se maximizar a verosimilhança. Contudo, em muitas situações, particularmente quando as f.d.p. apresentam vários parâmetros a estimar, a esperança condicional necessária no passo-E do algoritmo EM pode envolver integração numérica. Tal leva a tentar estimar esta esperança por simulação. É o que acontece no algoritmo EM estocástico, sugerido por Celeux, Diebolt (1985 ou 1986) – ver Diebolt e Ip (1996) para uma revisão recente e outras referências – e no algoritmo EM de Monte Carlo (Wei e Tanner 1990). A principal diferença entre estes algoritmos é que, enquanto no algoritmo EM estocástico se utiliza apenas uma simulação em cada iteração, o algoritmo EM de Monte Carlo envolve várias simulações para obter uma boa estimativa da esperança condicionada (pelo menos nas últimas iterações do algoritmo).

Nesta secção, faz-se uma breve referência a esses dois métodos.

3.3.4.3.1 Algoritmo EM estocástico

Em algumas aplicações do algoritmo EM, o passo-E é complexo e pode não admitir uma solução explícita para a computação da esperança condicionada da função de log-verosimilhança dos dados completos, isto é, a função $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$. Uma maneira de ultrapassar esta situação será passar à integração numérica ou, como alternativa, executar o passo-E por simulação. Broniatowski, Celeux e Diebolt (1983) e Celeux e Diebolt (1985, 1986a, 1986b), visando a "computação" de EMV para modelos de misturas finitas, consideraram uma versão modificada do algoritmo EM, a que deram o nome de algoritmo EM estocástico (EME). Claro que o método pode ser aplicado a qualquer algoritmo EM, quer se refira a misturas ou não.

Seja $\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T$, o vector dos dados completos, onde $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ é o vector das observações, sendo o vector dos dados "perdidos" $\boldsymbol{z}$.

A função-Q que deverá ser calculada no passo-E do algoritmo EM pode ser expressa por

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) &= E_{\boldsymbol{\theta}^{(k)}}[\ln L_C(\boldsymbol{\theta} \mid \boldsymbol{x}) \mid \boldsymbol{y}] = \\ &= \int_{X(y)} \ln L_C(\boldsymbol{\theta} \mid \boldsymbol{x}) k(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)}) d\boldsymbol{x} = \int_{X(y)} \ln L_C(\boldsymbol{\theta} \mid \boldsymbol{x}) \frac{\ln L_C(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{x})}{\ln L(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})} d\boldsymbol{x}, \end{aligned} \tag{3.117}$$

onde $\boldsymbol{x}$ corresponde aos dados completos, $\boldsymbol{y}$ aos dados observados, $\boldsymbol{z}$ aos dados "perdidos" e onde $X(\boldsymbol{y})$ é o subconjunto de X constituído pelos valores de $\boldsymbol{x}$ para os quais $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$.

O algoritmo EME toma a seguinte forma. A partir de um determinado valor inicial arbitrário $\boldsymbol{\theta}^{(0)}$, passando pelo passo-E estocástico (simulação) e pelo passo-M (maximização), é formada uma sequência $\{\boldsymbol{\theta}^{(k)}\}, k \in IN_0$.

Passo-E Estocástico. Dado um valor de $\boldsymbol{\theta}^{(k)}$, na iteração k, simular os dados completos $\boldsymbol{x}$ a partir de $k(\boldsymbol{x} \mid \boldsymbol{y};\boldsymbol{\theta}^{(k)})$ uma única vez. Seja essa simulação $\boldsymbol{x}^{(k)} = (\boldsymbol{y}^T, \boldsymbol{z}^{(k)T})^T$. Aproxima-se a função $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ por

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) \approx \ln L_C(\boldsymbol{\theta} \mid \boldsymbol{x}^{(k)}) \tag{3.118}$$

Passo-M. Maximizar a função de log-verosimilhança resultante $Q(\theta;\theta^{(k)})$ segundo θ, deixando o "maximizador" ser o próximo valor $\theta^{(k+1)}$.

O passo-E estocástico torna o conjunto de dados completo, e assim a maximização no passo-M torna-se numa estimação de MV de dados completos. O passo-M é normalmente de fácil resolução quer explicitamente, quer através de algoritmos como o de Newton-Raphson ou o método de Newton modificado ("scoring" de Fischer), utilizando a matriz $I_e(\theta^{(k)} \mid y)$ como uma aproximação à matriz $I(\theta^{(k)} \mid y)$.

Contrariamente ao algoritmo EM determinístico, simulando os dados em falta em cada iteração (dados pseudo-completos), a sequência de "maximizadores" $\{\theta^{(k)}\}, k \in IN_0$, do algoritmo EME é uma cadeia de Markov ergódica que, sob condições regulares, converge fracamente para a sua distribuição estacionária $\pi(\cdot)$ (Ip, 1994a e Gilks, Richardson e Spiegelhalter, 1996). Esta distribuição é aproximadamente centrada no EMV de θ e a variância reflecte a informação perdida que, neste caso, resulta não só do facto de se trabalhar com uma amostra e dados incompletos (variância devido ao modelo), mas também da simulação.

Na prática, é necessário deixar "correr" o algoritmo durante um determinado período (denominado o período de "queima") para permitir que a sequência $\{\theta^{(k)}\}$ se aproxime do seu regime estacionário. Esta sequência providencia uma região de interesse para θ que pode ser chamada de região plausível (Gilks, Richardson e Spiegelhalter,1996).

Sendo estocástico, cada vez que este algoritmo é activado providencia uma sequência $\{\theta^{(k)}\}, k \in IN_0$. Pode-se repeti-lo M vezes, considerando a média ao longo das várias (M) distribuições estacionárias correspondentes como uma estimativa de θ, melhorando assim a estimação dos parâmetros. Também se pode utilizar uma média de M valores consecutivos de uma sequência $\{\theta^{(k)}\}, k \in IN_0$ (após o periodo de queima). Quando tal acontece, chama-se a essa média a estimativa EME e denota-se por $\hat{\theta}_M$. O que se ganha por tomar esta média depende em grande medida da fracção de informação perdida. Se a fracção de informação perdida for grande então o ganho é pequeno, mas para pequenas fracções de informação perdida, o ganho pode ser bastante grande. Salvo para alguns exemplos simples, esta estimativa não coincide com o EMV. Resultados para a

consistência e normalidade assintótica do estimador $\hat{\theta}_M$ têm vindo a ser estabelecidos para exemplos específicos (Celeux e Diebolt, 1992).

Se as extracções aleatórias dos dados em falta forem substituídas pela média ou moda, obtém-se o que normalmente é conhecido, por um algoritmo tipo EM. Caso se utilize a média e se a função de log-verosimilhança dos dados completos para θ for linear em z, este algoritmo coincide com o algoritmo EM (Mclachlan e Krishnan, 1997).

3.3.4.3.1.1 Matriz de variâncias-covariâncias assintótica

Enquanto nos outros métodos apresentados a variância reflecte a informação perdida devido a se trabalhar com uma amostra e com dados incompletos, para algoritmos de simulação a variância dos estimadores pode ser dividida em duas partes, a parte do modelo, aproximada por $I^{-1}(\hat{\theta} \mid y)$, e a parte da simulação por, $I^{-1}(\hat{\theta} \mid y)\left[I - \{I + Inf_m(\hat{\theta} \mid y)\}^{-1}\right]$ (Nielsen, 2000). Assim, a variância assintótica de $\hat{\theta}$ é dada pelos elementos da diagonal principal da matriz

$$I^{-1}(\theta \mid y) + I^{-1}(\theta \mid y)\left[I - \{I + Inf_m(\theta \mid y)\}^{-1}\right] \tag{3.118}$$

para $\theta = \hat{\theta}$. Aqui $I(\hat{\theta} \mid y)$ é a matriz de informação observada, que pode ser aproximada por $I_e(\hat{\theta} \mid y)$ no caso de dados não agrupados (ver secção 3.3.4.2.1) ou por $I_{e,g}(\hat{\theta} \mid y)$ para dados agrupados (ver secção 3.3.4.2.3.1) e $Inf_m(\hat{\theta} \mid y) = \mathrm{cov}_\theta\left[S_C(\hat{\theta} \mid X) \mid y\right]$ é a matriz de informação perdida (3.68). Esta covariância pode ser aproximada substituindo-a pela covariância amostral, baseada nos valores simulados no passo-E, isto é

$$\mathrm{cov}_\theta\left[S_c(\hat{\theta} \mid X) \mid y\right] \approx \mathrm{cov}_\theta\left[\sum_{j=1}^{n} s(\hat{\theta} \mid y_j)\right], \tag{3.119}$$

(Gilks, Richardson e Spiegelhalter, 1996) com $\theta = (\theta)_i \ (i = 1, \ldots, p)$ e $s(\theta \mid y_j)$ dado por (3.29).

A variância adicional devido à simulação é uma função crescente da fracção de informação perdida. Assim escolhendo o modelo de dados completos de tal maneira que a fracção de informação perdida seja pequena pode-se controlar o aumento da variância.

A estimação pode ser melhorada considerando a média das últimas M iterações da cadeia de Markov $\{\theta^{(k)}\}, k \in IN_0$, o que reduz a variâncias dos estimadores. Prova-se que a variância assintótica de $\hat{\theta}_M$ (Nielsen, 2000) é dada pela diagonal principal da matriz

$$I^{-1}(\hat{\theta} \mid y) + \frac{1}{M} I^{-1}(\hat{\theta} \mid y)\left[I - \left\{I + Inf_m(\hat{\theta} \mid y)\right\}^{-1}\right] + \\ + \frac{2}{M} I^{-1}(\hat{\theta} \mid y)\left[I - \left\{I + Inf_m(\hat{\theta} \mid y)\right\}^{-1}\right] Inf_m(\hat{\theta} \mid y)\left(I - Inf_m(\hat{\theta} \mid y)\right)^{-1} - \\ - \frac{2}{M^2} I^{-1}(\hat{\theta} \mid y)\left[I - \left\{I + Inf_m(\hat{\theta} \mid y)\right\}^{-1}\right] Inf_m(\hat{\theta} \mid y)\left(I - Inf_m^M(\hat{\theta} \mid y)\right)\left(I - Inf_m(\hat{\theta} \mid y)\right)^{-2}$$

Repare-se que quanto maior for M menor vai ser a variância adicional devido à simulação, podendo-se mesmo desprezar o último termo para grandes valores de M. Tal facto é vantajoso devido à dificuldade do cálculo deste termo, pode, contudo, trazer algumas desvantagens, em termos computacionais, na construção do algoritmo.

3.3.4.3.1.2 Aplicação a um modelo de misturas de distribuições normais

Suponhamos que a f.d.p. de uma v.a. Y é uma mistura-φ em localização dada por

$$f(y \mid \Pi) = \sum_{i=1}^{d} \frac{\varphi_i}{\delta\sqrt{2\pi}} e^{-\frac{(y-\lambda_i)^2}{2\delta^2}},$$

onde d é o número de componentes da mistura, $\varphi_d = 1 - \sum_{i=1}^{d-1} \varphi_i$, $\lambda_i \in IR$ e $\delta > 0$.

Seja o vector das observações da mistura $y = (y_1, \ldots, y_n)^T$, o vector dos dados não observados $z = (z_1^T, \ldots, z_n^T)^T$ onde z_j é o vector composto por zeros e uns, indicando a componente à qual a observação de ordem j, y_j, pertence $(j = 1, \ldots, n)$.

Como foi visto na secção 3.3.4.2.2, $z_{ij}^{(k)} = \dfrac{\varphi_i^{(k)} e^{-\frac{1}{2}\left(\frac{y_j-\lambda_i^{(k)}}{\delta^{(k)}}\right)^2}}{\sum_{i=1}^{d} \varphi_i^{(k)} e^{-\frac{1}{2}\left(\frac{y_j-\lambda_i^{(k)}}{\delta^{(k)}}\right)^2}}$ é a probabilidade de $z_{ij}=1$ dadas as observações z quando $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(k)}$. Pode-se, então, formular o algoritmo EME:

Passo-E estocástico. Para cada j, simula-se $\boldsymbol{z}_j^{(k,1)} = \left(z_{1j}^{(k,1)},\ldots,z_{dj}^{(k,1)}\right)^T$ de modo que um destes tenha valor 1 e todos os outros valor 0. A escolha de i $(i=1,\ldots,d)$ tal que $z_{ij}^{(k,1)}=1$ faz-se de acordo com as probabilidades $z_{ij}^{(k)}$ $(i=1,\ldots,d)$.

Passo-M. Maximizar $Q\left(\boldsymbol{\Pi};\boldsymbol{\Pi}^{(k)}\right)$. Tal traduz-se, analogamente ao que foi referido na secção 3.3.4.2.2, ao cálculo de

$$\varphi_i^{(k+1)} = \sum_{j=1}^{n} \frac{z_{ij}^{(k,1)}}{n},\quad \lambda_i^{(k+1)} = \frac{\sum_{j=1}^{n} y_j z_{ij}^{(k,1)}}{\sum_{i=1}^{n} z_{ij}^{(k,1)}} \quad \text{e} \quad \delta^{2(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{d}\left(y_j-\lambda_i^{(k+1)}\right)^2 z_{ij}^{(k,1)}$$

Este algoritmo evita que a sequência fique próxima de um ponto de estacionaridade instável da função de log-verosimilhança, evitando, assim, os casos de convergência lenta que pode acontecer com a aplicação do algoritmo EM a misturas (vantajoso será, também neste caso por razões já explicadas, a combinação deste algoritmo com o método de Newton-Raphson). A sequência de estimativas é uma cadeia de Markov ergódica que converge fracamente para uma distribuição estacionária.

A matriz de informação perdida, visando o cálculo do erro padrão dos estimadores, pode ser aproximada pela expressão (3.119), enquanto a matriz de informação esperada pode ser aproximada por $I_e(\boldsymbol{\theta} \mid \boldsymbol{y})$.

3.3.4.3.1.3 Aplicação a dados agrupados

Caso os dados estejam agrupados. Seja $x=\left(y^T,w_1^T,\ldots,w_v^T\right)^T$ o vector dos dados completos, onde $y=(n_1,\ldots,n_v)^T$ é o vector das observações, $w_j=\left(w_{j1},\ldots,w_{jn_j}\right)^T$, $(j=1,\ldots,v)$ o vector dos dados "perdidos" e $z=\left(w_1^T,\ldots,w_v^T\right)^T$ (ver secção 3.3.4.2.3).

Tem-se

$$Q\left(\theta;\theta^{(k)}\right)=E_{\theta^{(k)}}\left[\ln L_C(\theta\mid x)\mid y\right]=\int_{w_1\in W_1^{n_1},\ldots,w_v\in W_v^{n_v}}\ln L_C(\theta\mid x)p\left(w_1,\ldots,w_v\mid y;\theta^{(k)}\right)\mathrm{d}w_1\ldots\mathrm{d}w_v\,,$$

Passo-E Estocástico. Dado um valor de $\theta^{(k)}$, na iteração *k*, simular os dados em falta $w_1,\ldots,w_v$ com densidade condicionada dado y segundo $\theta^{(k)}$. Cada w_{jl} $(l=1,\ldots,n_j)$ deve ser simulado uma vez, obtendo-se os valores $w_{jl}^{(k,1)}$. Para cada j, os valores $w_{j1}^{(k,1)},\ldots,w_{jn_j}^{(k,1)}$ são independentes e são simulados de acordo com a f.d.p.

$$h_j(w\mid\theta)=\frac{f(w\mid\theta)}{P_j(\theta)}\ (j=1,\ldots,v)\,. \tag{3.120}$$

Aproxima-se a função

$$Q\left(\theta;\theta^{(k)}\right)=\sum_{j=1}^{v}n_jQ_j\left(\theta;\theta^{(k)}\right) \tag{3.121}$$

por

$$Q\left(\theta;\theta^{(k)}\right)\approx\ln L_C\left(\theta\mid\begin{bmatrix}y\\w_1^{(k,1)}\\\vdots\\w_v^{(k,1)}\end{bmatrix}\right)=\sum_{j=1}^{v}\sum_{l=1}^{n_j}\ln f\left(w_{jl}^{(k,1)};\theta\right) \tag{3.122}$$

o que corresponde a utilizar a aproximação

$$Q_j\left(\theta;\theta^{(k)}\right)\approx\frac{1}{n_j}\sum_{l=1}^{n_j}\ln f\left(w_{jl}^{(k,1)};\theta\right). \tag{3.123}$$

Passo-M. Maximizar o valor aproximado resultante de $Q(\theta;\theta^{(k)})$ segundo θ, deixando o "maximizador" ser a próxima estimativa $\theta^{(k+1)}$.

Tal como na secção 3.3.4.2.3.1, pode-se utilizar o método de Newton-Rapshon,

$$\theta^{(k+1)} \approx \theta^{(k)} + I_{e,g}^{-1}(\theta^{(k)} \mid y) S(\theta^{(k)} \mid y). \tag{3.124}$$

A matriz de informação perdida, necessária para o cálculo do erro padrão dos estimadores, pode ser aproximada por

$$\mathrm{cov}_{\theta}\left[S_c(\hat{\theta} \mid X) \mid y\right] \approx \mathrm{cov}_{\theta}\left[\sum_{j=1}^{v} n_j s_j(\hat{\theta})\right], \tag{3.125}$$

(Gilks, Richardson e Spiegelhalter, 1996) com $\theta = (\theta)_i \ (i = 1,\ldots,p)$ e $s_j(\theta)$ dada por (3.112). A matriz de informação esperada pode ser aproximada por $I_{e,g}(\hat{\theta} \mid y)$.

3.3.4.3.2 Algoritmo EM de Monte Carlo

A formulação do algoritmo EMMC é idêntica à do algoritmo EME sendo formada na mesma a sequência $\{\theta^{(k)}\}, k \in IN_0$. A única diferença, na formulação, está no número de simulações requeridas no passo-E. Apresenta-se a estrutura do algoritmo EMMC para dados agrupados.

Passo-E Monte Carlo. Cada w_{jl} $(l = 1,\ldots,n_j)$ deve ser simulado M vezes, visando a sua utilização "melhorar" a estimativa da esperança condicional necessária no passo-E do algoritmo EM, obtendo-se os valores $w_{jl}^{(k,m)}$ i.i.d., $m = 1,\ldots,M$ e aproxima-se a função-Q, por

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)\approx\frac{1}{M}\sum_{m=1}^{M}\ln L_C\left(\boldsymbol{\theta}\,|\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{w}_1^{(k,m)}\\ \vdots\\ \boldsymbol{w}_v^{(k,m)}\end{bmatrix}\right)=\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{v}\sum_{l=1}^{n_j}\ln f\left(\boldsymbol{w}_{jl}^{(k,m)};\boldsymbol{\theta}\right). \quad (3.126)$$

Neste caso, tal equivale a tomar

$$Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)\approx\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_j}\sum_{l=1}^{n_j}\ln f\left(\boldsymbol{w}_{jl}^{(k,m)};\boldsymbol{\theta}\right) \quad (3.127)$$

Passo-M. Igual ao do algoritmo EME.

Em algoritmos iterativos deste género, a escolha de M e a consequente convergência do algoritmo pode-se tornar num problema difícil. Wei e Tanner (1990) recomendam que em vez de se escolher e fixar M para todo o algoritmo, se utilize, nas fases iniciais deste, pequenos valores para M que vão crescendo proporcionalmente à convergência. Para se "controlar" esta última, recomenda-se que os valores de $\boldsymbol{\theta}^{(k)}$ sejam tabelados ou visualizados graficamente para cada iteração e quando a convergência for indicada pela estabilização do processo com flutuações aleatórias à volta de um valor de $\hat{\boldsymbol{\theta}}$, o processo pode ser concluído ou continuado com um maior valor para M.

Trabalhando com o algoritmo EMMC perde-se a propriedade de monotonia, mas em certos casos, o algoritmo produz, com elevada probabilidade, valores perto do máximo; para detalhes ver Chan e Ledolter (1995).

3.3.4.3.2.1 Matriz de Variâncias-covariâncias assintótica

O raciocínio referente à variância das estimativas é análogo ao do algoritmo EME, contudo neste caso, é dada pela diagonal principal da matriz (Nielsen, 2000)

$$I_{e,g}^{-1}\left(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{y}\right)+\frac{1}{M}I_{e,g}^{-1}\left(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{y}\right)\left[I-\left\{I+Inf\left(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{y}\right)\right\}^{-1}\right]. \quad (3.128)$$

Enquanto $I_{e,g}^{-1}\left(\hat{\theta} \mid y\right)$ dá uma aproximação da variância devido ao modelo, $\frac{1}{M} I_{e,g}^{-1}\left(\hat{\theta} \mid y\right)\left[I - \left\{I + Inf\left(\hat{\theta} \mid y\right)\right\}^{-1}\right]$ dá uma aproximação da variância adicional devido à simulação. Repare-se que, com a aplicação do algoritmo EMMC, a redução na parte da variância devido à simulação é proporcional ao aumento do número de simulações.

3.3.4.3.2.2 Aplicação à distribuição normal com dados agrupados

Tendo em conta o que foi referido em cima pode-se formular o algoritmo do seguinte modo.

Seja W uma v.a. que segue uma distribuição normal com média μ e variância σ^2. Fazendo $\theta = \begin{bmatrix} \mu \\ S \end{bmatrix}$, com $S = \sigma^2$, a f.d.p. é,

$$f(w \mid \theta) = \frac{1}{\sqrt{2\pi S}} e^{-\frac{(w-\mu)^2}{2S}}.$$

O espaço amostral W de W, está dividido em v classes mutuamente exclusivas, W_j,

$$\mathsf{W}_j = [w_{j-1}, w_j[\ , \ j = 1, \ldots, v$$

onde $w_0 = -\infty$ e $w_v = \infty$.

Sejam $y = (n_1, \ldots, n_v)^T$ o vector das frequências observadas, $x = \left(y^T, w_1^T, \ldots, w_v^T\right)^T$ o vector dos dados completos onde $w_j = \left(w_{j1}, \ldots, w_{jn_j}\right)^T$ é o vector dos dados não observados na classe W_j.

Passo-E Monte Carlo. Em cada iteração k são geradas, para cada $j = 1, \ldots, v$, as variáveis aleatórias $w_{jl}^{(k,m)}$ $\left(l = 1, \ldots, n_j ; m = 1, \ldots, M\right)$ i.i.d.. Para as simulações, basta fazer

$$w_{jl}^{(k,m)} = \sqrt{S^{(k)}}\, t_{jl}^{(k,m)} + \mu^{(k)},$$

onde

$$t_{jl}^{(k,m)} = \Phi^{-1}\left(U_{jl}^{(k,m)}\left(G_{j,k} - G_{j-1,k}\right) + G_{j-1,k}\right),$$

com $G_{j,k} = \Phi\left(z_j\left(\boldsymbol{\theta}^{(k)}\right)\right)$, $z_j\left(\boldsymbol{\theta}^{(k)}\right) = \dfrac{w_j - \mu^{(k)}}{\sqrt{S^{(k)}}}$ e $U_{jl}^{(k,m)}$ números aleatórios $U[0,1)$ i.i.d (ver secção 4.3).

Como

$$\ln f\left(w_{jl}^{(k,m)} \mid \boldsymbol{\theta}\right) = -\frac{1}{2}\ln 2\pi S - \frac{\left(w_{jl}^{(k,m)} - \mu\right)^2}{2S},$$

resulta

$$Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) \approx \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_j}\sum_{l=1}^{n_j}\ln f\left(w_{jl}^{(k,m)};\boldsymbol{\theta}\right) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_j}\sum_{l=1}^{n_j}\left(-\frac{1}{2}\ln 2\pi S - \frac{1}{2S}\left(w_{jl}^{(k,m)} - \mu\right)^2\right) =$$

$$= \frac{1}{M}\left(-\frac{1}{2}M\ln(2\pi S) - \frac{1}{2S}\sum_{m=1}^{M}\frac{1}{n_j}\sum_{l=1}^{n_j}\left(w_{jl}^{(k,m)} - \mu\right)^2\right)$$

e, uma vez que

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) = \sum_{j=1}^{v} n_j Q_j\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right),$$

então

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) \approx \frac{1}{M}\left(-\frac{1}{2}Mn\ln(2\pi S) - \frac{1}{2S}\sum_{m=1}^{M}\sum_{j=1}^{v}\sum_{l=1}^{n_j}\left(w_{jl}^{(k,m)} - \mu\right)^2\right).$$

Passo-M. Pretende-se maximizar $Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)$ segundo $\boldsymbol{\theta}$. Então

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial \mu} = 0 \Rightarrow \mu^{(k+1)} = \frac{1}{Mn}\sum_{m=1}^{M}\sum_{j=1}^{v}\sum_{l=1}^{n_j} w_{jl}^{(k,m)}$$

$$\frac{\partial Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)}{\partial S} = 0 \Rightarrow S^{(k+1)} = \frac{1}{Mn}\sum_{m=1}^{M}\sum_{j=1}^{v}\sum_{l=1}^{n_j}\left(w_{jl}^{(k,m)} - \mu^{(k+1)}\right)^2,$$

o que completa a formulação do algoritmo EMMC para a distribuição normal.

Depois do período de "queima", obtém-se uma sequência de "maximizadores" $\{\theta^{(k)}\}, k \in IN_0$ que é uma cadeia de Markov que converge para a sua distribuição estacionária, podendo ser esta sequência utilizada para obter uma estimativa de θ, para k suficientemente grande.

Não se trabalhando com o método de Newton-Rapshon, o algoritmo não fornece automaticamente uma aproximação da matriz de variâncias-covariâncias assintótioca, então, através destas estimativas pode-se calcular $s_j(\hat{\theta})$, $\bar{s}(\hat{\theta})$, $I_{e,g}(\hat{\theta}; y)$ e, assim, $I_{e,g}^{-1}(\hat{\theta}; y)$ para matriz aproximada de variâncias-covariâncias dos estimadores.

Para a aplicação directa do método de Newton-Raphson ao algoritmo EMMC para dados agrupados (útil também quando o passo-M é difícil) vem

$$\theta^{(k+1)} \approx \theta^{(k)} + I_{e,g}^{-1}(\theta^{(k)}; y) S(y; \theta^{(k)}).$$

Neste caso as observações são os valores simulados no passo-E Monte Carlo.

3.3.4.3.3 Comparação entre o algoritmo EME e o algoritmo EMMC

Fazendo uma breve comparação entre os dois algoritmos, o algoritmo EME pode ser visto como uma versão simples do algoritmo EMMC, (Wei e Tanner 1990), onde $M = 1$. Este estimador do EMMC é assintoticamente equivalente a $\hat{\theta}_M$. Contudo, a redução na variância para a mesma escolha de M é sempre maior para as estimativas EMCM.

O tempo que se leva a completar uma iteração depende de algoritmo para algoritmo. Pode-se pensar que o algoritmo EMMC é mais rápido. Contudo, a maximização da log-verosimilhança poderá ser mais complicada e assim levar mais tempo do que o algoritmo EME. Está longe de ser óbvio qual dos algoritmos tem iterações mais rápidas ou qual necessita de menos iterações para convergir. Ambas as questões são de interesse. Enquanto o tempo de convergência é medido pelo CPU, o número de iterações é importante para se decidir qual dos algoritmos de deve utilizar.

4. Simulação de amostras recorrendo à técnica de Monte Carlo

O desenvolvimento da estatística moderna é inseparável do uso do computador. Os métodos de simulação que possibilitam a experimentação em estatística são disso exemplo eloquente.

Em muitas circunstâncias parte dos dados não estão disponíveis, tal facto pode acontecer, por exemplo, quando os dados se apresentam agrupados em classes, isto é, o conjunto de dados é considerado incompleto. Para os "completar", uma possibilidade será gerar artificialmente amostras - recorrendo à técnica de Monte Carlo - e, a partir das amostras geradas, estudar experimentalmente o comportamento das estatísticas em causa.

Como se verá, sabendo simular observações que tenham distribuição uniforme, é relativamente fácil gerar valores que tenham outra distribuição qualquer.

A geração de amostras aleatórias - ou, como frequentemente se diz, a geração de números aleatórios - passa, em geral, por duas fases:

(i) geração de números aleatórios seguindo uma distribuição uniforme no intervalo $[0,1)$, $U[0,1)$;

(ii) transformação daqueles números noutros igualmente aleatórios, mas seguindo uma outra distribuição qualquer pretendida.

4.1 Geração de amostras aleatórias com distribuição $U[0,1)$

A geração de números pseudo-aleatórios seguindo uma distribuição uniforme, $U[0,1)$, foi feita utilizando o gerador do "excel" aleatório(). Estes valores são obtidos por processos numéricos (deterministicamente). Contudo, apesar de não serem, de facto, aleatórios, têm um comportamento estatístico que os torna indistinguíveis de uma amostra de números aleatórios uniformes entre zero e um. É por esta razão que recebem a designação de números pseudo-aleatórios.

Visando testar o gerador do "excel", depois de se gerar uma quantidade suficientemente grande de números aleatórios, espera-se que a média e o desvio padrão

sejam aproximadamente 1/2 e $1/\sqrt{12}$, respectivamente . Por outro lado, agrupando estes números em classes mutuamente exclusivas da mesma amplitude, percorrendo o intervalo $[0,1[$, é de esperar que o aspecto gráfico das frequências absolutas dessas classes tenha um comportamento uniforme (ver figura 4.1).

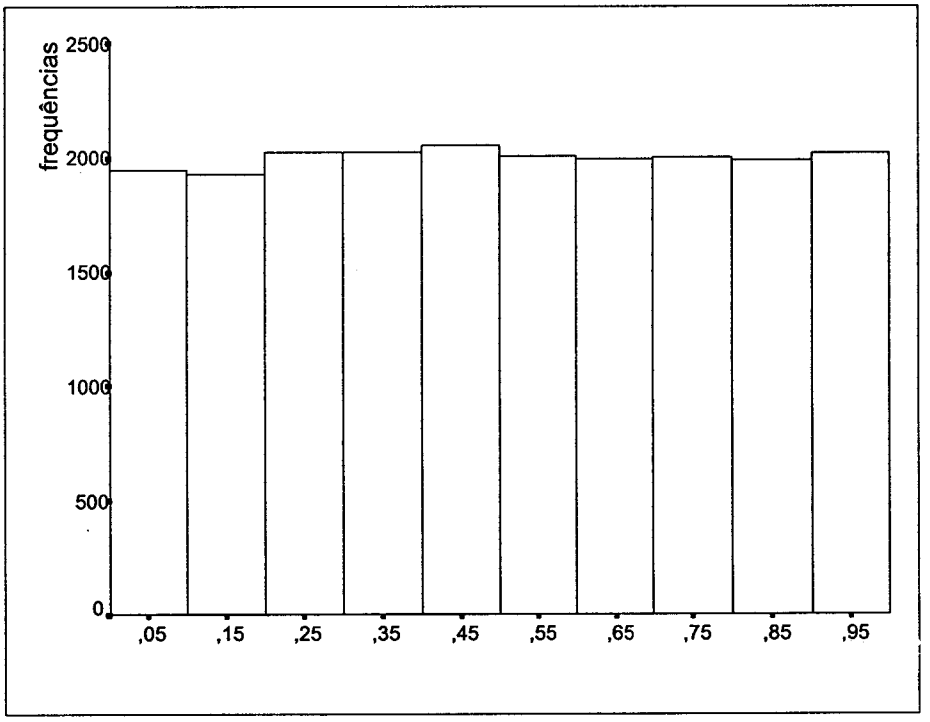

Figura4.1 – Um dos histogramas obtidos para o agrupamento de 20000 pseudo-números aleatórios uniformes [0,1).

Assim, depois de se gerar várias amostras de 20000 números pseudo-aleatórios a partir do gerador do "excel", verificou-se que a média e a variância são aproximadamente iguais ás pretendidas. Agrupando cada uma das amostras em 10 classes com amplitude 0,1 mutuamente exclusivas percorrendo o intervalo $[0,1[$, pode-se realizar um teste de ajustamento sobre estes números, não se rejeitando a hipótese de que eles sejam provenientes de uma distribuição $U[0,1)$.

Fez-se também um teste de autocorrelação para testar a independência dos sucessivos valores gerados. Como a função de autocorrelação está contida dentro da região de aceitação de autocorrelação nula aceita-se a hipótese de independência (ver figura 4.2)

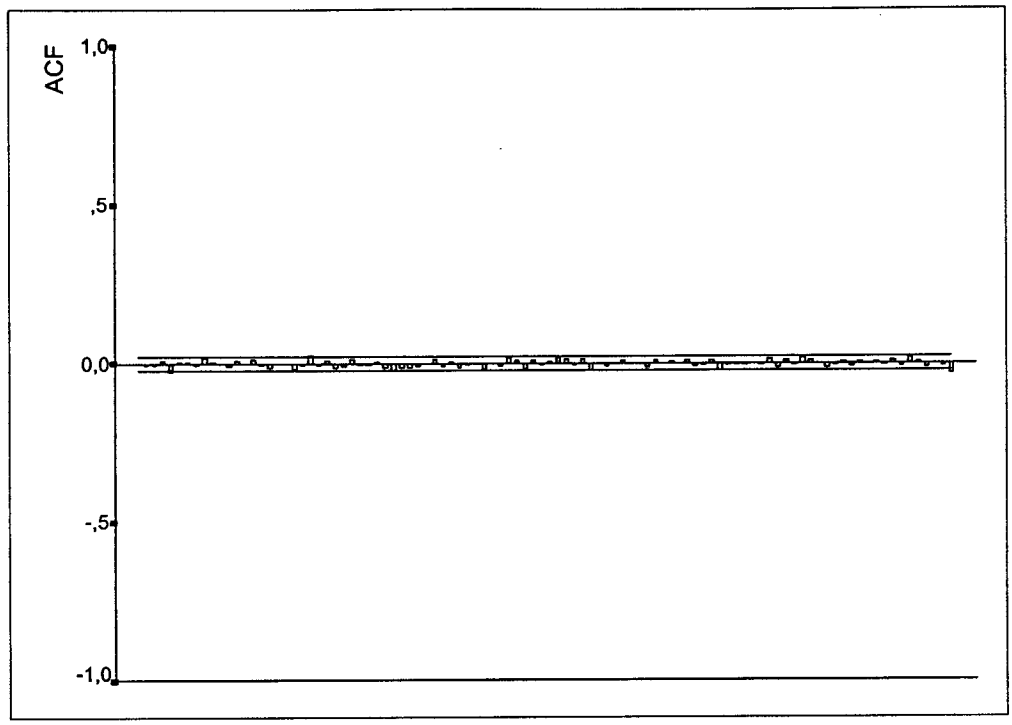

Figura4.2 – Gráfico relativo a um teste de correlação serial para números aleatórios pseudo-uniformes [0,1).

Dispondo de um gerador de números uniformes, põe-se o problema de gerar números aleatórios com uma distribuição qualquer.

4.2 Simulação de amostras aleatórias provenientes de uma população contínua qualquer

Uma técnica muito utilizada, conhecida por método da função inversa, consiste em gerar números aleatórios com uma distribuição qualquer usando como ponto de partida os números pseudo-aleatórios com distribuição $U[0,1)$.

Comece-se por considerar uma v.a. contínua X, com f.d.p. $f(x \mid \boldsymbol{\theta})$ e f.d. invertível

$$F(x \mid \boldsymbol{\theta}) = \int_{-\infty}^{x} f(t \mid \boldsymbol{\theta}) \mathrm{d}t . \tag{4.1}$$

Defina-se, por transformação da variável X, uma nova variável, $Z = F(X \mid \boldsymbol{\theta})$.

As funções distribuição, G(z), e densidade, g(z), da nova variável vêm dadas por

$$G(z) = P(Z < z) = P\left[X < x = F^{-1}(z)\right] = F\left[F^{-1}(z)\right] = z \tag{4.2}$$

e

$$g(z) = \frac{\mathrm{d}G(z)}{\mathrm{d}z} = 1, \tag{4.3}$$

no domínio $z \in [0,1]$. Ora a forma destas funções implica que $Z \sim U[0,1)$.

A conclusão a que se chegou é, portanto, a de que, se uma variável X segue uma distribuição com f.d.p. $f(x \mid \boldsymbol{\theta})$ e f.d. $F(x \mid \boldsymbol{\theta})$, invertível, então a variável transformada $Z = F(X \mid \boldsymbol{\theta})$ segue uma distribuição $U[0,1)$.

Inversamente, se uma variável Z segue uma distribuição $U[0,1)$, então a variável $X = F^{-1}(Z)$, onde F é a f.d., invertível, de uma v.a. contínua, segue uma distribuição com f.d.p. $f(x \mid \boldsymbol{\theta})$ e f.d. $F(x \mid \boldsymbol{\theta})$.

Algumas f.d.p.'s não são integráveis analiticamente, podendo a f.d. ser obtida, de forma aproximada, por integração numérica. No "excel" estão contempladas, entre várias, as f.d.'s para as distribuições normal, lognormal e gama, que se encontram nestas situação.

A geração de amostras aleatórias provenientes de populações contínuas com funções de distribuição não invertíveis pode ser resolvida de forma relativamente simples adaptando ligeiramente o procedimento anteriormente apresentado; por exemplo, quando forem gerados números aleatórios $U[0,1)$ iguais aos valores onde a função não é invertível ignoram-se esses valores e geram-se outros diferentes em sua substituição.

4.3 Simulação de variáveis aleatórias para dados agrupados

Considere-se o espaço amostral da v.a. X, com f.d.p. $f(x \mid \boldsymbol{\theta})$ e f.d. $F(x \mid \boldsymbol{\theta})$, dividido em v classes mutuamente exclusivas X_j $(j = 1 \ldots v)$. Suponha-se que $\mathrm{X}_j = [x_{j-1}, x_j[$. Pretende-se simular valores X condicionados a pertencerem à classe X_j, isto é, cada X é retirado de uma distribuição com densidade,

$$\begin{cases} \dfrac{f(x \mid \boldsymbol{\theta})}{P_j(\boldsymbol{\theta})} & \text{para } x \in \mathsf{X}_j \\ 0 & \text{caso contrário,} \end{cases} \tag{4.4}$$

onde $P_j(\boldsymbol{\theta}) = \int_{\mathsf{X}_j} f(x \mid \boldsymbol{\theta}) dx = F(x_j \mid \boldsymbol{\theta}) - F(x_{j-1} \mid \boldsymbol{\theta})$ é a probabilidade de X pertencer à classe X_j.

Passa-se a dar um exemplo para a distribuição normal $N(\mu, \sigma^2)$, para depois se apresentarem os resultados para as restantes distribuições teóricas utilizadas.

Pretende-se gerar, para cada X_j, v.a's X, i.i.d. com f.d.p.

$$\frac{\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\Phi(z_j(\boldsymbol{\theta})) - \Phi(z_{j-1}(\boldsymbol{\theta}))}, \text{ para } x \in [x_{j-1}, x_j[$$

onde $z_j(\boldsymbol{\theta}) = \dfrac{x_j - \mu}{\sigma}$ e $\boldsymbol{\theta} = (\mu, \sigma)$. Note-se que $z_j(\boldsymbol{\theta})$, condicionado a X pertencer à classe X_j, tem f.d.p.

$$\frac{\dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}{\Phi(z_j(\boldsymbol{\theta})) - \Phi(z_{j-1}(\boldsymbol{\theta}))} \text{ para } z \in [z_{j-1}(\boldsymbol{\theta}), z_j(\boldsymbol{\theta})[$$

e f.d. dada por

$$\frac{\Phi(z) - \Phi(z_{j-1}(\boldsymbol{\theta}))}{\Phi(z_j(\boldsymbol{\theta})) - \Phi(z_{j-1}(\boldsymbol{\theta}))} \text{ para } z \in [z_{j-1}(\boldsymbol{\theta}), z_j(\boldsymbol{\theta})[\ .$$

Pelo método da função inversa, obtém-se

$$Z = \Phi^{-1}\left[U\left(\Phi(z_j(\boldsymbol{\theta})) - \Phi(z_{j-1}(\boldsymbol{\theta}))\right) + \Phi(z_{j-1}(\boldsymbol{\theta}))\right],$$

onde U são números aleatórios $U[0,1)$, e assim $X = \sigma Z + \mu$ pertence à classe X_j.

Por raciocínio análogo, vem para as restantes distribuições:

- lognormal $LN(\lambda, \delta^2)$, basta utilizar $\ln x$ em vez de x e o raciocínio é idêntico ao anterior;
- gama $G(\lambda, \delta, \rho)$, $X = F^{-1}[UP_j(\theta) + E]$ com $E = F(x_{j-1} \mid \theta)$, onde F é a f.d. da distribuição gama referida;
- Burr1 $B1(\lambda, \delta, \rho)$, $X = \delta - \lambda \ln\left[\left(UP_j(\theta) + E\right)^{-\frac{1}{\rho}} - 1\right]$, com $E = F(x_{j-1} \mid \theta)$, onde F é a f.d. da distribuição Burr1 referida;
- Burr2 $B2(\lambda, \delta, \kappa, \rho)$, $X = \delta + \lambda \ln\left[\left(1 + \frac{UP_j(\theta) + E}{k(1 - UP_j(\theta) - E)}\right)^{\frac{1}{r}} - 1\right]$, com $E = F(x_{j-1} \mid \theta)$; onde F é a f.d. da distribuição Burr2 referida.

A simulação dos valores para as diferentes classes, que seguem um modelo que é a mistura de duas distribuições, é feita do seguinte modo:

(1) simular um valor aleatório $U[0,1)$;

(2) sendo φ a contribuição da primeira componente da mistura, se o número aleatório simulado em (1) for menor ou igual a φ, simulam-se os valores seguindo a distribuição da primeira componente da mistura (como em cima); caso contrário os valores simulados seguirão uma distribuição dada pela segunda componente da mistura. Este procedimento pode ser realizado com a função "SE" do "excel".

5. Tratamento dos dados

5.1 Objectivos

O presente estudo, insere-se no âmbito da investigação do ajustamento de distribuições paramétricas contínuas a dados organizados numa chave de comprimentos à idade. Estas são utilizadas em estudos de pescas com dois objectivos práticos:

1. Encontrar a estrutura etária da população a partir de uma distribuição de comprimentos;

2. Encontrar parâmetros e/ou taxas de crescimento, ajustando os valores médios do comprimento à idade a modelos de crescimento (comprimento em função do tempo/idade).

Os desvios em relação à normalidade são importantes do ponto de vista biológico, o que leva a estudar um conjunto de modelos (ver secção 2.2) alternativos à distribuição normal que permitam variações de assimetria na distribuição do comprimento à idade.

O objectivo principal deste trabalho será, assim, a escolha de modelos alternativos, biologicamente significativos, que permitam a análise das inúmeras chaves de comprimento à idade publicadas, com o fim de estudar e interpretar, do ponto de vista biológico, alterações de simetria na distribuição do comprimento à idade.

De uma maneira geral, estabeleceram-se as seguintes etapas:

1. Organização dos dados em chaves de comprimento à idade.

2. Cálculo das diferentes estatísticas amostrais necessárias ao estudo.

3. Estimação dos parâmetros das diferentes distribuições teóricas, utilizando para isso, quando possível, o método da máxima verosimilhança (MV) e métodos iterativos para o cálculo das estimativas de MV (métodos de Newton, algoritmo EM e algoritmo EM com simulação). O algoritmo EM é mesmo necessário se se quiser considerar os efeitos do agrupamento de dados ou de mistura de distribuições. Em

caso de impossibilidade de se utilizar métodos de MV, usa-se o método dos momentos.

4. Determinação da qualidade de ajustamento das diferentes distribuições às diferentes amostras através de testes de qualidade de ajuste, o teste de qui-quadrado (ou χ^2) e o teste de Kolmogorov-Smirnov (K-S), do critério de informação de Akaike (AIC) (Akaike 1973, 1985), e por visualização gráfica.

No subsequente tratamento de dados, foram utilizados os seguintes programas informáticos: excel, mathematica e spss.

5.2 Caracterização dos dados

Foram consideradas amostras aleatórias da população de interesse, e determinados o comprimento e idade de cada indivíduo. Os comprimentos (cm) são medidos através de um aparelho próprio de medida, o ictiómetro. Para cada indivíduo a idade (anos) é observada nas suas estruturas calcificadas (otólitos ou escamas). As frequências absolutas foram, então, organizadas numa chave de comprimento à idade (vector das observações).

Para testar a qualidade do ajuste de cada uma das distribuições teóricas, de entre uma compilação de 458 chaves de comprimento à idade (Erzini, 1990), foram utilizadas 4 dessas chaves relativas a uma espécie 'típica'. Esta espécie, a pescada (*Merlucius bilinearis*), tem taxas de crescimento e número de classes de idade, presentes na população, intermédios. Ao longo do trabalho verificou-se que os resultados eram análogos para todas as chaves estudadas, pelo que se apresentam apenas os resultados referentes a uma delas (tabela5.1).

Repare-se que a amplitude de cada classe é 1 cm. Para além disso, exceptuando para a Idade1, a assimetria é sempre positiva (os indivíduos crescem sem acção dos predadores) e a curtose é sempre superior a zero (distribuição menos achatada do que a distribuição normal). Pelo que se espera que alguma distribuição teórica que esteja nestas condições se ajuste melhor aos dados do que a distribuição normal.

Pescaria 3					
Comp.	Idade1	Idade2	Idade3	Idade4	Idade5
6	1				
7	7				
8	16				
9	7				
10	8				
11	10				
12	11				
13	7				
14	8				
15	7				
16	10				
17	9				
18	8				
19	21				
20	16	2			
21	11	8			
22	14	10			
23	2	21			
24	3	38	2		
25		35	2		
26	1	37	2		
27		29	8		
28		30	20	3	
29		33	20	4	1
30		18	24	4	
31		20	34	10	6
32		17	29	11	10
33		7	21	13	7
34		3	19	5	6
35		5	17	9	2
36		1	15	7	3
37		3	9	6	3
38			2	4	3
39		1	3	1	3
40		1	3		2
41			2	2	
42			1		4
43					2
44					4
45				1	1
46					1
47					1
n	177	319	234	80	59
$\bar{x}$	15,4520	27,2038	32,0342	33,6000	36,2373
s	5,0294	3,5287	3,3786	3,1926	4,7682
$\hat{y}_1$	-0,1809	0,5881	0,4983	0,7267	0,6556
$\hat{y}_2$	-1,2175	0,34445	0,4834	1,1222	0,81171

Tabela5.1- Chave comprimento à idade, com algumas estatísticas amostrais. Para cada idade (anos) e classe de comprimentos (cm), indica-se o nº de peixes da amostra com essa idade cujo comprimento está na classe indicada.

5.3 Comparação dos diferentes métodos de estimação para a distribuição normal

5.3.1 Log-verosimilhança e AIC

Optou-se, para estudar a qualidade das aproximações fornecidas pelos diferentes algoritmos, fazer uma comparação dos mesmos para o caso da distribuição normal. Apresenta-se de seguida, para além das estimativas dos parâmetros, os valores da log-verosimilhança e do critério de informação de Akakie (AIC) (ver apêndice A2). Utilizaram-se mais algarismos significativos do que seria adequado à natureza dos dados para uma melhor percepção das pequenas entre os métodos. A unidade de medida para $\hat{\mu}$ e $\hat{\sigma}$ é o centímetro.

normal	$\hat{\mu}$	$\hat{\sigma}^2$	$V(\hat{\theta} \mid y)$	AIC
Idade1	15,4520	25,1512	-536,5578	1077,1156
Idade2	27,2038	12,4129	-854,3815	1712,7630
Idade3	32,0342	11,3663	-616,4185	1236,8370
Idade4	33,6000	10,0648	-205,8776	415,7552
Idade5	36,2373	22,3502	-175,3694	354,7388

Tabela5.2 – Estimativas de MV dos parâmetros da distribuição normal utilizando o algoritmo de Newton-Raphson para dados não agrupados.

Obs: Neste caso considera-se *y* como o vector das observações, uma vez que se consideram os dados incompletos.

normal	$\hat{\mu}$	$\hat{\sigma}^2$	$V(\hat{\theta} \mid x)$	AIC
Idade1	15,4520	25,0680	-536,5583	1077,1166
Idade2	27,2038	12,3297	-854,3851	1712,7702
Idade3	32,0342	11,2829	-616,4217	1236,8434
Idade4	33,6045	10,0343	-205,8779	415,7558
Idade5	36,2373	22,2671	-175,3696	354,7392

Tabela5.3 – Estimativas de MV dos parâmetros da distribuição normal utilizando o algoritmo EM para dados agrupados com utilização do método de Newton-Raphson.

Convém salientar que os resultados obtidos com a utilização do algoritmo EM para dados agrupados são idênticos quer se utilize ou não o método de Newton-Raphson. E uma vez que os dados estão agrupados em classes de 1 cm de amplitude, serão estes os valores de referência em relação aos quais se estudará a validade dos métodos aproximados. Uma vez que a utilização do método Newton-Raphson providencia automaticamente a matriz de variâncias-covariâncias, opta-se por este último.

normal	$\hat{\mu}_{30}$	$\hat{\sigma}^2{}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1	15,4625	25,1693	-536,5583	1077,11651
Idade2	27,2110	12,3911	-854,3824	1712,7648
Idade3	32,1886	11,3374	-616,66487	1237,3297
Idade4	33,5917	10,0312	-205,87815	415,7563
Idade5	36,2600	22,2143	-175,37062	354,7412

Tabela5.4 – Estimativas de MV dos parâmetros da distribuição normal utilizando o algoritmo EME para dados agrupados e fazendo a média de 30 valores consecutivos de sequências da iteração, com utilização do método de Newton-Raphson. O período de "queima" foi de 500 iterações.

normal	$\hat{\mu}_{30}$	$\hat{\sigma}^2{}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1	15,4672	25,2074	-536,5589	1077,1177
Idade2	27,2109	12,3897	-854,3824	1712,7648
Idade3	32,0313	11,3172	-616,41973	1236,8395
Idade4	33,6004	9,9742	-205,8793	415,7586
Idade5	36,2669	22,2841	-175,3707	354,7413

Tabela5.5 – Estimativas de MV dos parâmetros da distribuição normal utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson. O período de queima" foi de 500 iterações.

Algumas conclusões podem ser retiradas da comparação das diferentes tabelas, resumidamente:

- As médias pouco se alteram;
- As variâncias diminuem um pouco com a utilização do algoritmo EM, pois este tem em conta o facto de os dados estarem agrupados;
- A log-verosimilhança e o *AIC* são praticamente iguais utilizando qualquer um dos métodos mencionados.

- Os algoritmos estocásticos dão resultados muito próximos do algoritmo EM propriamente dito, o que dá uma certa tranquilidade nas situações em que o algoritmo EM não é praticável. Deve-se ter em conta que cada vez que os algoritmo estocásticos são activados obtêm-se estimativas um pouco diferentes, facto este que não interfere de uma maneira significativa nos valores obtidos para os dois coeficientes analisados.

Assim, parece que, independentemente do método, as conclusões a tirar serão as mesmas. Contudo, como também se irá trabalhar com misturas, cuja estimação dos parâmetros é bastante facilitada com a aplicação do algoritmo EM, e uma vez que, estando os dados agrupados, a aplicação do algoritmo EM para dados agrupados só é viável para as distribuições normal e lognormal (devido às dificuldades encontradas no cálculo da esperança matemática condicionada necessária no passo-M), optou-se pela apresentação dos resultados obtidos através do algoritmo EMMC $(M = 30)$ com utilização do método de Newton-Raphson. Exceptuando para as misturas, considera-se o caso de dados agrupados.

A utilização de algoritmos EM estocásticos provoca, no entanto, como se irá verificar de seguida, o aumento da amplitude dos intervalos de confiança, em virtude da variância adicional devido à simulação. Esta variância adicional pode controlar-se aumentando M mas tal torna os algoritmos mais lentos.

5.3.2 Intervalos de confiança

Relativamente aos intervalos de confiança, apresenta-se um exemplo ilustrativo referente à Idade2 com a aplicação dos diferentes algoritmos visando estimar os parâmetros da distribuição normal. Para 95% de confiança, os extremos dos intervalos de confiança aproximados para cada parâmetro são dados por $\hat{\theta}_i \pm 1{,}96 SE\left(\hat{\theta}_i\right)$ $(i = 1, \ldots, p)$. De facto, as estimativas de MV são assintoticamente normais pelo que utilizam os valores críticos desta distribuição como aproximação. Utilizam-se também valores aproximados de $SE\left(\hat{\theta}_i\right)$, que resulta do facto destes erros padrão se basearem na matriz de variâncias-covariâncias assintótica (inversa da matriz de informação esperada) que é aproximada pela

matriz de informação observada $I(\hat{\theta} \mid y)$, $I_e(\hat{\theta} \mid x)$ ou $I_{eg}(\hat{\theta} \mid x)$ consoante se utilize o método de MV, o algoritmo EM ou o algoritmo EM para dados agrupados (claro que se tem, também, a variância aproximada devido à simulação nos algoritmos estocásticos).

Considerando os parâmetros da distribuição μ e σ^2, vem:

- Método de MV

$\hat{\mu} = 27{,}2038$ e $\hat{\sigma}^2 = 12{,}4130$, $I^{-1}(\hat{\theta} \mid y) = \begin{bmatrix} 0{,}0389 & 0{,}0000 \\ 0{,}0000 & 0{,}9660 \end{bmatrix}$, $\theta = [\hat{\mu} \quad \hat{\sigma}^2]^T$.

$SE(\hat{\mu}) = 0{,}1973$, $SE(\hat{\sigma}^2) = 0{,}9829$. Assim os intervalos de confiança aproximados são $\mu : [26{,}8171; 27{,}5904]$ e $\sigma^2 : [10{,}4866; 14{,}3395]$.

- Algoritmo EM com Newton-Raphson e dados agrupados

$\hat{\mu} = 27{,}2038$ e $\hat{\sigma}^2 = 12{,}3297$, $I_{e,g}^{-1}(\hat{\theta} \mid x) = \begin{bmatrix} 0{,}0457 & -0{,}0812 \\ -0{,}0812 & 0{,}9769 \end{bmatrix}$

$SE(\hat{\mu}) = 0{,}2137$, e $SE(\hat{\sigma}^2) = 0{,}9884$. Resulta para intervalos de confiança aproximados $\mu : [26{,}7850; 27{,}6226]$ e $\sigma^2 : [10{,}3925; 14{,}2669]$.

- Algoritmo EME com Newton-Raphson e dados agrupados

$\hat{\mu}_{30} = 27{,}2110$ e $\hat{\sigma}_{30}^2 = 12{,}3911$, vem $I_{e,g}^{-1}(\hat{\theta} \mid x) = \begin{bmatrix} 0{,}0457 & -0{,}0814 \\ -0{,}0814 & 0{,}9879 \end{bmatrix}$ e a variância adicional devido à simulação 0,0420 e 0,3107, respectivamente, donde $SE(\hat{\mu}_{30}) = 0{,}2962$ e $SE(\hat{\sigma}_{30}^2) = 1{,}1396$. Vem para intervalos de confiança aproximados $\mu : [26{,}6477; 27{,}8089]$ e $\sigma^2 : [10{,}1599; 14{,}6268]$.

- Algoritmo EMMC com Newton-Raphson e dados agrupados

$\hat{\mu}_{30} = 27{,}2109$ e $\hat{\sigma}_{30}^2 = 12{,}3897$, vem $I_{e,g}^{-1}(\hat{\theta} \mid x) = \begin{bmatrix} 0{,}0466 & -0{,}0846 \\ -0{,}0846 & 0{,}9942 \end{bmatrix}$ e a variância adicional devido à simulação 0,0014 e 0,0104, respectivamente, para μ e σ^2. Donde

$SE(\hat{\mu}_{30}) = 0{,}2190$ e $SE(\hat{\sigma}^2_{30}) = 1{,}0023$. Os intervalos de confiança aproximados são $\mu : [26{,}7707 ; 27{,}6293]$ e $\sigma^2 : [10{,}4595 ; 14{,}3884]$

Os erros padrão para os parâmetros aumentam pouco com a aplicação do algoritmo EM comparativamente à aplicação do método de MV. Em relação ao algoritmo EM a estimativa aproximada da variância devida ao modelo é aproximadamente igual aplicando os três métodos alternativos. Contudo, nos algoritmos estocásticos, a variância dos parâmetros é inflacionada uma vez que inclui a variância adicional devida à simulação. Esta última é menor para o algoritmo EMMC por razões já mencionadas.

5. 4 Resultados gerais

5.4.1 Log-verosimilhança e AIC

Com a aplicação do algoritmo EMMC com $M = 30$ os resultados obtidos são resumidos nas seguintes tabelas:

normal	$\hat{\mu}_{30}$	$\hat{\sigma}^2{}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1	15,4672	25,2074	-536,5589	1077,1177
Idade2	27,2109	12,3897	-854,3824	1712,7648
Idade3	32,0313	11,3172	-616,41973	1236,8395
Idade4	33,6004	9,9742	-205,8793	415,7586
Idade5	36,2669	22,2841	-175,3707	354,7413

Tabela5.6 – Estimativas de MV dos parâmetros da distribuição normal utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson.

lognormal	$\hat{\lambda}_{30}$	$\hat{\delta}^2_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1	2,6766	0,1326	-546,0759	1096,1518
Idade2	3,2952	0,0162	-846,0033	1696,0066
Idade3	3,4614	0,0108	-612,4490	1228,8980
Idade4	3,5105	0,0087	-204,6459	413,2918
Idade5	3,5819	0,0160	-173,1362	350,2723

Tabela5.7 – Estimativas de MV dos parâmetros da distribuição lognormal utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson.

Gama	$\hat{\lambda}_{30}$	$\hat{\delta}_{30}$	$\hat{\rho}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1*					
Idade2	16,6656	1,1897	8,8582	-843,6163	1693,2326
Idade3	16,5418	0,7299	21,2243	-612,1022	1230,2296
Idade4	24,4207	1,0999	8,3459	-202,7076	411,4152
Idade5	28,5045	3,0628	2,5248	-168,4894	342,9788

Tabela5.8 – Estimativas de MV dos parâmetros da distribuição gama utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson.

*Não foi possível obter estimativas de MV para a Idade1 para os parâmetros da distribuição gama. Utilizando o método dos momentos, vem: $\hat{\lambda} = -40{,}1496$, $\hat{\delta} = 0{,}4549$, $\hat{\rho} = 122{,}2215$, $V(\hat{\theta} \mid x) = -537{,}5560$ e $AIC = 1081{,}1120$.

Burr1	$\hat{\lambda}_{30}$	$\hat{\delta}_{30}$	$\hat{\rho}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1	21,0157	1,1973	0,2028	-541,4244	1088,8488
Idade2	20,9055	2,7318	6,1782	-845,7734	1697,5468
Idade3	28,9706	2,3639	2,5427	-612,2219	1230,4438
Idade4	28,8697	2,3755	4,6343	-203,0792	412,1584
Idade5*					

Tabela5.9 – Estimativas de MV dos parâmetros da distribuição Burr1 utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson.

*Não foi possível obter estimativas de MV para a Idade5 para os parâmetros da distribuição Burr1. Utilizando o método dos momentos, vem: $\hat{\lambda} = 32{,}4278$, $\hat{\delta} = 3{,}2254$, $\hat{\rho} = 2{,}3071$, $V(\hat{\theta} \mid x) = -173{,}6579$ e $AIC = 353{,}3158$.

Burr2	$\hat{\delta}_{30}$	$\hat{\lambda}_{30}$	$\hat{\kappa}_{30}$	$\hat{\rho}_{30}$	$V(\hat{\theta} \mid x)$	AIC
Idade1*						
Idade2	21,3838	0,6560	0,2275	0,2738	-844,7096	1697,4192
Idade3	27,0041	1,0048	0,2734	0,4481	-611,9828	1231,9656
Idade4	27,9071	0,0241	0,1836	0,0116	-201,0867	410,1734
Idade5	30,7916	0,4783	1,4821	0,0999	-167,0418	342,0836

Tabela5.10 – Estimativas de MV dos parâmetros da distribuição Burr2 utilizando o algoritmo EMMC para dados agrupados para M=30 com utilização do método de Newton-Raphson. *Para a idade1 não foi possível obter estimativas dos parâmetros da distribuição Burr2 (por qualquer um dos métodos aqui apresentados).

Da análise das tabelas em apresentadas podem-se tirar as seguintes conclusões para cada uma das idades.

- Idade1: Esta amostra é a única que apresenta assimetria negativa. Seria de esperar que uma das distribuições de Burr dessem os melhores valores da função log-verosimilhança e do *AIC*, uma vez que estas distribuições são as únicas que contemplam a assimetria negativa. É, contudo, a distribuição normal que apresenta os valores mais elevados do módulo para a função de log-verosimilhança e do critério *AIC*. Devido à assimetria negativa da amostra, como seria de esperar, não foi possível obter estimativas de MV para a distribuição gama, passando-se a utilizar as estimativas obtidas pelo método dos momentos, que dão os valores mais próximos (em relação à distribuição normal) para a função de log-verosimilhança e para o *AIC*. Para a distribuição Burr2 não foi possível obter quaisquer estimativas pelos métodos apresentados (é, contudo, possível obter valores dos parâmetros que, apesar de anularem aproximadamente o gradiente da função de log-verosimilhança, não permitem que os algoritmo utilizados convirjam). Como se verá na análise gráfica, esta amostra apresenta uma distribuição que difere no padrão das demais utilizadas, cujas possíveis razões se apresentarão então;
- Idade2: Neste caso é a distribuição gama que apresenta os melhores valores para a função de log-verosimilhança e para o *AIC*. O valor da função de log-verosimilhança para a distribuição Burr2 é o mais próximo do obtido para a distribuição gama, o que não acontece com o valor para o AIC é o valor da distribuição lognormal o que mais se aproxima (a distribuição Burr2 tem mais dois parâmetros do que a distribuição lognormal);
- Idade3: O melhor valor para função de log-verosimilhança dá-se para a distribuição Burr2 seguindo-se a distribuição gama. Contudo, exceptuando o caso da distribuição normal, os restantes valores obtidos para esta função estão muito próximos. Em relação ao *AIC* verifica-se que o melhor valor é para a distribuição lognormal, seguindo-se a distribuição gama e distribuição Burr1, respectivamente (naturalmente devido ao número de parâmetros das distribuições);
- Idade4: Mais uma vez é a distribuição Burr2 que apresenta o melhor valor para a função de log-verosimilhança, aqui acentua-se um pouco a diferença em relação às

outras distribuições, sendo o valor mais próximo o correspondente à distribuição gama. Apesar do número de parâmetros ser 4, é a distribuição Burr2 que tem o maior valor do AIC, mas muito próximo do obtido para a distribuição gama;

- Idade5: Tal como na Idade4, a distribuição Burr2 apresenta os melhores resultados para a função de log-verosimilhança e para o AIC (repare-se que esta amostra é de tamanho reduzido, $n = 59$), aparecendo a distribuição gama com os segundos melhores valores para estes dois coeficientes.

Portanto, são as distribuições gama e Burr2 que se apresentam mais vezes referenciadas na análise efectuada em cima. Repare-se que as amostras de menores dimensões apresentam os melhores valores do *AIC* para a distribuição Burr2 (maior número de parâmetros). Não se pode, para já, prever um modelo mais adequado para a generalidade das idades, mas sim, diferentes modelos para diferentes amostras. Na prática, o *AIC* é frequentemente utilizado para ordenar os modelos. Como se viu, alguns modelos poderão ter valores do *AIC* tão próximos que pode não se justificar ou adequar a escolha de um único modelo. Pode-se, pois, escolher um modelo mais "adequado" (o que tiver menor *AIC*) e especificar um conjunto de modelos que poderão ser igualmente "bons" para os dados, sendo desejável, investigá-los numa segunda etapa.

5.4.2 Testes de qualidade de ajuste

Depois de estimados os parâmetros das distribuições teóricas pelos diferentes métodos apresentados e considerados os valores para a função de log-verosimilhança e para o critério *AIC*, pretende-se verificar o grau de ajustamento dessas distribuições relativamente às amostras (podendo tirar-se conclusões da comparação dos mesmos). Os testes de qualidade de ajuste utilizados, os testes de qui-quadrado e de Kolmogorov-Smirnov (K-S) (ver apêndice A1), permitem verificar hipóteses acerca da *forma* da distribuição da população de onde provém uma qualquer amostra ou avaliar se diferentes amostras são provenientes de uma população comum.

O ajustamento das distribuições foi feito utilizando as estimativas obtidas em 5.4.1. Depois de se formular a seguinte hipótese nula:

H_0 : a amostra provém da distribuição utilizada,

os resultados obtidos para a teste de χ^2 e para o teste K-S $(\alpha = 0{,}05)$, foram os seguintes:

normal	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	49,0677	13	0,0000	Rej. H_0	0,1226	177	0,1022	Rej. H_0
Idade2	27,2051	13	0,0117	Rej. H_0	0,0534	319	0,0761	Ac. H_0
Idade3	16,2035	12	0,2821	Ac. H_0	0,047	234	0,0889	Ac. H_0
Idade4	5,5871	7	0,5871	Ac. H_0	0,075	80	0,1521	Ac. H_0
Idade5	*	*	*	*	0,1547	59	0,1771	Ac. H_0

Tabela5.11 – Testes de qui-quadrado e de Kolmogorov-Smirnov para a distribuição normal com α=0,05.

lognormal	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	69,8231	13	0,000	Rej. H_0	0,1412	177	0,1022	Rej. H_0
Idade2	14,8384	13	0,3180	Ac. H_0	0,0297	319	0,0761	Ac. H_0
Idade3	9,9477	12	0,6205	Ac. H_0	0,0270	234	0,0889	Ac. H_0
Idade4	4,752	7	0,6900	Ac. H_0	0,0582	80	0,1521	Ac. H_0
Idade5	*	*	*	*	0,1660	59	0,1771	Ac. H_0

Tabela5.12 – Testes de qui-quadrado e de Kolmogorov-Smirnov para a distribuição lognormal com α=0,05.

gama	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	55,6681	12	0,000	Rej. H_0	0,1296	177	0,1022	Rej. H_0
Idade2	11,2939	12	0,5039	Ac. H_0	0,0263	319	0,0761	Ac. H_0
Idade3	7,8978	11	0,7224	Ac. H_0	0,0244	234	0,0889	Ac. H_0
Idade4	4,9900	6	0,5451	Ac. H_0	0,0512	80	0,1521	Ac. H_0
Idade5	*	*	*	*	0,1199	59	0,1771	Ac. H_0

Tabela5.13 – Testes de qui-quadrado e de Kolmogorov-Smirnov para a distribuição gama com α=0,05.

Burr1	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	26,424	12	0,0093	Rej. H_0	0,1044	177	0,1022	Rej. H_0
Idade2	14,2878	12	0,2827	Ac. H_0	0,0374	319	0,0761	Ac. H_0
Idade3	8,2973	11	0,6865	Ac. H_0	0,0264	234	0,0889	Ac. H_0
Idade4	5,1133	6	0,5294	Ac. H_0	0,0610	80	0,1521	Ac. H_0
Idade5	*	*	*	*	0,1316	59	0,1771	Ac. H_0

Tabela5.14– Testes de qui-quadrado e de Kolmogorov-Smirnov para a distribuição Burr1 com α=0,05.

Burr2	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1*								
Idade2	12,8646	11	0,3023	Ac. H_0	0,0253	319	0,0761	Ac. H_0
Idade3	7,5488	10	0,6728	Ac. H_0	0,0238	234	0,0889	Ac. H_0
Idade4	4,9354	5	0,4238	Ac. H_0	0,0529	80	0,1521	Ac. H_0
Idade5	**	**	**	**	0,1154	59	0,1771	Ac. H_0

Tabela5.15– Testes de qui-quadrado e de Kolmogorov-Smirnov para a distribuição Burr2 com α=0,05. **Não foi possível obter quaisquer estimativas pelos métodos utilizados.

* Não se efectuaram testes para o ajustamento da distribuição Burr2 à Idade1 uma vez que não foi possível obter estimativas dos parâmetros desta distribuição por qualquer um dos métodos de estimação aqui utilizados.

**Em virtude do tamanho da amostra correspondente à idade5 ser $n = 59$, sendo a amplitude da mesma "extensa", torna-se difícil a utilização do teste de qui-quadrado; deste modo são apresentados apenas os resultados obtidos pelo teste de Kolmogorov-Smirnov para esta amostra.

Da comparação dos resultados obtidos para os dois testes de ajustamento, convém salientar que para a Idade2, enquanto o teste de qui-quadrado rejeita H_0 para a distribuição normal (repare-se que o valor de p-value é 0,0117, isto é, H_0 seria aceite para $\alpha = 0{,}01$), o teste K-S aceita esta hipótese para esta mesma distribuição. Para todos os restantes casos os dois testes estão em consonância, aceitando H_0 para as diferentes distribuições. Estes resultados diferem, apenas, nalguns casos quanto à escolha da distribuição mais "adequada", os quais se citarão nas conclusões que se seguem.

- Idade1: Os testes de ajustamento rejeitam H_0 para qualquer uma das distribuições apresentadas;
- Idade2: Tanto o teste de χ^2 como o teste de K-S evidenciam o melhor ajustamento da distribuição gama a esta idade, estando de acordo com os resultados obtidos para a função de verosimilhança e para o *AIC*;
- Idade3: Enquanto o teste χ^2 privilegia o ajustamento da distribuição gama a esta amostra, o teste K-S privilegia o ajustamento à distribuição Burr2. Contudo, os

valores de $D_{obs.}$ para estas duas distribuições estão muito próximos. O melhor valor da função de log-verosimilhança foi obtido para a distribuição Burr2, enquanto o melhor valor do *AIC* foi para a distribuição lognormal;

- Idade4: O teste de χ^2 privilegia o ajustamento da distribuição lognormal a esta amostra, enquanto o teste de K-S privilegia o ajustamento da distribuição gama. Foi a para a distribuição Burr2 que se obtiveram os melhores resultados para a função de log-verosimilhança e para o *AIC*;
- Idade5: O teste K-S (único utilizado por razões referidas) evidencia o ajustamento da distribuição Burr2 a esta amostra. Contudo o resultado obtido para $D_{obs.}$ é muito próximo do resultado de $D_{obs.}$ obtido para a distribuição gama. Apesar deste facto, como se verá na análise gráfica, não parece que qualquer uma destas distribuições se ajuste adequadamente a estes dados. Também neste caso, os melhores resultados para a função de log-verosimilhança e para o *AIC* foram obtidos para a distribuição Burr2.

As distribuições apresentados parecem biologicamente significativas para este conjunto de dados, uma vez que, exceptuando a Idade1, se aceita a hipótese das amostras resultarem das diferentes distribuições. Em particular, a distribuição gama é a que aparece referenciada como distribuição mais "adequada" em todos estes casos. Ora pelo teste de χ^2, ora pelo teste de K-S, ou pelos dois. Parecem pois estarem encontradas distribuições alternativas à distribuição normal, o que está de acordo com o objectivo principal deste trabalho. Deve-se, porém, ter em conta casos como o da Idade1.

5.4.3 Misturas

Como foi referido na alínea anterior, a Idade1 não é ajustada por qualquer das distribuições apresentadas. Para se tentar ter alguma ideia sobre a distribuição a ajustar apresenta-se de seguida o histograma correspondente a esta idade.

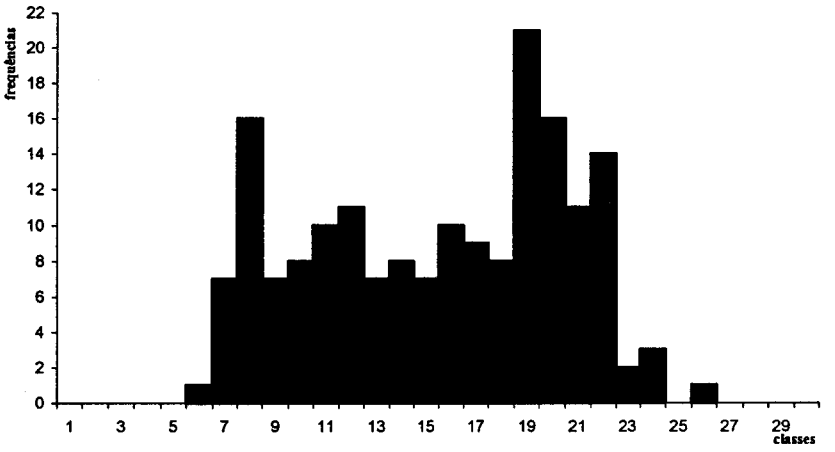

Figura5.1 - Histograma referente à Idade1.

De facto, a análise deste gráfico não sugere o ajustamento de qualquer uma das distribuições apresentadas até aqui. Optou-se por uma mistura de distribuições normais. Com base na técnica desenvolvida por Cabral (1987a) introduzida na secção 2.2.6, pretende-se identificar a existência ou não de mistura e, em caso afirmativo, determinar o número de componentes e estimar os parâmetros da mesma. Obtiveram-se os seguintes quadros de resultados:

- Mistura em escala

n	T_n	c_n	Decisão
		1,221318 $(\alpha = 0{,}01)$	
177	0,499484	1,156484 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
		1,121921 $(\alpha = 0{,}1)$	

Tabela5.16 – Teste de existência de mistura em escala de distribuições normais relativamente à Idade1.

Como o valor da estatística de teste T_n é menor que o valor crítico c_n para $\alpha = 0{,}1$, rejeita-se a hipótese nula da existência de uma mistura em escala de distribuições normais para a idade1.

- Mistura em localização

n	T_n	c_n	decisão
		1,000619 $(\alpha = 0{,}01)$	
177	1,062108	1,000438 $(\alpha = 0{,}05)$	mistura $(k = 2)$
		1,000341 $(\alpha = 0{,}1)$	
subamostras			
		1,000951 $(\alpha = 0{,}01)$	
75	1,010274	1,000672 $(\alpha = 0{,}05)$	mistura $(k = 2)$
		1,000524 $(\alpha = 0{,}1)$	
		1,000815 $(\alpha = 0{,}01)$	
102	1,011663	1,000577 $(\alpha = 0{,}05)$	mistura $(k = 2)$
		1,000449 $(\alpha = 0{,}1)$	

Tabela5.17 – Teste de existência de mistura em localização de distribuições normais relativamente à Idade1.

Numa primeira observação é sugerido que a amostra em estudo se divida em duas subamostras, onde, para cada uma delas, continua a haver indicações da existência de mistura em localização. Um modelo de mistura em localização de duas distribuições normais tem 5 parâmetros $\varphi_1, \varphi_2, \lambda_1, \lambda_2$ e δ, que se podem reduzir a 4 uma vez que $\varphi_1 + \varphi_2 = 1$, donde $\varphi = \varphi_1$ e $\varphi_2 = 1 - \varphi$. O número máximo de parâmetros considerados até aqui foram 4. Tendo-se em conta o princípio de parcimónia, faz-se um estudo do ajustamento desta mistura às observações da idade1.

Utilizando as estimativas obtidas pelo método dos momentos como valores iniciais, para a aplicação do algoritmo EMMC $(M = 30)$ com Newton-Raphson (considerando os dados não agrupados), obtêm-se as seguintes estimativas de MV, $\hat{\varphi}_1 = 0{,}5734$, $\hat{\varphi}_2 = 0{,}4266$, $\hat{\lambda}_1 = 19{,}2071$, $\hat{\lambda}_2 = 10{,}3465$ e $\hat{\delta} = 2{,}4895$. O valor da função de log-verosimilhança é $V(\hat{\Pi} \mid x) = -514{,}4432$ e o critério de informação de Akaike $AIC = 1036{,}8864$.

Comparando estes resultados com os obtidos para as outras distribuições pode-se concluir que estes são os melhores, sendo por isso de esperar um ajustamento de uma

mistura em localização de duas distribuições normais para a Idade1. Para o confirmar apresenta-se os resultados dos testes de qualidades de ajuste.

mistura	$\chi^2_{obs.}$	Gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	16,2313	11	0,1327	Ac. H_0	0,0445	177	0,1022	Ac. H_0

Tabela5.18 – Testes de qui-quadrado e de Kolmogorov-Smirnov para uma mistura em localização de 2 distribuições normais ajustada à Idade1 (α=0,05).

Pelos resultados apresentados na tabela5.18 pode-se concluir que a mistura em localização de duas distribuições normais se ajusta melhor a esta amostra do que as outras distribuições apresentadas (veja-se o valor de p-value).

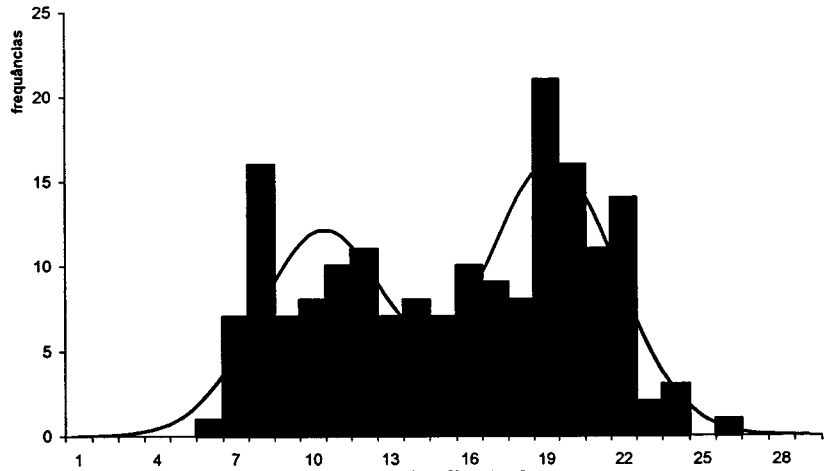

Figura5.2 - Ajustamento de uma mistura em localização de duas distribuições normais à Idade1.

De facto, visualmente, parece ser esta a distribuição que melhor se ajusta à Idade1. Convém relembrar que, fazendo $1-\hat{\varphi}$ para a proporção da mistura, trocando a ordem dos parâmetros de localização e mantendo o valor do parâmetro de escala, os valores assim obtidos correspondem a outra estimativa dos parâmetros da mistura, mas que corresponde à mesma distribuição (ao mesmo modelo). O algoritmo EM com Newton-Raphson garante que se atinja o máximo global.

Apesar de se ter testado o ajustamento de uma mistura de duas distribuições normais, para qualquer das subamostras apresentadas na tabela5.17, $k=2$, isto é, deve-se continuar a fazer o teste de existência de mistura em localização até que se tenha $k=1$ para todas as subamostras. Assim,

n		T_n	c_n	Decisão
			1,001319 $(\alpha = 0{,}01)$	
	39	1,000364	1,000932 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
75			1,000726 $(\alpha = 0{,}1)$	
			1,001373 $(\alpha = 0{,}01)$	
	36	1,000534	1,00097 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
			1,000756 $(\alpha = 0{,}1)$	
			1,000977 $(\alpha = 0{,}01)$	
	71	1,004684	1,000691 $(\alpha = 0{,}05)$	mistura $(k = 2)$
102			1,000538 $(\alpha = 0{,}1)$	
			1,001479 $(\alpha = 0{,}01)$	
	31	1,000521	1,001046 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
			1,000815 $(\alpha = 0{,}1)$	

Tabela5.19 – Continuação do teste de existência de mistura em localização de distribuições normais relativamente à Idade1.

Como o número de componentes da mistura é dado pelo número de subamostras com $k = 1$, a tabela anterior indica, para já, o ajustamento a uma mistura de três distribuições normais. Como para a subamostra $n = 71$, $k = 2$, então

n		T_n	c_n	decisão
			1,001615 $(\alpha = 0{,}01)$	
	26	0,999033	1,001142 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
71			1,00089 $(\alpha = 0{,}1)$	
			1,001228 $(\alpha = 0{,}01)$	
	45	0,998769	1,000868 $(\alpha = 0{,}05)$	não mistura $(k = 1)$
			1,000676 $(\alpha = 0{,}1)$	

Tabela5.20 – Conclusão do teste de existência de mistura em localização de distribuições normais relativamente à Idade1.

Agora sim, pode-se concluir que o teste de existência de misturas indica o ajustamento de uma mistura em localização de cinco distribuições normais. Uma distribuição nestas condições envolve 10 parâmetros, o que poderá levar a um sobreajustamento e consequentemente à perca de precisão. A título de exemplo, e por se achar razoável, tenta-se o ajustamento a uma mistura em localização de três distribuições normais (6 parâmetros). Aplicando o algoritmo EMMC (como em cima), obteve-se $\hat{\varphi}_1 = 0{,}2150$, $\hat{\varphi}_2 = 0{,}3157$, $\hat{\varphi}_3 = 1 - \hat{\varphi}_1 - \hat{\varphi}_2 = 0{,}4693$, $\hat{\lambda}_1 = 14{,}4773$, $\hat{\lambda}_2 = 9{,}4040$, $\hat{\lambda}_3 = 19{,}9679$ e $\hat{\delta} = 1{,}9565$, $V\left(\hat{\mathbf{\Pi}} \mid x\right) = -508{,}8158$ e $AIC = 1029{,}6315$. Estes dois últimos coeficientes são os mais baixos até aqui apresentados, ou seja tudo indica que este modelo de mistura é o que melhor se ajusta a esta amostra. De facto, aceita-se a hipótese do ajustamento de uma mistura em localização de três distribuições normais à Idade1.

mistura	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1	12,7685	9	0,1734	Ac. H_0	0,0337	177	0,1022	Ac. H_0

Tabela5.21 – Testes de qui-quadrado e de Kolmogorov-Smirnov para uma mistura em localização de 3 distribuições normais ajustada à Idade1 (α=0,05).

Os resultados da tabela5.21 estão de acordo com os valores obtidos para a função de verosimilhança e para o *AIC*.

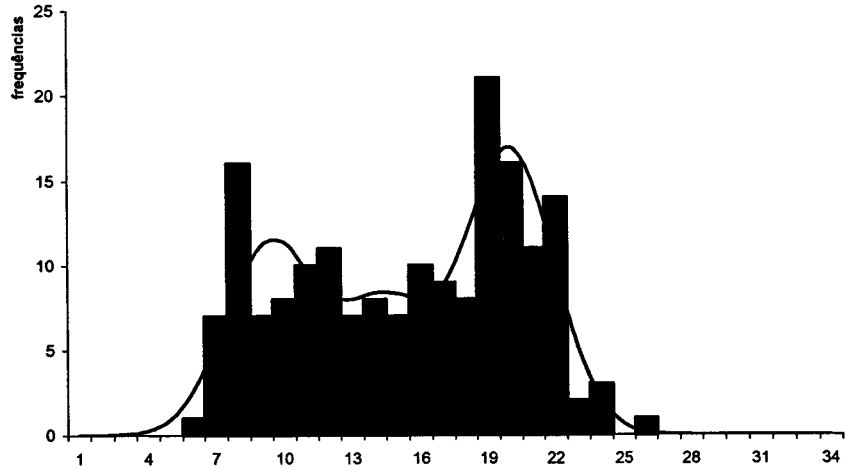

Figura5.3 - Ajustamento de uma mistura em localização de três distribuições normais à Idade1

Graficamente pode-se observar que de facto este é o modelo, de entre os apresentados, que melhor se ajusta a esta amostra.

Por raciocínio análogo ao utilizado para a Idade1, obtiveram-se os seguintes resultados para o ajustamento de misturas em localização de distribuições normais ao resto das idades.

mistura normais	nº componentes	$V(\hat{\Pi} \mid x)$	*AIC*
Idade2	2	-847,6752	1703,3505
Idade3	2	-612,1161	1232,2322
Idade4	2	-203,3246	414,6493
Idade5	2	-165,5910	339,1822
Idade5	3	-160,4028	332,8056

Tabela5.22 – Número de componentes da mistura de distribuições normais relativamente a cada idade, valores da função de log-verosimilhança e do AIC.

Mistura normais	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade2	14,4627	11	0,2084	Ac. H_0	0,0332	319	0,0761	Ac. H_0
Idade3	9,3751	10	0,4969	Ac. H_0	0,0275	234	0,0889	Ac. H_0
Idade4	4,9598	5	0,4208	Ac. H_0	0,0773	80	0,1521	Ac. H_0
Idade5 2 componentes de mistura	*	*	*	*	0,0951	59	0,1771	Ac. H_0
Idade5 3 componentes de mistura	*	*	*	*	0,0573	59	0,1771	Ac. H_0

Tabela5.23 – Testes de qui-quadrado e de Kolmogorov-Smirnov (K-S) para mistura em localização de distribuições normais relativamente a cada Idade. *Por razões já apontadas não é conveniente a aplicação do teste de qui-quadrado.

Repare-se que, o ajustamento de um modelo de misturas em localização de distribuições normais dá os melhores resultados para a Idade1 e Idade5. Apesar de se obterem bons resultados para a mistura de 2 distribuições normais, estes resultados melhoram ao se considerar a mistura de 3 destas distribuições. Resta saber se tal facto não resulta num sobreajustamento e consequentemente à perda de precisão no que diz respeito à inferência. Infelizmente, pela razão explicada na página 24, não é possível fazer um teste de razão de verosimilhanças para decidir se a terceira componente da mistura é ou não significativa.

Analogamente ao ajustamento de misturas de distribuições normais, fazendo $\ln Y$ em vez de Y (prop. 2.1 pág.23), considera-se o ajustamento a misturas de distribuições lognormais. O teste rejeita a hipótese de misturas em escala, obtendo-se os seguintes resultados para as misturas em localização (por razões referidas consideram-se três o número máximo de componentes da mistura),

mistura lognormais	nº componentes	$V(\hat{\Pi} \mid y)$	AIC
Idade1	2	-520,5804	1049,1609
Idade1	3	-511,7042	1035,4083
Idade2	2	-843,976	1695,9510
Idade3	2	-611,8097	1231,6197
Idade4	2	-202,9088	413,8176
Idade5	2	-164,3260	336,6518
Idade5	3	-160,1205	332,2411

Tabela5.24 – Número de componentes para uma mistura em localização de distribuições log-normais relativamente a cada Idade, valores da função log-verosimilhança e do AIC.

mistura lognormais	$\chi^2_{obs.}$	gl	p-value	decisão	$D_{obs.}$ K-S	n	$D_{0,05}$	decisão
Idade1 2 componentes de mistura	37,3435	11	0,0001	Rej. H_0	0,0767	177	0,1022	Ac. H_0
Idade1 3 componentes de mistura	14,8438	9	0,0953	Ac. H_0	0,0390	177	0,1022	Ac. H_0
Idade2	10,2752	11	0,5058	Ac. H_0	0,0217	319	0,0761	Ac. H_0
Idade3	7,8317	10	0,6453	Ac. H_0	0,0221	234	0,0889	Ac. H_0
Idade4	4,7017	5	0,4534	Ac. H_0	0,0655	80	0,1521	Ac. H_0
Idade5 2 componentes de mistura	*	*	*	*	0,0785	59	0,1771	Ac. H_0
Idade5 3 componentes de mistura	*	*	*	*	0,0491	59	0,1771	Ac. H_0

Tabela5.25 – Testes de qui-quadrado e de Kolmogorov-Smirnov (K-S) para mistura em localização de distribuições lognormais relativamente a cada Idade. *Por razões já apontadas não é conveniente a aplicação do teste de qui-quadrado.

Comparando os resultados obtidos para misturas de distribuições normais e para misturas de distribuições lognormais (o número de componentes de cada uma destas misturas, consideradas para cada amostra, é idêntico nos dois casos), pode-se concluir que:

- Idade1: Relembre-se que esta amostra tem assimetria negativa e que a sua distribuição não segue o padrão das demais amostras estudadas. Tanto os resultados obtidos para o *AIC*, como para os testes de qualidade de ajustamento, favorecerem o ajustamento de uma mistura em localização de 3 distribuições normais.
- Idade2, Idade3 e Idade4: Os resultados obtidos favorecem o ajustamento de uma mistura em localização de 2 distribuições lognormais.
- Idade5: Os testes apresentados deixam adivinhar, embora ligeiramente, o melhor ajustamento para uma mistura em localização de 3 distribuições lognormais.

Para qualquer uma das idades os modelos de misturas em localização aqui apresentados evidenciam bons resultados para o ajustamento. Em comparação com os resultados obtidos para as outras distribuições, convém salientar que os testes prevêem o melhor ajustamento relativamente a modelos de misturas, para a Idade1 (em particular uma mistura de 3 distribuições normais) e para a Idade5 (mistura em localização 3 distribuições lognormais). Em relação à Idade2, os testes de ajustamento favorecem ligeiramente o ajustamento de uma mistura em localização de 2 distribuições lognormais, enquanto os valores obtidos para a função de log-verosimilhança são aproximadamente iguais o valor do *AIC* é menor para a distribuição gama (recorde-se que o *AIC* aumenta com o número de parâmetros, a mistura tem mais 3 parâmetros do que a distribuição gama o que pode levar a um sobreajustamento).

5.4.4 Análise gráfica

Apresentam-se de seguida algumas conclusões do ajustamento das diferentes distribuições a cada uma das idades após visualização gráfica. Estes gráficos são constituídos pelos histogramas das amostras referentes a cada idade e pelas curvas das f.d.p.'s ajustadas correspondentes.

- Idade1 (figuras5.4-5.11): Note-se que o comportamento desta amostras não segue o padrão apresentado pelas outras amostras. Podem-se apontar algumas justificações para tal facto: um erro de amostragem; mistura de dois grupos sujeitos a condições de crescimento diferentes; os sexos dos peixes se encontrarem misturados e terem tamanhos médios muito diferentes nesta idade; dois pulsos de recrutamento no mesmo ano, isto é, como o crescimento dos peixes nos primeiros meses é muito rápido poderão ser agrupados peixes que são classificados como tendo um ano mas que têm uma idade diferente em meses e, por isso, diferem bastante no comprimento médio.

  Como se pode observar são, de facto, os modelos de mistura em localização com 3 componentes que parecem ajustar-se aos dados. O facto destas distribuições apresentarem 6 parâmetros não parece sobrecarregar o ajustamento e consequentemente levar a uma perda de precisão (uma vez que apresentam os valores mais baixos do *AIC*). Como foi referido, os diferentes testes rejeitaram a hipótese de ajustamento a distribuições que não fossem misturas. No caso da mistura de 3 distribuições lognormais, $AIC = 1035{,}4083$ e $\text{p-value} = 0{,}0953$, enquanto para uma mistura de 3 distribuições normais $AIC = 1029{,}6315$ e $\text{p-value} = 0{,}1734$. Graficamente, qualquer um destes ajustamentos parece razoável (figura5.8 e 5.9).

  Repare-se que o gráfico da distribuição Burr1 (figura5.11), reflecte a assimetria negativa da amostra (esta distribuição permite uma alteração da assimetria nos dois sentidos, seccção2.2.4). Apesar da distribuição Burr2 também permitir uma alteração na assimetria no sentido negativo (ver secção2.2.5), não foi possível a obtenção de estimativas para os parâmetros desta distribuição a Idade1.

É possível apresentar um gráfico com valores dos parâmetros que, apesar de anularem aproximadamente o gradiente da função de log-verosimilhança, não permitem que os algoritmos utilizados convirjam; poderá tratar-se de um ponto de sela;

- Idade2 (figuras5.12-5.18): Esta amostra apresenta assimetria positiva; como tal, são as distribuições que possibilitam este desvio em relação à distribuição normal que melhor se ajustam. Repare-se que os gráficos das distribuições gama (figura5.16), Burr2 (figura5.18) e mistura em localização de duas lognormais (figura5.15) são praticamente iguais. Analiticamente os melhores resultados foram obtidos para a distribuição gama (com $AIC = 1693{,}2326$) e para a mistura referida (com $AIC = 1695{,}9510$). Relembre-se que a Burr2 (com $AIC = 1697{,}4192$) e a mistura têm mais um parâmetro do que a gama, podendo evidenciar um maior sobreajustamento (indicado pelo maior valor do *AIC* para estas duas distribuições);

- Idade3 (figuras5.19-5.25): Relativamente a esta amostra os gráficos são todos muito similares. Repare-se que os testes de ajustamento dão resultados razoável para todas as distribuições. A que tem melhor *AIC* é a distribuição lognormal (figura5.21), contudo os testes de ajustamento privilegiam a distribuições gama (figura5.23) e Burr2 (figura5.25);

- Idade4 (figuras5.26-5.32): Esta amostra é de pequena dimensão. É o ajustamento das distribuições gama (figura5.30) e Burr2 (figura5.32) que mais evidenciam a assimetria positiva da amostra. Um destes ajustamentos parece o melhor para esta amostra. Contudo, como já foi referido, enquanto o teste de qui-quadrado privilegia o ajustamento à distribuição lognormal (figura5.28), o teste de Kolmogorov-Smirnov favorece o ajustamento à distribuição gama. É a distribuição Burr2 que apresenta o melhor valor para o *AIC*;

- Idade5 (figuras5.33-5.40): Recorde-se que os testes de ajustamento dão resultados razoável para todas as distribuições. Parece não haver dúvidas, graficamente, que são as misturas que melhor se ajustam à amostra correspondente à Idade5. Pode-se observar, que os gráficos correspondentes às misturas de 3 distribuições normais

(figura5.34) e de 3 distribuições lognormais (figura5.36) são praticamente idênticos, assim como, os valores do *AIC*. Os testes de ajustamento favorecem ligeiramente a mistura de lognormais.

Visualmente, é possível ter uma ideia do que foi dito em relação ao ajustamento analítico das distribuições. Relativamente à Idade1 (assimetria negativa) são de facto as misturas que melhor se parecem ajustar aos dados. Confirmam-se, assim, os resultados obtidos pelos testes de ajustamentos. Para as idades intermédias (assimetria positiva), o ajustamento de qualquer uma das distribuições parece razoável. Para a Idade5 (assimetria positiva e $n = 59$), contrariamente ao resultados do teste Kolmogorov-Smirnov, todas as distribuições consideradas, exceptuando os modelos de misturas, não parecem ajustar-se aos dados. Os resultados obtidos estão de acordo com o objectivo principal deste estudo, encontrar distribuições alternativas à distribuição normal para dados de comprimento à idade referentes à captura de pescado como base para a inferência, uma vez que todas as distribuições aqui apresentadas se comportam melhor, em termos de ajustamento, do que a distribuição normal.

Nas páginas que se seguem apresentam-se os gráficos referentes ao ajustamento das diferentes distribuições estudadas para cada Idade

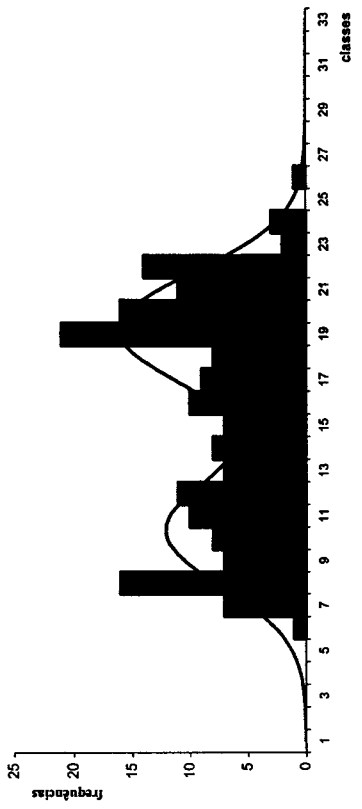

Figura5.5 - Ajustamento de uma mistura de duas dist. normais à Idade1

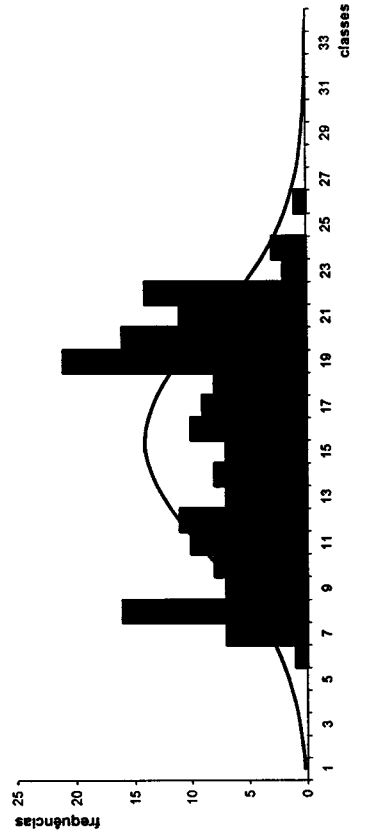

Figura5.4 - Ajustamento da distribuição normal à Idade1

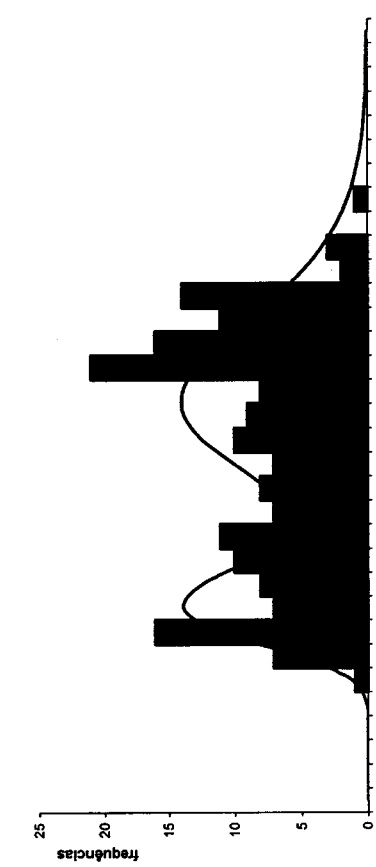

Figura5.7 - Ajustamento de uma mistura de duas dist. lognormais à Idade1

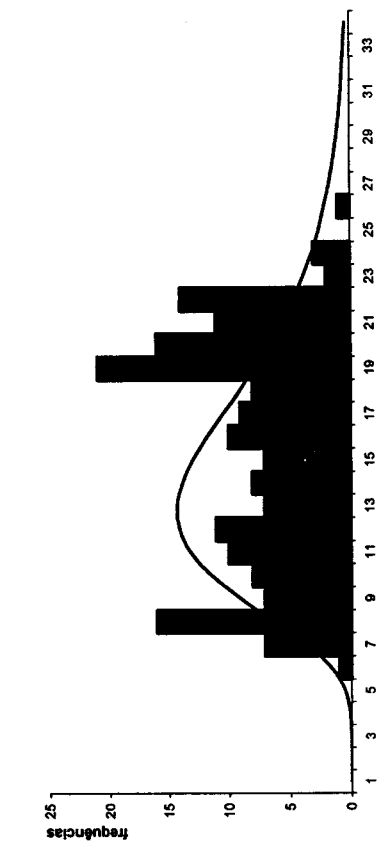

Figura5.6 - Ajustamento da distribuição lognormal à Idade1

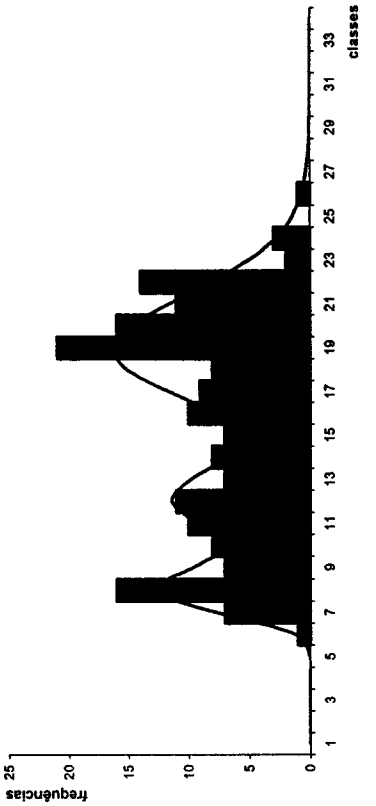

Figura5.9 - Ajustamento de uma mistura de três dist. lognormais à Idade1

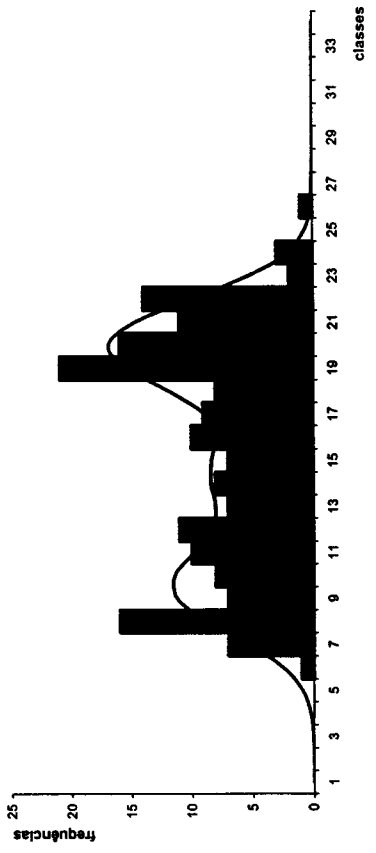

Figura5.8 - Ajustamento de uma mistura de três dist. normais à Idade1

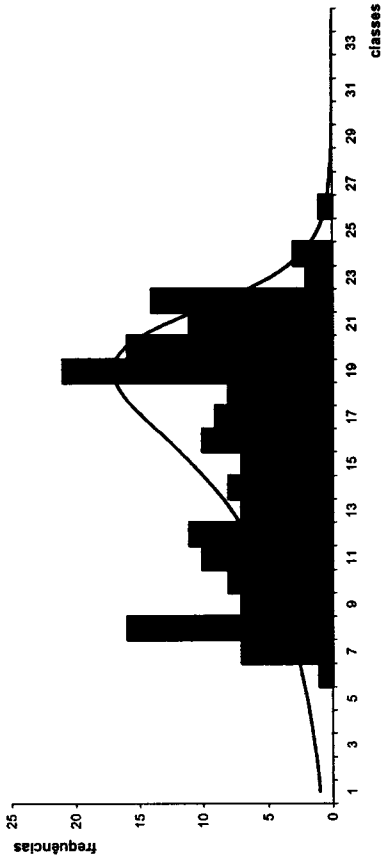

Figura5.11 - Ajustamento da distribuição Burr1 à Idade1

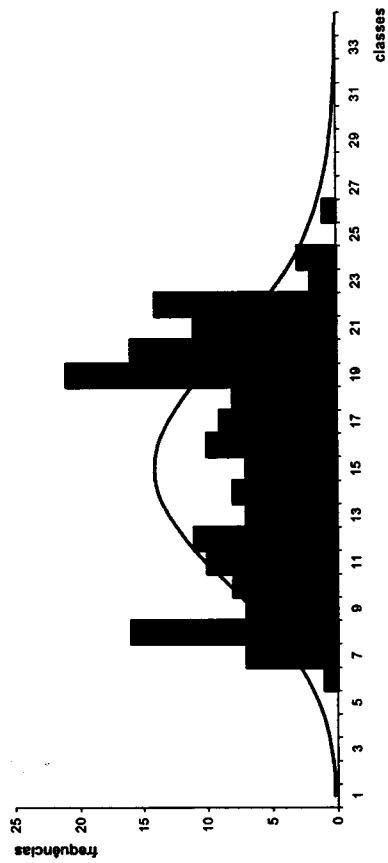

Figura5.10 - Ajustamento da distribuição gama à Idade1

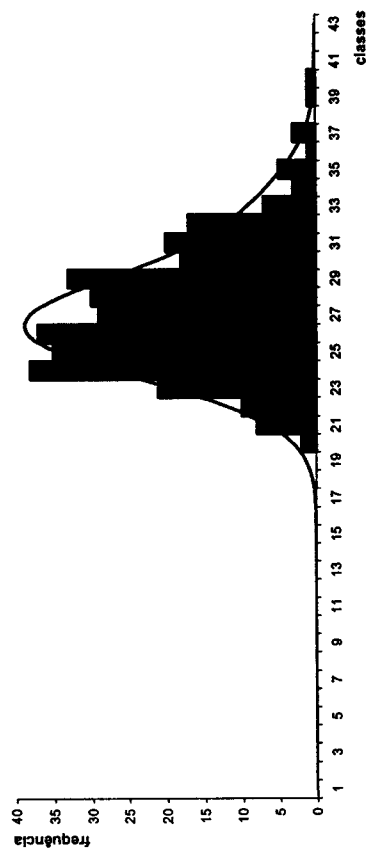

Figura5.13 - Ajustamento de uma mistura de duas dist. normais à Idade2

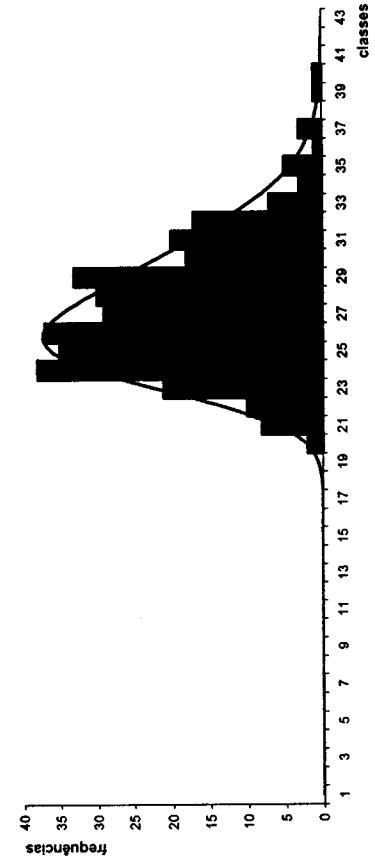

Figura5.15 - Ajustamento de uma mistura de duas dist. lognormais à Idade2

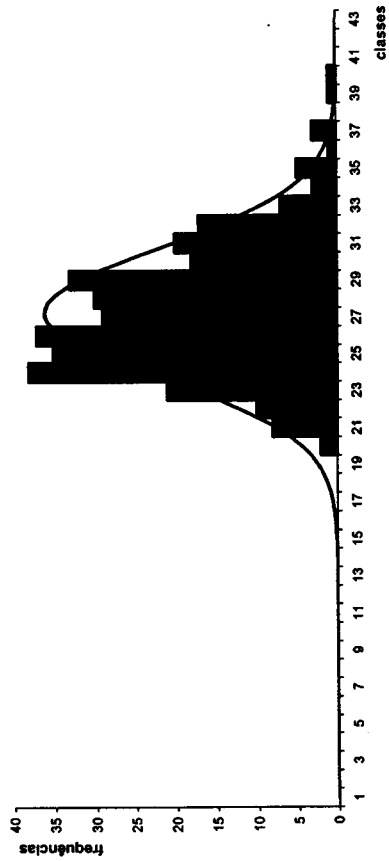

Figura5.12 - Ajustamento da distribuição normal à Idade2

Figura5.14 - Ajustamento da distribuição lognormal à Idade2

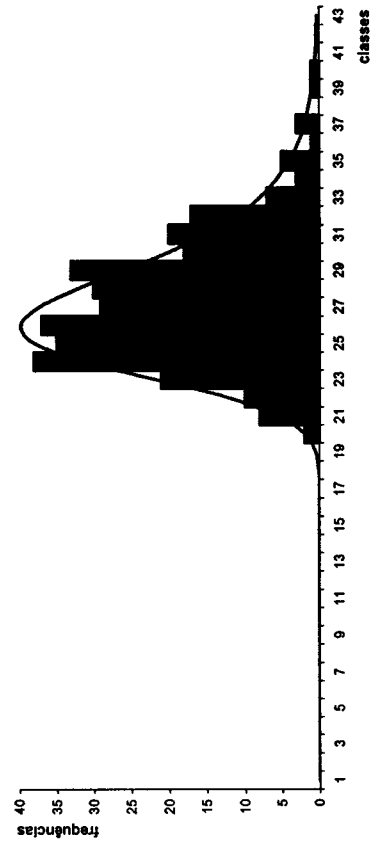

Figura5.17 - Ajustamento da distribuição Burr1 à Idade2

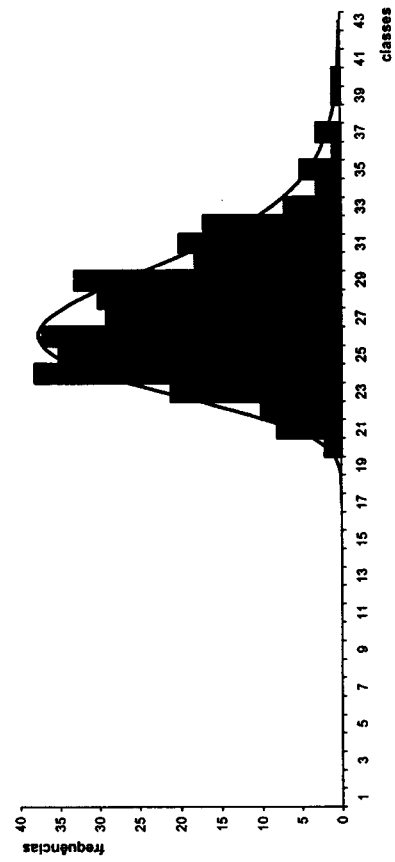

Figura5.16 - Ajustamento da distribuição gama à Idade2

Figura5.18 - Ajustamento da distribuição Burr2 à Idade2

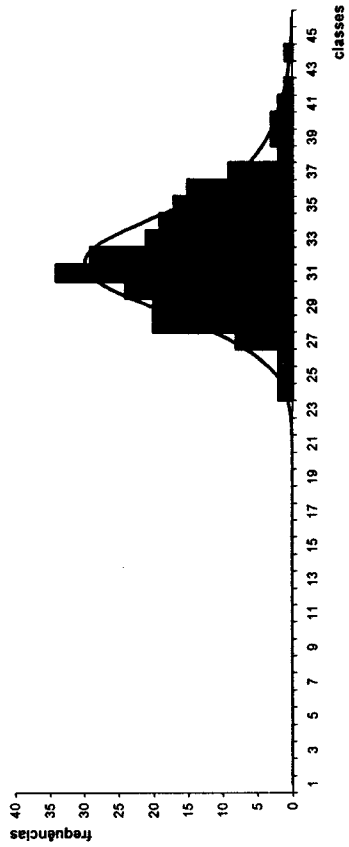

Figura5.20- Ajustamento de uma mistura de duas dist. normais à Idade3

Figura5.19 - Ajustamento da distribuição normal à Idade3

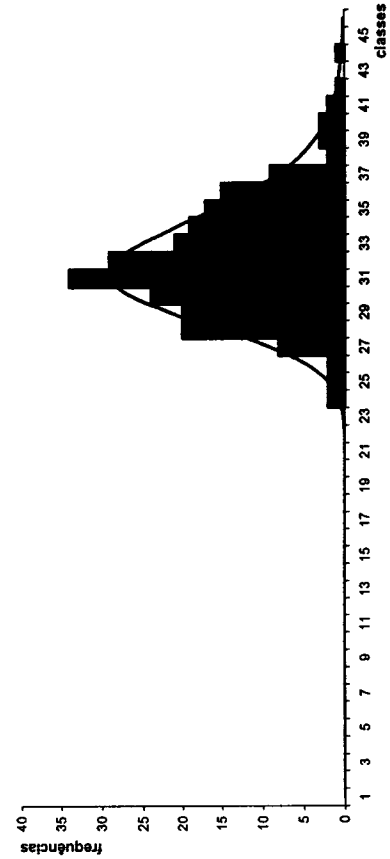

Figura5.22 - Ajustamento de uma mistura de duas dist. lognormais à Idade3

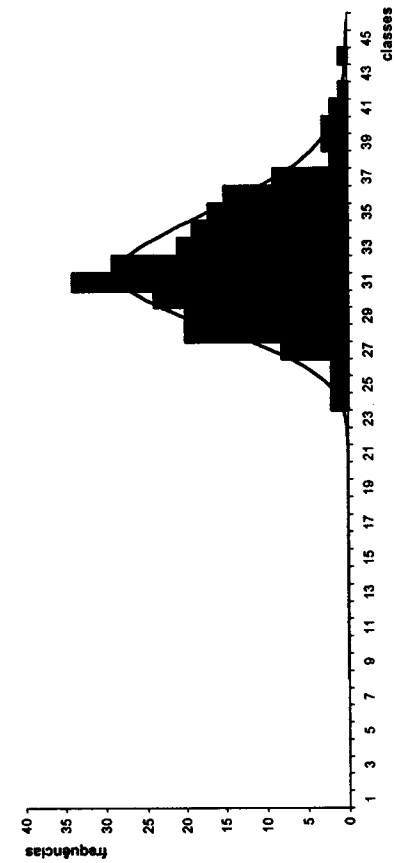

Figura5.21 - Ajustamento da distribuição lognormal à Idade3

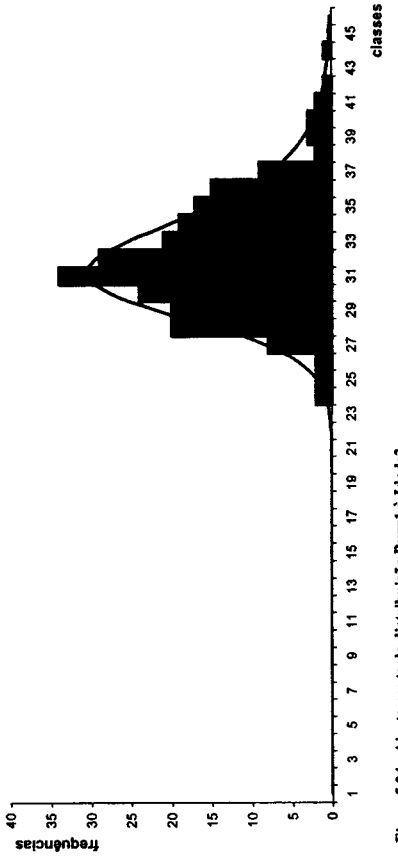

Figura5.24 - Ajustamento da distribuição Burr1 à Idade3

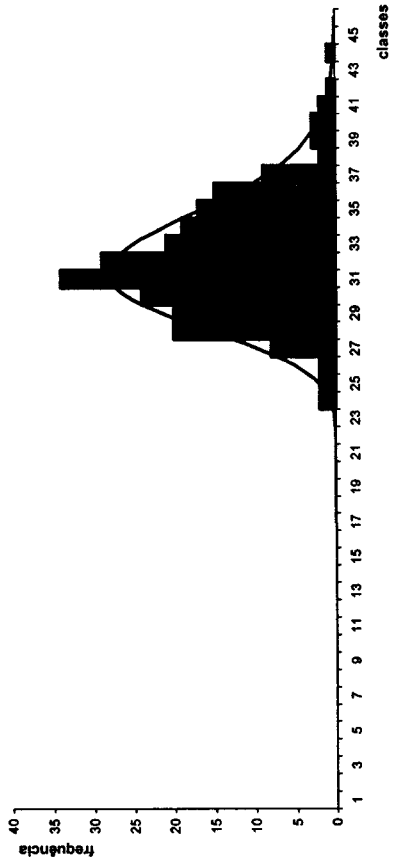

Figura5.23 - ajustamento da distribuição gama à Idade3

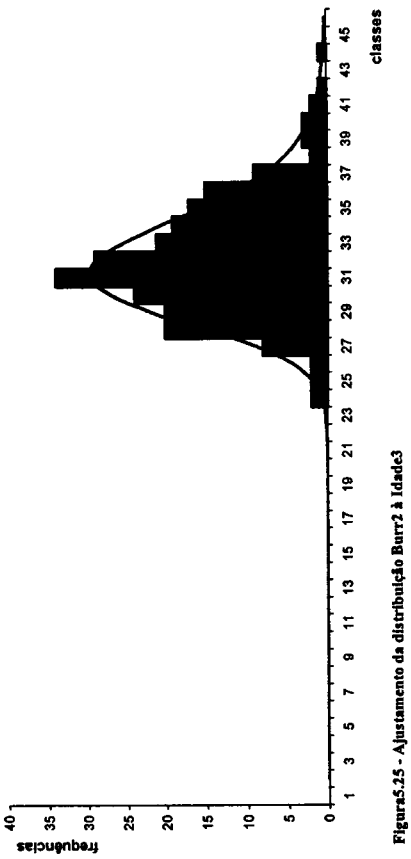

Figura5.25 - Ajustamento da distribuição Burr2 à Idade3

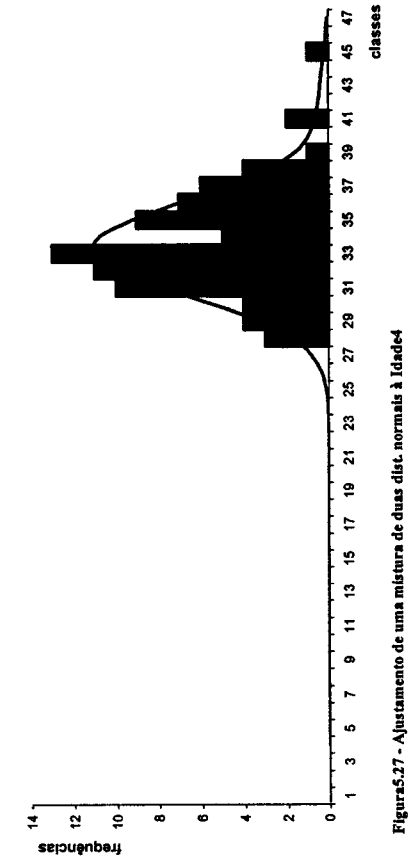

Figura5.27 - Ajustamento de uma mistura de duas dist. normais à Idade4

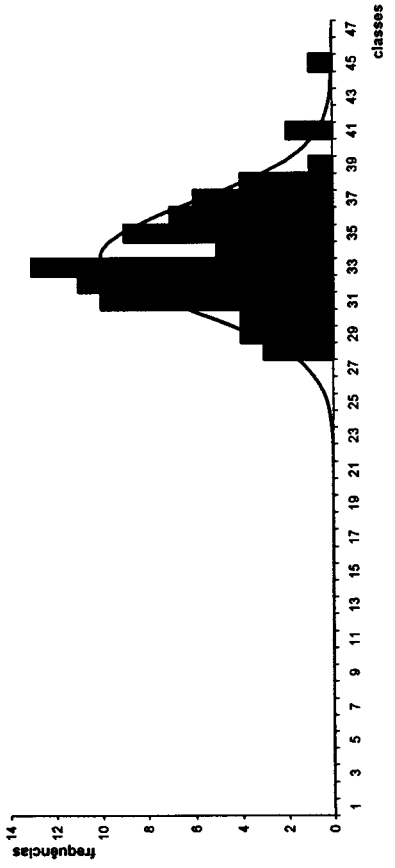

Figura5.26 - Ajustamento da distribuição normal à Idade4

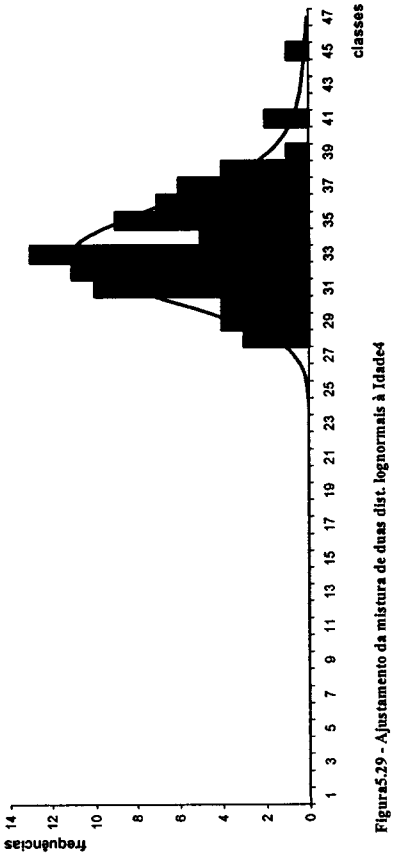

Figura5.29 - Ajustamento da mistura de duas dist. lognormais à Idade4

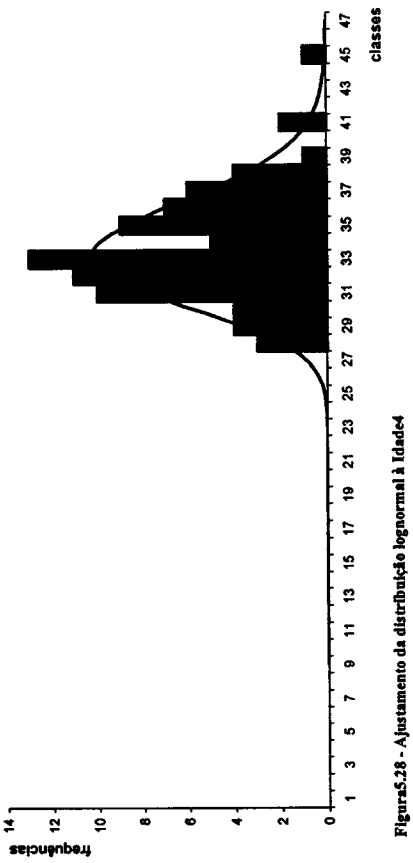

Figura5.28 - Ajustamento da distribuição lognormal à Idade4

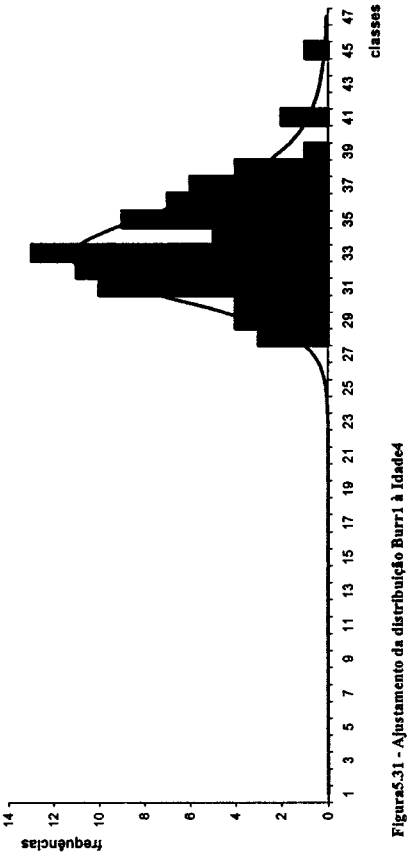

Figura5.31 - Ajustamento da distribuição Burr1 à Idade4

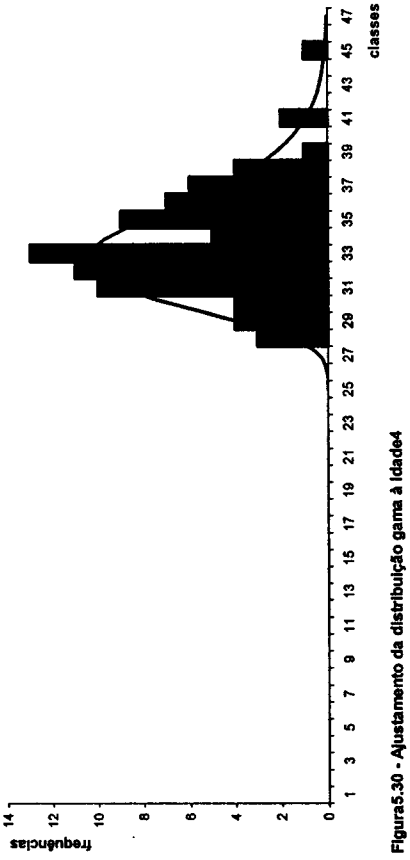

Figura5.30 - Ajustamento da distribuição gama à Idade4

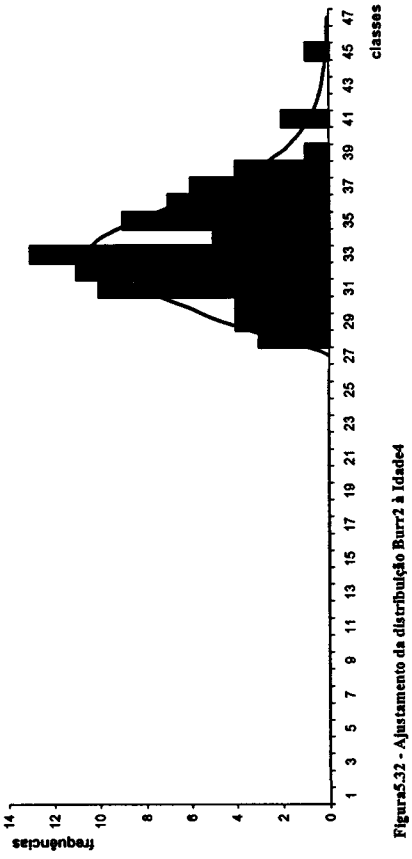

Figura5.32 - Ajustamento da distribuição Burr2 à Idade4

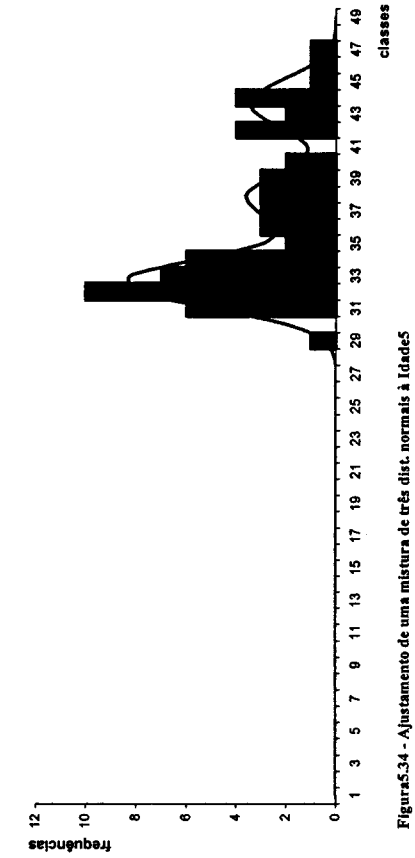

Figura5.34 - Ajustamento de uma mistura de três dist. normais à Idade5

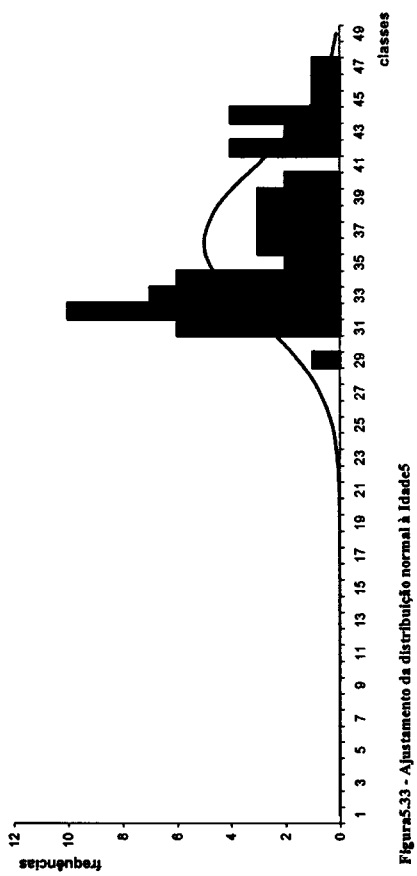

Figura5.33 - Ajustamento da distribuição normal à Idade5

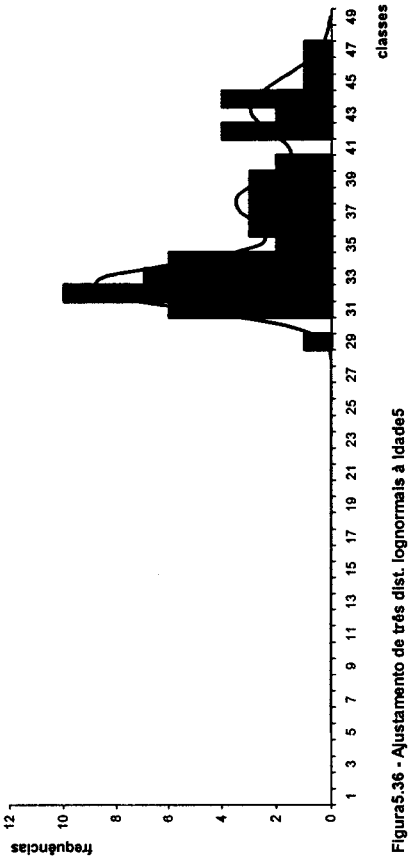

Figura5.36 - Ajustamento de três dist. lognormais à Idade5

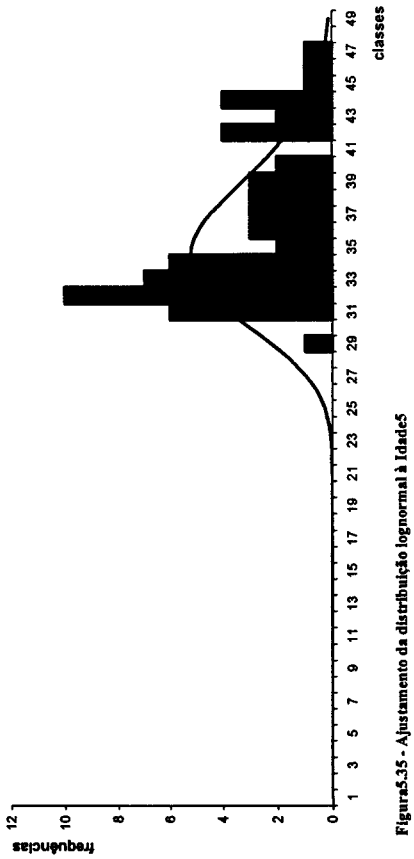

Figura5.35 - Ajustamento da distribuição lognormal à Idade5

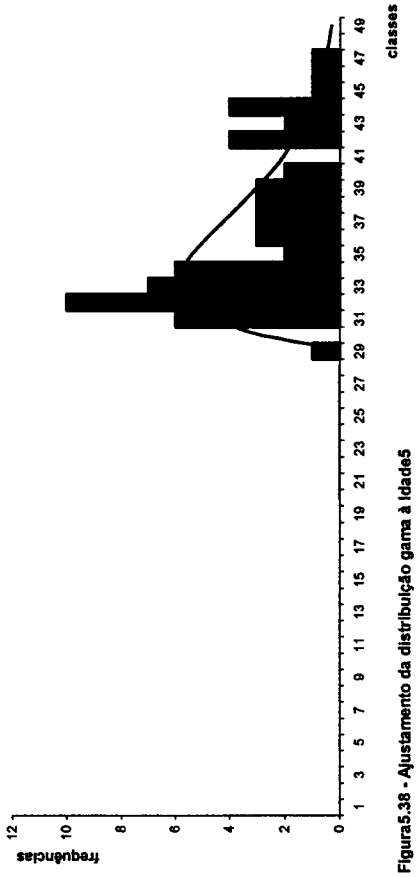

Figura5.38 - Ajustamento da distribuição gama à Idade5

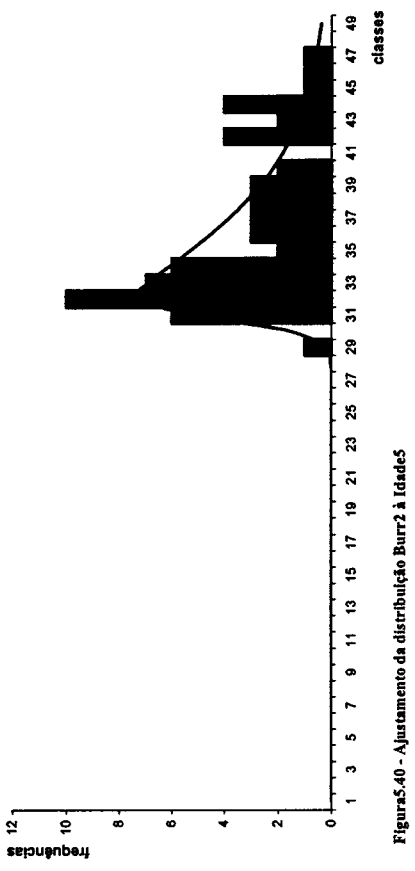

Figura5.40 - Ajustamento da distribuição Burr2 à Idade5

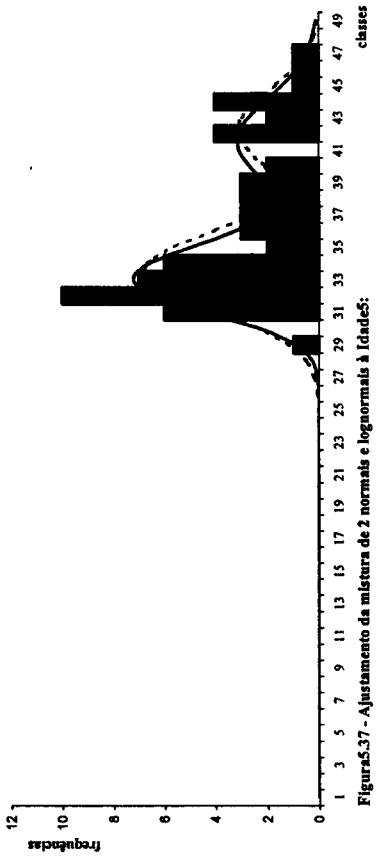

Figura5.37 - Ajustamento da mistura de 2 normais e lognormais à Idade5: tracejado (normal), negrito (lognormal)

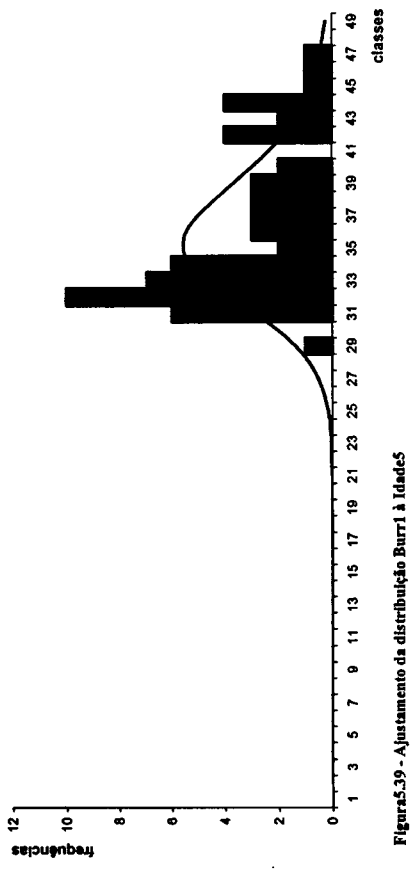

Figura5.39 - Ajustamento da distribuição Burr1 à Idade5

5.4.5 Intervalos de confiança

Cada um dos métodos de estimação pontual analisados permite associar a cada parâmetro populacional um estimador. A cada estimador estarão associadas tantas estimativas diferentes quantas as amostras que forem utilizadas para o seu cálculo. Na generalidade dos casos, nenhuma dessas estimativas coincidirá com o valor do parâmetro.

A grande limitação dos métodos de estimação pontual é a de não fornecerem qualquer informação relativa ao rigor ou à confiança das estimativas que através deles são obtidas. Esta dificuldade é ultrapassada recorrendo aos métodos de estimação por intervalos. Deve-se ter em conta que a precisão (que varia na razão inversa da sua amplitude) e o grau de confiança dos intervalo de confiança aumentam com o tamanho da amostra.

Apresentam-se de seguida, para as amostras da tabela1.1, os intervalos de confiança aproximados a 95% para os parâmetros das diferentes distribuições teóricas apresentadas. As estimativas foram obtidas através da aplicação do algoritmo EMMC ($M = 30$) após um período de "queima" de 500 iterações. Exceptuando para as misturas, os dados foram considerados agrupados.

normal	Idade1	Idade2	Idade3	Idade4	Idade5
μ	15,47 ± 0,75	27,21 ± 0,43	32,03 ± 0,46	33,60 ± 0,77	36,27 ± 1,21
σ^2	24,17 ± 5,24	12,39 ± 1,96	11,32 ± 1,99	9,97 ± 2,78	22,28 ± 8,01

Tabela5.26 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da distribuição normal.

Relativamente à distribuição normal, pode-se reparar que, como era de esperar, as amostras de menor dimensão têm maiores amplitudes do intervalo de confiança, podendo-se traduzir numa menor precisão destas estimativas.

No caso da mistura em localização de distribuições normais (tabela5.31, pag.141), de uma maneira geral as amplitudes dos intervalos são pequenas. No caso do número de componentes da mistura ser 2, a amplitude dos intervalos de confiança é maior para os maiores valores de λ_i $(i = 1,2)$. Porém, para as Idades 2,3 e 4, verifica-se que o peso de uma das componentes da mistura de duas componentes é próxima de zero e que o respectivo intervalo de confiança inclui zero (no caso das Idades 3 e 4), o que indicia que

esta componente poderá não ser significativa, isto é, que será desnecessário considerar um modelo de mistura de distribuições normais.

lognormal	Idade1	Idade2	Idade3	Idade4	Idade5
λ	2,67 ± 0,05	3,30 ± 0,01	3,46 ± 0,01	3,51 ± 0,02	3,58 ± 0,03
δ^2	0,1326 ± 0,0453	0,0162 ± 0,0029	0,0108 ± 0,0021	0,0087 ± 0,0027	0,0160 ± 0,0096

Tabela5.27 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da distribuição lognormal.

Para a distribuição lognormal, a amostra com maior amplitude é a correspondente à Idade1, esta amostra tem assimetria negativa e distribuição lognormal é assimétrica positiva. Para as restantes amostras a amplitude vai diminuindo com o aumento da dimensão da amostra.

Para as misturas em localização de distribuições lognormais (tabela5.32, pag.141), podem-se tirar conclusões idênticas ao caso da mistura de distribuições normais. Contudo, como o valor dos parâmetros é mais pequeno, a amplitude dos intervalos é menor. Porém, para as Idades 3 e 4, verifica-se que o peso de uma das componentes da mistura (de duas componentes) é próxima de zero e que o respectivo intervalo de confiança inclui zero, o que indicia que esta componente poderá não ser significativa, isto é, que será desnecessário considerar um modelo de mistura de lognormais. Para a Idade2, já tal não sucede, mas verifica-se, em contrapartida, que os valores do parâmetro de localização das duas componentes da mistura estão muito próximos um do outro, o que também aponta para o pouco interesse da consideração de um modelo de mistura de lognormais; esta é também a situação que se verifica na Idade5, quer para o modelo de mistura de duas lognormais, quer para o modelo de mistura de três lognormais.

gama	Idade1*	Idade2	Idade3	Idade4	Idade5
λ		16,67 ± 3,51	16,54 ± 8,71	24,42 ± 6,11	28,50 ± 1,15
δ		1,19 ± 0,50	0,73 ± 0,44	1,10 ± 0,93	3,06 ± 1,45
ρ		8,86 ± 6,55	21,22 ± 24,63	8,35 ± 12,40	2,52 ± 1,40

Tabela5.28 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da distribuição gama. *Intervalos de confiança obtidos pelo método de "Jackknife", em virtude de não se terem obtido estimativas de MV.

Consideremos agora a distribuição gama. Para a Idade1 (assimetria negativa) não se obtiveram estimativas de MV (como é sabido a distribuição gama apresenta sempre

assimetria positiva). Apresenta-se as estimativas obtidas pelo método dos momentos, $\hat{\lambda} = -40{,}15$, $\hat{\delta} = 0{,}46$ e $\hat{\rho} = 122{,}22$. Assim utilizou-se o método "Jackknife" (ver apêndice A3) para a estimação dos intervalos de confiança, e obtiveram-se estimativas $\hat{\lambda}_{Jk} = -40{,}28$, $\hat{\delta}_{Jk} = 0{,}455$ e $\hat{\rho}_{Jk} = 123{,}09$. Os erros padrão aproximados são, respectivamente, $SE(\hat{\lambda}_{Jk}) = 2{,}76$, $SE(\hat{\delta}_{Jk}) = 0{,}022$ e $SE(\hat{\rho}_{Jk}) = 11{,}87$. Como $t_{0{,}05[176]} = 1{,}9735$ então $\lambda : [-45{,}73 ; -34{,}84]$, $\delta : [0{,}412 ; 0{,}498]$ e $\rho : [99{,}65 ; 146{,}53]$.

Com excepção da Idade5 (amostra de menor amplitude), os intervalos de confiança apresentam grandes amplitudes para os parâmetros populacionais da distribuição gama. Tal facto pode-se dever à utilização de aproximações para a função digama na aplicação dos métodos iterativos. Repare-se que a ordem de grandeza da amplitude dos intervalos obtidos pelo método "Jackknife" é similar à dos restantes intervalos

Burr1	Idade1	Idade2	Idade3	Idade4	Idade5*
λ	21,02 ± 0,71	20,91 ± 4,45	28,97 ± 2,26	28,87 ± 6,67	
δ	1,20 ± 0,03	2,73 ± 0,33	2,36 ± 0,32	2,38 ± 0,62	
ρ	0,20 ± 0,04	6,18 ± 8,01	2,54 ± 1,72	4,64 ± 9,97	

Tabela5.29 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da distribuição Burr1. *Intervalos de confiança obtidos pelo método de "Jackknife", em virtude de não se terem obtido estimativas de MV.

Consideremos agora a distribuição Burr1. Para a Idade5 não se obtiveram estimativas de MV para os parâmetros dessa distribuição. A partir das estimativas resultantes do método dos momentos $\hat{\lambda} = 36{,}24$, $\hat{\delta} = 4{,}77$ e $\hat{\rho} = 0{,}66$, obtiveram-se estimativas "Jackknife" $\hat{\lambda}_{Jk} = 33{,}88$, $\hat{\delta}_{Jk} = 3{,}80$ e $\hat{\rho}_{Jk} = 0{,}69$. Os erros padrão aproximados são, respectivamente, $SE(\hat{\lambda}_{Jk}) = 2{,}44$, $SE(\hat{\delta}_{Jk}) = 0{,}24$ e $SE(\hat{\rho}_{Jk}) = 1{,}09$. Uma vez que $t_{0{,}05[176]} = 2{,}0017$, os intervalos de confiança aproximados são $\lambda : [28{,}99 ; 38{,}78]$, $\delta : [3{,}32 ; 4{,}29]$ e $\rho : [-1{,}49 ; 2{,}87]$. Repare-se que o extremo inferior do intervalo de confiança para ρ é negativo, indicando que com estas estimativas o ajustamento desta distribuição ($\rho > 0$) à Idade5 não fará muito sentido.

Para as Idades 2 e 4 os intervalos de confiança têm amplitudes consideráveis.

Burr2	Idade1*	Idade2	Idade3	Idade4	Idade5
λ		21,38 ± 1,94	27,00 ± 2,54	27,83 ± 0,39	30,79 ± 0,77
δ		0,6560 ± 0,5791	1,0048 ± 0,5619	0,0614 ± 0,1392	0,4783 ± 0,0075
κ		0,23 ± 0,30	0,27 ± 0,46	0,15 ± 0,13	1,48 ± 0,83
ρ		0,2738 ± 0,2197	0,4481 ± 0,2450	0,0299 ± 0,0688	0,0999 ± 0,0003

Tabela5.30 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da distribuição Burr2. *Não foi possível obter estimativas para a Idade1 por qualquer um dos métodos de estimação utilizados.

Os intervalos de confiança para as estimativas dos parâmetros da distribuição Burr2 parecem razoáveis.

De um modo geral são as estimativas obtidas para os modelos lognormal, misturas e Burr2 que apresentam as menores amplitudes dos intervalos de confiança. A distribuição gama, que apesar de ser uma das distribuições que melhor se ajusta, na generalidade, aos dados, é a que apresenta maiores amplitudes para os intervalos de confiança (provavelmente devido às aproximações feitas para tornar possível a aplicação do algoritmo EMMC à distribuição gama).

Na página seguinte apresentam-se as tabelas com as estimativas e limites de confiança aproximados a 95% para os parâmetros das misturas em localização de distribuições normais e lognormais.

mistura normais	Idade1 2 comp. mistura	Idade1 3 comp. mistura	Idade2 2 comp. mistura	Idade3 2 comp. mistura	Idade4 2 comp. mistura	Idade5 2 comp. mistura	Idade5 3 comp. mistura
φ_1	0,429 ± 0,084	0,215 ± 0,038	0,837 ± 0,120	0,913 ± 0,125	0,963 ± 0,078	0,312 ± 0,140	0,548 ± 0,096
φ_2	0,571 ± 0,084	0,469 ± 0,037	0,163 ± 0,119	0,087 ± 0,125	0,037 ± 0,078	0,688 ± 0,140	0,222 ± 0,085
φ_3		0,316 ± 0,039					0,230 ± 0,088
λ_1	10,433 ± 0,706	14,477 ± 0,574	26,262 ± 0,608	31,486 ± 0,661	33,294 ± 0,702	42,301 ± 1,282	32,559 ± 0,296
λ_2	19,228 ± 0,586	19,968 ± 0,202	32,047 ± 1,434	37,815 ± 3,155	41,622 ± 5,931	33,482 ± 0,973	43,631 ± 0,581
λ_3		9,404 ± 0,319					37,841 ± 0,633
δ	2,490 ± 0,353	1,957 ± 0,160	2,802 ± 0,437	2,863 ± 0,448	2,758 ± 0,651	2,376 ± 0,588	1,479 ± 0,231

Tabela5.31 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da mistura em localização de distribuições normais.

mistura lognormais	Idade1 2 comp. mistura	Idade1 3 comp. mistura	Idade2 2 comp. mistura	Idade3 2 comp. mistura	Idade4 2 comp. mistura	Idade5 2 comp. mistura	Idade5 3 comp. Mistura
φ_1	0,659 ± 0,083	0,557 ± 0,036	0,653 ± 0,164	0,888 ± 0,520	0,965 ± 0,164	0,351 ± 0,145	0,235 ± 0,090
φ_2	0,341 ± 0,083	0,194 ± 0,022	0,347 ± 0,174	0,112 ± 0,520	0,035 ± 0,164	0,649 ± 0,145	0,536 ± 0,090
φ_3		0,249 ± 0,029					0,229 ± 0,087
λ_1	2,902 ± 0,046	2,956 ± 0,011	3,236 ± 0,032	3,446 ± 0,058	3,503 ± 0,029	3,731 ± 0,033	3,628 ± 0,019
λ_2	2,240 ± 0,052	2,094 ± 0,026	3,406 ± 0,043	3,584 ± 0,211	3,710 ± 0,365		3,480 ± 0,008
λ_3		2,506 ± 0,020				3,501 ± 0,026	3,772 ± 0,020
δ	0,185 ± 0,029	0,126 ± 0,007	0,098 ± 0,017	0,095 ± 0,024	0,084 ± 0,024	0,063 ± 0,016	0,041 ± 0,006

Tabela5.32 – Estimativas e limites de confiança aproximados a 95% para os parâmetros da mistura em localização de distribuições lognormais.

5.5 Regressão

Cada idade representa uma população estatística diferente. Até aqui o estudo realizado refere-se ao ajustamento de distribuições teóricas a cada uma idades. Tenta-se se seguida o ajustamento de um modelo à totalidade dos dados (à pescaria) visando obter alguma ideia sobre o crescimento desta espécie. Assim comparam-se o ajustamento de uma recta (a recta dos mínimos quadrados) e uma curva dada pela equação de Von Bertalanffy (Beverton e Hold, 1957) , aos dados referentes a esta pescaria.

A relação entre idade e comprimento pode ser adequadamente descrita pela curva de crescimento de Von Bertalanffy, dada por

$$l_t = L_\infty\left(1-e^{-Kt)}\right)$$

onde l_t é o comprimento no tempo t, L_∞ é o comprimento máximo teórico da espécie e K é a taxa instantânea de crescimento. Fazendo um ajustamento desta equação aos dados (pelo algoritmo de Levenberg-Marquardt, disponível no SPSS), obteve-se $L_\infty = 38{,}7428$, $K = 0{,}5728$. Os limites dos intervalos de confiança assintóticos aproximados a 95%, são $38{,}7428 \pm 1{,}0729$ para L_∞, e $0{,}5728 \pm 0{,}0357$ para K.

A tabela Anova é a seguinte:

F.V.	SQ	gl	MQ
Entre	37981,66	4	9495,42
regressão	37279,76	2	18639,88
Desvio	701,90	2	350,95
Dentro	13195,20	864	15,27
Total	51176,86	868	

Tabela5.33 – Análise de variância para o ajustamento da curva de Von Bertalanffy.

Da tabela, resulta $r^2 = \dfrac{SQ_{regressão}}{SQ_{total}} \approx 0{,}73$ (coeficiente de determinação), ou seja a variação no crescimento é explicada pelo modelo (pela idade) em cerca de 73%.

A forma mais simples de relação estocástica entre duas variáveis T e Y chama-se modelo de regressão linear simples. Este toma a forma

$$Y_i = \alpha + \beta T_i + \varepsilon_i .$$

Mas a especificação plena de um modelo de regressão linear simples compõe-se, não só de equação de regressão acima apresentada, como também de determinadas premissas acerca do termo erro (ε_i) e da forma como os valores de T são determinados.

Essas premissas, que se consideram aplicadas a todas as observações, são:

(1) Os erros ou desvios seguem uma distribuição normal.

(2) A média dos erros é nula: $E[\varepsilon_i] = 0, \forall i$.

(3) Há homogeneidade de variâncias: $E[\varepsilon_i^2] = \sigma^2, \forall i$.

(4) Os erros são independentes uns dos outros: $E[\varepsilon_i \varepsilon_j] = 0, \forall i \neq j$.

(5) T não é uma v.a., toma valores fixos e conhecidos em amostras repetidas, de forma tal que, para cada dimensão da amostra, o valor de $\frac{1}{n}\sum_{i=1}^{n}(T_i - \overline{T})^2 \neq 0$ e finito.

Assim tem-se

$$E[Y_i] = \alpha + \beta T_i \text{ e } \mathrm{Var}[Y_i] = E[\varepsilon_i]^2 = \sigma^2 ,$$

donde $Y_i \sim N(\alpha + \beta T_i, \sigma^2)$.

Fazendo o teste de homocedasticidade de variâncias de Bartlett (apêndice A4) obtêm-se $X^2\text{corrigido} = 52{,}6233 > \chi^2_{0,05[4]} = 9{,}4877$, rejeitando-se para um nível de sigificância de 5% a hipótese de homocedasticidade das variâncias. Apesar disso e de a distribuição normal não se ajustar à idade 1, vamos proceder à análise de variância associada ao modelo considerado.

A recta de regressão é $\hat{Y} = 5{,}3442T + 14{,}2020$ e os extremos dos intervalos de confiança assintóticos aproximados a 95% são $5{,}3442 \pm 0{,}2877$ e $14{,}2020 \pm 0{,}7755$, respectivamente.

A tabela Anova completada com a regressão é:

F.V.	SQ	gl	MQ	F_s
Entre	37981,6610	4	9495,4153	621,7441***
Regressão	30941,0340	1	30941,0340	13,1839*
Desvio	7040,6270	3	2346,8757	153,6695***
Dentro	13195,2020	864	15,2722	
Total	51176,8630	868		

Tabela5.34 – Análise de variância completada com a regressão linear simples.

Note-se que $F_{0,01}[4;864]=3,3409$, $F_{0,05}[1;3]=10,1280$ e $F_{0,01}[1;3]=34,1161$. Da tabela5.35 podem-se tirar as seguintes conclusões:

- Há diferenças significativas no comprimento entre as diferentes idades;
- Não se aceita a hipótese nula de que o modelo de regressão seja linear (os desvios da regressão são altamente significativos);
- Rejeita-se a hipótese nula de que o declive da recta de regressão seja nulo;
- $r^2 = 0,6046$.

Comparando graficamente os resultados alcançados analiticamente, vem

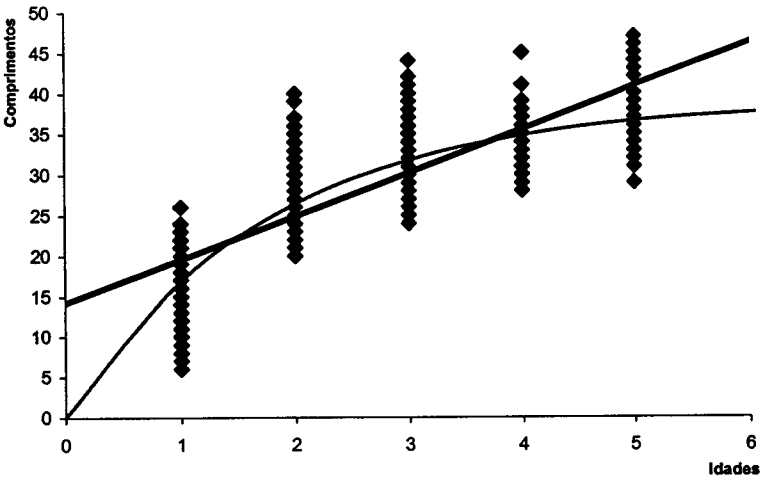

Figura5.41 - Ajustamento da recta de regressão simples e da curva de Von Bertalanffy (t=0) às diferentes amostras.

Tendo em conta a figura5.41 e os resultados obtidos para r^2 , conclui-se, como seria de esperar, que a curva de Von Bertalanffy se ajusta melhor aos dados do que a recta dos mínimos quadrados (apesar do ajustamento não ser o melhor). De facto, uma vez que a razão entre a entrada de energia absorvida pelo indivíduo (alimento) e o seu consumo diminui, o crescimento dos peixes vai abrandar ao longo do tempo (Beverton e Holt, 1957).

Em virtude da grande amplitude das amostras o ajustamento quer da recta quer da curva não é o melhor, facto este que pode ser reforçado pelos resultados obtidos para r^2. Este ajustamento pode ser melhorado ignorando-se as variações individuais e considerando apenas a média dos dados para cada idade (perdendo-se, assim, informação em relação às amostra). Neste caso a amplitude de cada amostra reduz-se a um ponto. Como se pode observar graficamente o ajustamento da curva de Von Bertalanffy é quase perfeito

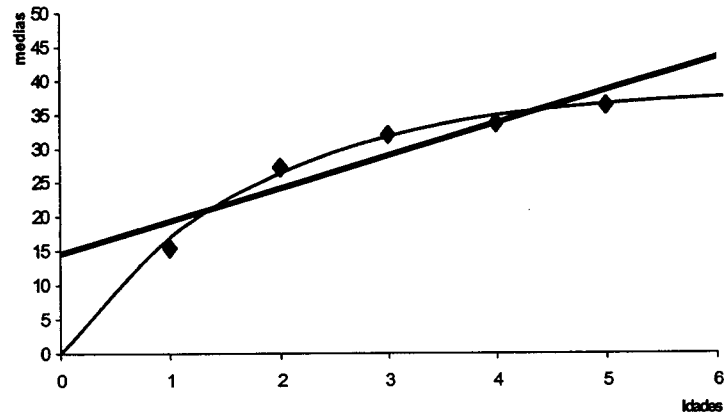

Figura 5.42 - Ajustamento da recta de regressão simples e da curva de Von Bertalanffy à média das amostras.

Para a equação de Von Bertalanffy, vem $L_\infty = 38{,}3801$, $K = 0{,}5701$ e $r^2 = 0{,}9859$. Repare-se que os valores obtido para os dois primeiros coeficientes são aproximadamente iguais nas duas situações.

Em relação à regressão linear simples obtém-se, $Y = 4{,}7967T + 14{,}5150$ e $r^2 = 0{,}8538$, pior resultado do que para a curva de Von Bertalanffy.

6. Conclusões gerais

Em biologia das pescas, assume-se, na generalidade dos casos, que a distribuição normal descreve adequadamente o comportamento dos dados referentes aos comprimento dos peixes com determinada idade. Na verdade, verifica-se que as amostras correspondentes a estes dados apresentam variações na assimetria. Em particular, neste estudo 83% (considerando as 4 pescarias) das amostras apresenta assimetria positiva (a mortalidade natural é superior em indivíduos mais pequenos e a predação não afecta sobretudo os indivíduos de maiores dimensões). Espera-se, portanto, que um modelo que permita variações na assimetria se ajuste melhor, a este conjunto de dados, do que a distribuição normal. O objectivo principal deste trabalho foi o estudo de modelos, biologicamente significativos, alternativos à distribuição normal. Foram consideradas, para além da distribuição normal, a distribuição lognormal (uma vez que o processo de crescimento dos peixes se tratar de um fenómeno de carácter multiplicativo a nível celular), a distribuição gama (pela sua flexibilidade e por ser distribuição assintótica de certos modelos de crescimento em ambientes aleatórios), as distribuições de Burr (Burr, 1954), que foram sugerida por investigadores da área de biologia e pescas (por nunca terem sido estudadas neste campo e por permitirem assimetria tanto negativa como positiva) e as misturas de distribuições normais e lognormais, (por poder haver mistura de subpopulações e por estarem bem descritas na literatura da especialidade, sendo estas utilizadas como modelos biológicos)

A estimação dos parâmetros das diferentes distribuições teóricas foi feita, utilizando, quando possível, o método da máxima verosimilhança (MV) e métodos iterativos para o cálculo das estimativas de MV (métodos de Newton, algoritmo EM e algoritmo EM com simulação). A aplicação do algoritmo EM tem a vantagem de se poderem considerar os efeitos do agrupamento dos dados e de simplificar o processo de obtenção das estimativas de MV para a mistura de distribuições. O método dos momentos é utilizado para se obterem valores iniciais para a aplicação dos algoritmo iterativos, e sempre que a utilização de métodos de MV seja impossível.

As amostras estudadas (organizadas numa chave de comprimento à idade) foram denominadas, por ordem de faixa etária, respectivamente Idade1, Idade2, Idade3, Idade4 e Idade5 e correspondem ao comprimento dos peixes para cada ano.

Em relação aos métodos utilizados podem-se tirar as seguintes conclusões:

- Apenas para as distribuições normal e lognormal foi possível obter expressões explícitas para os estimadores de MV;
- Foi possível a aplicação do método "scoring" de Fisher para todas as distribuições;
- Foi possível a aplicação do método de Newton-Rapshon, para a obtenção das estimativas de MV, para todas as distribuições com excepção das misturas;
- A aplicação do algoritmo EM (considerando os dados não agrupados) simplifica bastante o cálculo das estimativas de MV no caso dos modelos de misturas. Utilizando este algoritmo com Newton-Rapshon, isto é, aproximando $I(\theta \mid x)$ por $I_e(\theta \mid x)$, uma vez que existem várias soluções para a equação de log-verosimilhança, este método tem a vantagem de convergir apenas para o máximo global.
- O algoritmo EM para dados agrupados só é viável para as distribuições normal e lognormal, devido à dificuldade de obter expressões explícitas para as esperanças matemáticas no passo-E do algoritmo;
- Os algoritmos EM com simulação (EME e EMMC) funcionam para todos os modelos. Na maior parte das situações a convergência das iteradas $\theta^{(k)}$ para o seu regime estacionário foi relativamente rápida. Nestes caso, deve-se ter em consideração o aumento da variância devido à simulação.
- A escolha dos valores iniciais para a aplicação dos métodos iterativos de MV pode ser reduzida às estimativas obtidas, quando possível, pelo método do momento. Em particular, reduzem o período de "queima" dos algoritmos com simulação;
- Cada vez que os algoritmos EME e EMMC são activados obtêm-se sequências de estimativas dos parâmetros que diferem um pouco. Apesar deste facto os resultados obtidos permitem conclusões idênticas.
- Não foi possível obter expressões explícitas dos parâmetros da distribuição Burr2 pelo método dos momentos;
- Quando a assimetria das amostras é negativa não é possível obter estimativas de MV para a distribuição gama, apenas pelo método dos momentos. Deve-se ter atenção que podem obter-se, para esta distribuição, estimativas inadequadas do

parâmetro de localização (superiores ao menor comprimento da amostra), o que acontece em alguns casos pelo método dos momentos quando a assimetria da amostra é negativa;

- Para a Idade1 (assimetria negativa), por qualquer um dos métodos de estimação, não se obtiveram quaisquer estimativas para a distribuição Burr2;
- Para a Idade5, para a distribuição Burr1, só foi possível obter estimativas pelo método dos momentos.

Uma vez que foram utilizados vários métodos de estimação, interessa apresentar os resultados daquele (ou daqueles) que possa ser utilizado para a maioria das distribuições. Assim, compararam-se as estimativas de MV obtidas para a distribuição normal utilizando os diferentes métodos. Podem-se tirar as seguintes conclusões:

- As médias pouco se alteram;
- As variâncias diminuem um pouco com a utilização do algoritmo EM, pois este tem em conta o facto de os dados estarem agrupados;
- A log-verosimilhança e o *AIC* são praticamente iguais utilizando qualquer um dos métodos mencionados.
- Os algoritmos estocásticos dão resultados muito próximos do algoritmo EM propriamente dito, o que dá uma certa tranquilidade em situações em que não é possível aplicar o algoritmo EM.
- Os erros padrão para os parâmetros aumentam pouco com a aplicação do algoritmo EM comparativamente à aplicação do método de MV. Contudo, nos algoritmos estocásticos, a variância dos parâmetros é inflacionada uma vez que inclui a variância adicional devida à simulação. Esta última é menor para o algoritmo EMMC. A variância adicional pode ser controlada usando um número M suficientemente grande de valores simulados.

Nestes termos, apresentaram-se os resultados obtidos considerando as estimativas dos parâmetros obtidas pelo algoritmo EMMC ($M = 30$) com Newton-Rapshon ou seja aproximando $I(\boldsymbol{\theta} \mid \boldsymbol{x})$ por $I_e(\boldsymbol{\theta} \mid \boldsymbol{x})$ ou por $I_{e,g}(\boldsymbol{\theta} \mid \boldsymbol{x})$ caso os dados sejam considerados agrupados ou não (obtendo-se automaticamente a variância das estimativas devida ao modelo). Exceptuando para as misturas, considera-se o caso de dados agrupados.

A selecção dos modelos foi baseada em testes de qualidade de ajuste e nos valores obtidos para a função de verosimilhança e para *AIC*. Deve-se ter em conta que, o aumento do número de parâmetros tende a aumentar o valor da função de log-verosimilhança e do valor do *AIC*. Nalguns casos obtiveram-se valores *AIC* muito próximos, o que pode levantar dúvidas em relação à escolha da distribuição mais "adequada" para cada amostra. Foram também utilizados testes de qualidade de ajuste; o teste de χ^2 e o teste de Kolmogorov-Smirnov (K-S), estes dois testes estão, geralmente, em consonância em termos de aceitar ou rejeitar H_0, diferindo, apenas, nalguns casos quanto à escolha da distribuição mais "adequada". Convém salientar os seguintes resultados:

- Idade1: Os testes evidenciam o melhor ajustamento para as misturas em localização de 3 distribuições normais e de 3 distribuições lognormais.
- Idade2: A distribuição gama e a mistura em localização de 2 distribuições lognormais são as que melhores resultados dão quanto ao ajustamento. Contudo, para a mistura de 2 lognormais verifica-se que os valores do parâmetro de localização das duas componentes da mistura estão muito próximos um do outro, o que aponta para o pouco interesse da consideração de um modelo de mistura de lognormais.
- Idade3: Enquanto o teste χ^2 privilegia o ajustamento da distribuição gama a esta amostra, o teste K-S privilegia o ajustamento à distribuição Burr2. Contudo, os valores de $D_{obs.}$ para estas duas distribuições estão muito próximos. O melhor valor da função de log-verosimilhança foi obtido para a distribuição Burr2, enquanto o melhor valor do *AIC* foi para a distribuição lognormal (relembre-se que o *AIC* aumenta com o número de parâmetros e que a distribuição Burr2 tem mais dois parâmetros do que a distribuição lognormal).
- Idade4: Apesar, dos melhores resultados para a função de log-verosimilhança e para o *AIC* se terem obtido para a distribuição Burr2, o teste de χ^2 privilegia o ajustamento da distribuição lognormal a esta amostra, enquanto o teste de K-S privilegia o ajustamento da distribuição gama..
- Idade5: O teste K-S (único utilizado por razões referidas) evidencia o ajustamento de uma mistura em localização de distribuições lognormais ou de distribuições normais. Contudo, verifica-se que os valores dos parâmetros de localização das

componentes da mistura de lognormais estão próximos, o que aponta para o pouco interesse da consideração de um modelo de mistura de lognormais. A distribuições que mais se aproximam destas, em termos de qualidade de ajuste, são a Burr2 e a gama cujos de $D_{obs.}$ são muito próximos.

Tendo presente que a diferença que existe, em considerar a distribuição normal, em vez de outra qualquer que melhor se ajuste aos dados, pode permitir más conclusões em termos de inferência, pode-se concluir que as distribuições apresentadas ajustam-se melhor às amostras, referentes aos dados de comprimentos à idade de peixes para as 4 pescarias estudadas neste trabalho, do que a distribuição normal.

Quando a assimetria das amostra é negativa (que no nosso caso foi de 17% do total das amostras) não faz muito sentido ajustar uma distribuição gama (por esta apresentar sempre assimetria positiva). Exceptuando para Idade1, para todas as outras amostras, tanto as misturas (aqui apresentadas) como as distribuições de Burr dão resultados satisfatórios em relação ao ajustamento. Para a Idade1 (provavelmente por não seguir o padrão das demais amostras nas mesmas condições, por razões já mencionadas) são as misturas que melhor se ajustam.

Em relação às amostras com assimetria positiva, apesar de todas as distribuições se ajustarem melhor do que a normal, pode-se evidenciar as distribuições gama e Burr2 (note-se que a distribuição Burr2 tem mais um parâmetro do que a distribuição gama).

As distribuições de Burr, apesar de não darem os melhores resultados em termos do ajustamento, ajustam-se razoavelmente às amostras quer estas tenham assimetria negativa ou positiva. Deve-se ter em conta que, enquanto para a distribuição Burr1 é sempre possível obter estimativas dos seus parâmetros pelos métodos aqui apresentados, para a distribuição Burr2 isso nem sempre é possível (lembre-se que não foram encontrados expressões explícitas para os estimadores através do método dos momentos).

Todas as outras distribuições têm a vantagem de ser relativamente "fácil" a obtenção de estimativas para os parâmetros (por um qualquer método aqui apresentado). Em particular, para as misturas de distribuições normais e lognormais, utilizando as estimativas saídas do método dos momentos, a aplicação do algoritmo EM facilita bastante o cálculo das estimativas de MV.

Depois de se ter estudado o ajustamento das diferentes distribuições para cada amostra, tentou-se a análise de regressão à totalidade das amostras. Foram comparadas a recta de regressão simples e a curva Von Bertalanffy, como era de esperar, uma vez que o crescimento dos peixes abranda ao longo do tempo, observou-se que a curva se ajusta melhor aos dados do que a recta.

Pensa-se que o objectivo principal proposto para este estudo foi alcançado, isto é, conseguiu-se, de entre um conjunto de distribuições biologicamente significativas para os dados referentes ao comprimento desta espécie (a pescada), distribuições alternativas à distribuição normal (habitualmente utilizada) com melhores ajustamentos. Privilegiou-se o método da MV, mas de forma a ter em conta os efeitos de agrupamento dos dados (algoritmo EM) e utilizaram-se versões estocásticas do algoritmo que permitem obter os estimadores com razoável rapidez e facilidade e que dão resultados semelhantes às versões determinísticas, conforme se constatou pelo estudo comparativo feito usando como referência a distribuição normal. Desenvolveu-se, assim, uma metodologia que permite estudar o ajustamento de distribuições, adequadas para um determinado conjunto de dados, como base para a inferência (bastante mais desenvolvida na literatura da especialidade).

Referências bibliográficas

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle".Pages 267-281 in B.N. Petra, and F. Csáki, editors. International symposium on information theory, second edition. Akadémiai iadi, Budapest, Hungary.

Akaike, H. (1985). "Prediction and entropy". Pages 1-24 in A.C. Atkinson, and S.E. Fienberg, editors. A celebration of statistics: the ISI Centenary Volume. Springer-Verlag, New York, New York, USA.

Behboodian, J. (1975). "Structural properties and statistics of finite mixtures." G. P. Patil el al. (eds.), Statistical Distributions in Scientific Work, **1**, 103-12.

Beverton, R.J.H. and Holt, S.J. (1957). "On the dynamics of exploited fish populations". Ministry of Agriculture Fisheries and Food, United Kingdom.

Boes, D.J. (1966). "On the estimation of mixing distributions". Ann. Of Math. Stats., **37**, 177-188.

Böhning, D., and Lindsay, B. (1988). "Monotonicity of quadratic approximation algorithms". Annals of the Institute of Statistical Mathematics., **40**, 641-663.

Broniatowski, M., Celeux, G., and Diebolt, J. (1983). "Reconaissance de densités par um algorithme d'apprentissage probabiliste". In Data Analysis and Informatics Vol. 3. Amsterdam: North-Holland, pp. 359-374.

Burnham, K.P. and Anderson, D.R.. "Data-base selection of an appropriate biological model: the key to modern data analysis". U.S. Fish & Wildlife Service, Colorado Cooperative Fish & Wildlife Research Unit, Colorado State University, Fort Collins, Colorado USA 80523.

Burr, I.W. (1954). "Cumulative frequency functions". Annals of Mathematical Statistics, XIII (2): 215-232.

Cabral, M.S. (1987a). "Inferência estatística em misturas de populações". Tese de doutoramento, não publicada, Universidade de Lisboa.

Cabral, M.S. (1987b). "Teste de existência de misturas em escala ou em localização". Actas das XII Jornadas Luso-Espanholas de Matemática. Vol.III, 71-72.

Cabral, M.S. (1993). "Identificação de um Modelo de Misturas Finitas: Aplicação a Dados Hidrológicos". I Congresso Anual da Sociedade Portuguesa de Estatística. A Estatística e o futuro e o Futuro e a Estatística, 413-424

Celeux, G., and Diebolt, J. (1985). "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem". Computational Statistics Quarterly, 2, 73-82.

Celeux, G., and Diebolt, J. (1986a). "The SEM and EM algorithms for mixtures: numerical and statistical aspects". Proceedings of the 7th Franco-Belgium Meeting of statistics. Bruxelles: Publication des Facultés Universitaries St. Louis.
Celeux, G., and Diebolt, J. (1986b). "L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités". Revue de Statistique Appliquée, 34, 35-52.

Celeux, G., and Diebolt, J. (1992). "A stochastic approximation type EM algorithm for the mixture problem". Stochastics and Stochastic Reports, **41**, 119-134.

Chan, K.S., and Ledolter, J. (1995). "Monte Carlo estimation for time series models involving counts". Journal of American Statistical Association, **90**, 242-252.

Crámer, H. (1946). "Mathematical Methods os Statistics". Princeton, New Jersey; Princeton University Press.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B, **39**, 1-38.

Diebolt, J., e Ip, E.H.S. (1996). "Stochastic EM: method and application". In Markov Chain Monte Carlo in practice, W.R. Gilks, S. Richardson, e D.J. Spiegelhalter (Eds.). London: Chapman & Hall, pp.259-273.

Dijkstra, T.K., editor. (1988). "On model uncertainty and its statistical implications". Lecture notes in economics and mathematical statistics. Springer-Verlag, New York, New York, USA.

Erzini, K. (1990). "Variability in length - at - age in marine fishes". Tese de doutoramento, não publicada. University of Rhode Island.

Fischer, R.A. (1922). "On the mathematical foundations of theoretical statistics". Philosophical Transactions of the Royal Society of London, Series A, **222**, 309-368.

Fischer, R.A. (1925). "Theory of statistical estimation". Proceedings of The Cambridge Philosophical Society, **22**, Pt.5, 700-725.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., (Eds.). (1996). "Markov Chain Monte Carlo in Practice". London: Chpman and Hall.

Gnanadesikan, R., editor. (1983). "Statistical data analysis." Proceedings of the symposia in applied mathematics, volume 28. American Mathematics Society, Providence, Rhode Island, USA.

Goodman, L.A. (1984). "The analysis of cross-classified data having ordered categories". Harvard University Press, Cambridge, Massachusetts, USA.

Gumbel, E.J. (1962). "Produits et quotients de deux plus grandes valeurs indépendantes". Comptes Rendus de l'Académie des Sciences, Paris, **254**, 2132-2134.

Hampel, F.R. (1974). "The influence curve and its role in robust estimation". Journal of the American Statistical Association, **69**, 383-393.

Huber, P.J. (1972). "Robust statistics: a review", Annals of Mathematical Statistics, **43**, 1041-1067.

Johnson, N.I. and Kotz, S. (1969). "Continuous univariate distributions - 1". Houghton Mifflin company, Boston, USA.

Johnson, N.I. and S. Kotz (1969). "Continuous univariate distributions - 2". Houghton Mifflin company, Boston, USA.

Lebreton, J.D., Brunham, K.P., Clobert, J. and Anderson, D.R. (1991). "Modelling survival and testing biological hypotheses using marked animals": case studies and recent advances. Ecological Monographs. *In press*.

Lehmann, E.L. (1983). "Theory of Point Estimation". New York: Wiley.

Linhart, H. and Zucchini, W. (1986). "Model selection". John Wiley and Sons, New York, New York, USA.

Louis, T.A. (1982). "Finding the observed information matrox when using the EM algorithm". Journal of the Royal Statistics Society B, **44**, 226-223.

McCullagh, P. e Nelder, J.A.. (1989). "Generalized linear models". Chapman e Hall, New York, New York, USA.

McLachlan, G.J., and Basford, K.E. (1988). "Mixture Models: Inference and applications to Clustering. New York: Marcel Dekker.

McLachlan, G.J. (1997). "Recent Advances in Finite Mixture Models". New York. Wiley and sons, to appear.

McLachlan, G.J., Krishnan, T. (1997). "The EM algorithm and extensions". John Wiley and Sons, New York, New York, USA.

Meilijson, I. (1989). "A fast improvement to the EM algorithm on its own terms". Journal of the Royal Statistical Society B, **51**, 127-138.

Mosteller, F. e Tukey, J. W. (1977). "Data Analysis and Regression". Readin, M.A: Addison-Wesley.

Nielsen, S.F. (2000). "The stochastic EM algorithm: Estimation and asymptotic results". Bernoulli, **6**(3), 457-489.

Orchard, T., and Woodbury, M.A. (1972). "A missing information principle: theory and applications". In Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability Vol.1. Berkeley, California: University of California Press, pp. 697-715.

Rider, P.R. (1961). "The method of moments applied to a mixture of two exponencial distributions". Annals of Mathematical Statistics , **32**, 143-47.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). "Akaike information criterion statistics" KTK Scientific Publishers, Tokyo, Japan.

Shibata, R. (1989). "Statistical aspects of model selection". Pages 214-240 in J.C. Williams, editor. From data to model. Springer-Verlag, New York, New York, USA.

Sokal, R.R. and Rohlf, F.J. (1995). "Biometry". Third edtion. W. H. Freeman and company, NewYork, USA.

Stuart, A. and Ord, J.K. (1987). "Kendall's Advanced Theory of Statistics: Distribution theory". Fifth edition. Charles Griffin & Company Limited, London.

Stuart, A. and Ord, J.K. (1987). "Kendall's Advanced Theory of Statistics: Inference and Relationship". Third edition. Charles Griffin & Company Limited, London.

Tanner, M. A. (1996). "Tools for statistical inference". Third edition. Springer-Verlag, New York, USA.

Teicher, H. (1963). "Identifiability of finite mixtures." Annals of Mathematical Statistics, **32**, 244-48.

Yakowitz, S. J. and Spragins, J.D. (1968). "On the Identifiability of finite mixtures." Annals of Mathematical Statistics, **39**, 209-14.

Yakowitz, S. J. (1969). "A consistent estimation of the identification of finite mixtures". Annals of Mathematical Statistics, **40**, 1728-1735.

Wei, G.C.G., and Tanner, M.A. (1990). "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms". Journal of the American Statistics Association, 85, 699-704.

Wu, C.F.J. (1983). "On the convergence properties of the EM algorithm". Annals of Statistics, **11**, 95-103.

Apêndice A1. Testes de qualidade de ajuste

Os testes de qualidade de ajuste são bem conhecidos e podem ser encontrados na literatura da especialidade. Faz-se aqui uma breve referencia ao teste de qui-quadrado (χ^2) e ao teste de Kolmogorov-Smirnov (K-S).

Depois de considerada a hipótese nula:

H_0 : A população segue uma determinada distribuição teórica

Os referidos testes são formulados da maneira que se segue:

A1.1 Teste de qui-quadrado

O teste de χ^2 para a avaliação da qualidade de ajuste baseia-se na comparação da distribuição dos dados amostrais com a distribuição teórica à qual se supõe pertencer a amostra. A metodologia que se adopta no teste inclui os passos que se descrevem de seguida.

(1) Calculam-se as frequências absolutas das observações amostrais nas diferentes classes X_j $(j=1,\ldots,v)$ mutuamente exclusivas. Tais frequências, denotam-se por n_j e satisfazem a condição, $\sum_{j=1}^{v} n_j = n$.

(2) Determinam-se as frequências esperadas (n_{ej}) para cada classe X_j $(j=1,\ldots,v)$ supondo H_0 verdadeira. Com $n_{ej} = np_j$, onde p_j representa a probabilidade da v.a. X tomar valores pertencentes à classe X_j. Note-se que $\sum_{j=1}^{v} n_{ej} = n$.

(3) A estatística de teste é construída com base numa medida dada por $\chi^2_{obs.} = \sum_{j=1}^{v} \frac{(n_j - n_{ej})^2}{n_{ej}}$.

(4) Uma vez fixado o nível de significância α, a rejeição ou não rejeição de H_0 será feita com base na comparação entre o valor que a estatística de teste toma e $\chi^2_{\alpha[v-p-1]}$ onde v representa o número de classes e p o número de parâmetros da distribuição populacional estimados a partir da amostra.

Sendo H_0 verdadeira, a estatística $\chi^2_{obs.}$ terá uma distribuição tanto mais próxima da distribuição $\chi^2_{\alpha[\nu-p-1]}$ quanto maior for a dimensão da amostra e maiores forem os números de observações esperadas nas diferentes classes (n_{ei}). A título indicativo, apresenta-se em seguida uma regra prática que permite utilizar este teste com confiança:

- dimensão da amostra não inferior a 30 $(n \geq 30)$;
- frequência esperada em cada classe não inferior a 5 $(n_{ei} \geq 5)$.

Se esta última condição não prevalecer, o teste pode ainda ser utilizado, embora com moderada confiança, se não mais de 20% dos valores de n_{ei} forem inferiores a 5 e nenhum for inferior a 1. Quando tal não se verificar, procuram-se agregar classes adjacentes, por forma a obter novas classes que satisfaçam esta condição.

A1.2 Teste de Kolmogorov-Smirnov

Podem ser apontadas duas vantagens do teste de K-S em relação ao teste de χ^2. Em primeiro lugar, quando a distribuição populacional é contínua e se conhecem a forma e os parâmetros da sua f.d.p., a distribuição da estatística do teste é definida rigorosamente (ao contrário do que sucede com a estatística $\chi^2_{obs.}$, cuja distribuição é aproximada). Esta vantagem é tanto mais nítida quanto menor for a dimensão da amostra. Em segundo lugar, o teste K-S é, na maioria das situações, mais potente do que o teste de χ^2. Em contrapartida, o teste K-S exige distribuições populacionais contínuas e completamente especificadas (o que não sucede com o teste de χ^2).

Para uma v.a. X, o teste K-S tem por base a análise do ajuste entre a f.d. populacional (teórica), $F_0(x)$, que é admitida em H_0 e a f.d. empírica ou da amostra, $S(x)$ (para qualquer valor particular x da variável X, esta função expressa a soma das frequências relativas dos dados com valores menores ou iguais a x).

No teste K-S de qualidade de ajuste adopta-se o procedimento que se descreve em seguida:

(1) Uma vez determinada a função de distribuição empírica, $S(x)$, calcula-se a estatística de teste $D_{obs} = \sup_x |S(x) - F_0(x)|$.

(2) Uma vez especificado o nível de significância do teste, o valor D_{obs} observado é comparado com o respectivo valor crítico D_α, rejeitando-se H_0 se $D_{obs} > D_\alpha$.

O supremo de $|S(x) - F_0(x)|$ não é necessariamente o maior valor que $|S(x) - F_0(x)|$ toma quando se consideram apenas os valores *observados* de *X*. De facto, dado que a função $F_0(x)$ é contínua e $S(x)$ é uma função em escada, o valor máximo daquela diferença absoluta deve ser procurado na vizinhança de cada valor observado de *X*.

É possível demonstrar que, se a amostra é aleatória e provém de uma distribuição contínua conhecida, a estatística D_{obs} só depende da dimensão da amostra (n), sendo irrelevante a forma da f.d. $F_0(x)$.

Para amostras de grandes dimensões $D_\alpha \approx \sqrt{\frac{-\ln(1/2\alpha)}{2n}}$.

Tal como se afirmou atrás, o teste K-S é exacto (ou seja, o risco α está definido rigorosamente) quando a função $F_0(x)$ se encontra perfeitamente especificada e, em particular, se conhecem os seus parâmetros. O teste pode, no entanto, ser utilizado quando os parâmetros de $F_0(x)$ são estimados a partir da amostra. Porém, nestas circunstâncias, deverá ter-se em conta que o nível de significância com que se realiza o teste é menor do que aquele que é especificado e que a potência do teste também diminui de uma quantidade não conhecida.

Apêndice A2. Critério de informação de Akaike

Fundamentalmente, a selecção de modelos não é um problema de testes de hipóteses; é um problema de optimização de um qualquer critério segundo um conjunto de modelos a serem considerados. No contexto da estimação de MV, dado um conjunto de modelos a justar o Critério de Informação de Akaike (*AIC*) é dado pela expressão:

$$AIC = -2V(\hat{\theta} \mid x) + 2p,$$

onde $V(\hat{\theta} \mid x)$ é o valor da função de log-verosimilhança para a estimativa dos parâmetros, e p é o número parâmetros a estimar no modelo. Calcula-se o *AIC* para todos os modelos de interesse sendo o modelo mais "adequado" aquele que tiver menor *AIC*. Este critério envolve duas componentes, sendo $-2V(\hat{\theta} \mid x)$ uma medida da discrepância do ajustamento entre os dados e o modelo; quantos mais parâmetros tiver o modelo melhor será o ajustamento e $-2V(\hat{\theta} \mid x)$ decrescerá monoticamente. Assim, se a selecção for baseada apenas no melhor ajustamento, pode-se acabar por seleccionar o modelo com o maior número de parâmetros; o que normalmente sobreajusta os dados. O segundo termo do *AIC* é uma penalização que cresce proporcionalmente com o aumento no número de parâmetros. Quanto maior for esta para o número de parâmetros, mais ênfase efectivamente se dá à "parcimónia". Assim, há uma tendência para se aceitar algum enviesamento visando aumentar a precisão dos estimadores. O não enviesamento é uma propriedade desejável. Infelizmente, na prática, o custo (nos termos de aumentar a variância amostral) é muitas vezes grande se se procurar um total não enviesamento.

Apêndice A3. O método de "Jackknife"

O método "jackknife" é um método não paramétrico destinado a estimar o enviesamento e a variância dos estimadores em condições teoricamente complexas ou em que não se tem confiança no modelo especificado. É um método de re-amostragem pois baseia-se na construção de subamostras da amostra inicial. Pode ser resumido nos seguintes passos:

1. Calcular a estimativa $\hat{\theta}$ do parâmetro θ baseada na amostra (de tamanho n);
2. Calcular estimativas $\hat{\theta}_i$ $(i = 1,\ldots,n)$ baseadas nas n possíveis amostras de dimensão $n-1$ obtidas através da eliminação consecutiva observação i $(i = 1,\ldots,n)$;
3. Calcular os pseudovalores, $\varphi_i = n\hat{\theta} - (n-1)\hat{\theta}_i$;
4. A estimativa "Jackknife" do parâmetro será $\hat{\theta}_{Jk} = \frac{1}{n}\sum \varphi_i = \overline{\varphi}$;

5. O erro padrão aproximado de $\hat{\theta}_{Jk}$ será dado por $SE\left(\hat{\theta}_{Jk}\right)=\sqrt{\frac{\sum_{i=1}^{n}(\varphi_i-\overline{\varphi})^2}{n(n-1)}}$.

Os limites de confiança são, então, dados por $\hat{\theta}_{Jk} \pm t_{\alpha[n-1]} SE\left(\hat{\theta}_{Jk}\right)$, onde α é o nível de significância do teste.

Apêndice A4. Teste de homogeneidade de variâncias de Bartlett

A aplicação deste teste exige a existência para cada amostra da variável independente de mais do que uma observação para variável dependente.

Considere-se a amostras de observações, cada um deles correspondendo a valores idênticos para as variáveis independentes, e o número de observações para cada amostra a n_i.

O teste é então feito como se segue:

1. Para cada amostra, calcula-se a variância dos valores de Y, s_i^2;
2. Calculam-se as quantidades

$$s^2=\frac{\sum_{i=1}^{a}(n_i-1)s_i^2}{\sum_{i=1}^{a}(n_i-1)} \text{ e } X^2=\sum_{i=1}^{a}(n_i-1)\ln s^2-\sum_{i=1}^{a}(n_i-1)\ln s_i^2 .$$

Na hipótese de haver homocedasticidade, X^2 terá uma distribuição aproximadamente $\chi^2_{[a-1]}$. Esta aproximação será melhorada dividindo X^2 pelo seguinte factor de correcção:

$$C=1+\frac{1}{3(a-1)}\left[\sum_{i=1}^{a}\frac{1}{n_i-1}-\frac{1}{\sum_{i=1}^{a}n_i-1}\right].$$

Aceita-se a hipótese de homocedasticidade das variâncias, para um nível de significância α, se $X^2 < \chi^2_{\alpha[a-1]}$.