XXIII Congresso da
**Sociedade Portuguesa de Estatística**

# ATAS DO XXIII CONGRESSO

## Da Sociedade Portuguesa de Estatística

**Editores:**
Maria de Fátima Salgueiro
Paula Vicente
Teresa Calapez
Catarina Marques
Maria Eduarda Silva

iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

SPE Sociedade Portuguesa de Estatística

# ATAS DO XXIII CONGRESSO DA SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Lisboa, 18 a 21 de outubro de 2017

**Editores**
*Maria de Fátima Salgueiro*
*Paula Vicente*
*Teresa Calapez*
*Catarina Marques*
*Maria Eduarda Silva*

Janeiro, 2020
**Edições SPE**

**Editores:** Maria de Fátima Salgueiro, Paula Vicente, Teresa Calapez, Catarina Marques e Maria Eduarda Silva

**Título:** Atas do XXIII Congresso da Sociedade Portuguesa de Estatística

**Editora:** Sociedade Portuguesa de Estatística

**Conceção Gráfica da Capa:** Andreia Garcia (Iscte - Instituto Universitário de Lisboa)

**ISBN:** 978-972-8890-46-9

## Prefácio

Este é o Livro de Atas do XXIII Congresso da Sociedade Portuguesa de Estatística (SPE), que se realizou em Lisboa entre 18 e 21 de Outubro de 2017, nas instalações do ISCTE-Instituto Universitário de Lisboa.
Lisboa foi desta feita escolhida pela Sociedade Portuguesa de Estatística (SPE) para acolher o seu Congresso de 2017.

Lisboa, Janeiro de 2020
*Os Editores*

## Agradecimentos

Aos seguintes colegas, pelo generoso trabalho de revisão dos artigos submetidos a este Livro de Atas, que em muito valorizou o conteúdo desta publicação:

- **Ana Paula Amorim**, Universidade do Minho
- **Ana Sousa Ferreira**, Universidade de Lisboa
- **Antónia Turkman**, Universidade de Lisboa
- **Carlos Tenreiro**, Universidade de Coimbra
- **Cláudia Silvestre**, Instituto Politécnico de Lisboa
- **Conceição Amado**, IST, Universidade de Lisboa
- **Cristina Miranda**, Universidade de Aveiro
- **Esmeralda Gonçalves**, Universidade de Coimbra
- **Graça Trindade**, Iscte - Instituto Universitário de Lisboa
- **Helena Ferreira**, Universidade da Beira Interior
- **Helena Mouriño**, Universidade de Lisboa
- **Isabel Alves Rodrigues**, IST, Universidade de Lisboa
- **Isabel Barão**, Universidade de Lisboa
- **Isabel Pereira**, Universidade de Aveiro
- **Isabel Silva Magalhães**, Universidade do Porto
- **Joana Leite**, Instituto Politécnico de Coimbra
- **José Dias Curto**, Iscte - Instituto Universitário de Lisboa
- **José Manuel G.Dias**, Iscte - Instituto Universitário de Lisboa

- **Lisete Sousa**, Universidade de Lisboa
- **Luís Antunes**, Universidade do Porto
- **Luís Machado**, Universidade do Minho
- **Manuel Scotto**, IST, Universidade de Lisboa
- **Manuela Neves**, ISA, Universidade de Lisboa
- **Margarida Cardoso**, Iscte - Instituto Universitário de Lisboa
- **Maria Almeida Silva**, Universidade de Lisboa
- **Maria da Graça Temido**, Universidade de Coimbra
- **Maria do Carmo Botelho**, Iscte - Instituto Universitário de Lisboa
- **Helena Carvalho**, Iscte - Instituto Universitário de Lisboa
- **Marília Antunes**, Universidade de Lisboa
- **Miguel Pereira**, Imperial College, London
- **Nazaré Mendes-Lopes**, Universidade de Coimbra
- **Paula Milheiro-Oliveira**, Universidade do Porto
- **Rui Menezes**, Iscte - Instituto Universitário de Lisboa
- **Sandra Dias**, Universidade de Trás-os-Montes e Alto Douro
- **Sebestyan Szabolcs**, Iscte - Instituto Universitário de Lisboa
- **Sofia Azevedo**, Faculdade de Ciências, Universidade de Lisboa

Um **agradecimento especial** é também devido aos colegas da **Direção da Sociedade Portuguesa de Estatística** que colaboraram diretamente na realização deste congresso e aos colegas das Comissões Científica e Organizadora do XXIII Congresso da Sociedade Portuguesa de Estatística.

## Comissão Científica

- **Maria Eduarda Silva**, *Presidente da Sociedade Portuguesa de Estatística*, Faculdade de Economia, Universidade do Porto
- **Maria de Fátima Salgueiro**, Iscte - Instituto Universitário de Lisboa
- **Nazaré Mendes–Lopes**, Universidade de Coimbra
- **Conceição Amado**, Instituto Superior Técnico
- **Paulo M.M. Rodrigues**, Nova School of Business and Economics
- **José Manuel G. Dias**, Iscte - Instituto Universitário de Lisboa

## Comissão Organizadora

- **Maria de Fátima Salgueiro**
- **Paula Vicente**
- **Teresa Calapez**
- **Catarina Marques**
- **Elizabeth Reis**

*Iscte - Instituto Universitário de Lisboa*
e *Business Research Unit (BRU - Iscte)*

## Agradecimentos

Agradecemos às seguintes entidades o valioso apoio concedido para a realização do XXIII Congresso da SPE

- **Banco de Portugal**
- **Edições Sílabo**
- **EPAL - Grupo Águas de Portugal**
- **Escolar Editora**
- **Fundação para a Ciência e a Tecnologia**
- **Instituto Nacional de Estatística**
- **Iscte - Executive Education**
- **Iscte - Instituto Universitário de Lisboa**
- **Produtos e Serviços de Estatística, PSE**
- **Sociedade Portuguesa de Estatística**
- **Turismo de Lisboa**

# O critério *Minimum Message Lenght* na estimação de modelos de mistura sobre dados mistos

Cláudia Silvestre
Escola Superior de Comunicação Social-Instituto Politécnico de Lisboa, *csilvestre@escs.ipl.pt*

Margarida G. M. S. Cardoso
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, *margarida.cardoso@iscte.pt*

Mário A. T. Figueiredo
Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal, *mario.figueiredo@lx.it.pt*

**Palavras–chave**: Classificação não supervisionada; Análise de Agrupamento; Modelos de Mistura Finita; Dados Mistos; Minimum Message Lenght.

**Resumo**: Neste trabalho propomos uma nova variante do algoritmo *Expectation-Maximization* para agrupar dados mistos que simultaneamente estima o número de grupos. Recorremos aos modelos de mistura finita, pressupondo que os dados categoriais são modelados por distribuições multinomiais e os métricos por distribuições gaussianas. Para estimar o número de componentes de mistura baseamo-nos no critério *Minimum Message Lenght*. O desempenho do algoritmo proposto, designado por EM-MML-mix, é comparado com o de outros critérios usados frequentemente para a seleção de modelos de mistura. Desta análise comparativa, realizada sobre dados simulados e sobre um conjunto de dados reais provenientes do *European Social Survey*, salienta-se o reduzido tempo de computação para a obtenção da solução mediante a metodologia proposta.

# 1 Introdução

O agrupamento sobre dados mistos é um problema prático comum, nomeadamente no âmbito das ciências sociais. Este pode referir-se, por exemplo, à constituição de segmentos homogéneos de indivíduos, considerando as suas características métricas ou qualitativas. As abordagens metodológicas a este problema têm sido diversas. Por exemplo, Chiu et al. [8] propõem um algoritmo incremental e Ahmad e Dey [1] propõem um novo algoritmo K-Médias, ambos capazes de lidar com dados métricos e categoriais.

No âmbito do agrupamento com modelos de mistura finita, uma primeira proposta considerando dados mistos deve-se a Everitt [10]. A vantagem desta abordagem para segmentação, reside na sua capacidade de analisar diversos tipos de variáveis, de modelar relações entre elas, de integrar diversos critérios de seleção dos modelos e ainda de selecionar o número de segmentos (componentes da mistura).

Um modelo de mistura finita considera uma distribuição conjunta para as variáveis base de segmentação como uma soma ponderada de distribuições intra-segmentos, atendendo à natureza diversa dos atributos. A sua estimação viabiliza a construção de uma estrutura probabilística de segmentos e, em simultâneo, a obtenção de estimativas dos parâmetros distribucionais intra-segmentos. Neste âmbito, Hunt e Jorgensen [13] modelam a distribuição conjunta de uma variável categorial e de multinormais, permitindo, nestas últimas, que as médias dependam das categorias da variável qualitativa (sendo as covariâncias comuns). Outros trabalhos integram, nos modelos de mistura finita, a modelação conjunta de variáveis mistas considerando diversas distribuições, admitindo correlações intra-grupos de variáveis métricas ou mesmo de variáveis métricas contínuas (por exemplo, [20] e [15]). O critério que habitualmente orienta a estimação destes modelos é o da máxima verosimilhança. No entanto, incorporando informação *a priori*, podem também adotar-se métodos bayesianos.

Neste trabalho, consideramos o agrupamento de dados mistos, usando um modelo de mistura e propondo o uso do critério *Minimum Mes-*

*sage Lenght* (MML) [21] para a sua estimação. Este critério advém da teoria da informação, considerando como modelo mais adequado aquele que permite uma descrição mais sucinta das observações. Figueiredo e Jain [11] foram pioneiros na utilização deste critério para estimação de misturas de gaussianas e uma primeira proposta para a utilização do MML em misturas de multinomiais foi proposta por Silvestre et al. [19]. Este critério também foi usado em agrupamento de dados *fuzzy* em [16], onde os autores consideraram misturas de gaussianas e usaram o MML para estimar as variáveis relevantes e identificar o número de componentes de mistura.
A presente análise integra dados mistos considerando uma mistura de gaussianas e multinomiais, bem como um algoritmo que é uma variante do conhecido *Expectation-Maximization* (EM). A metodologia é testada comparativamente com critérios comuns para a seleção de modelos de mistura, nomeadamente o Integrated Completed Lilelihood, o qual é particularmente adequado neste contexto [12]. A análise é efetuada sobre dados sintéticos e um conjunto de dados reais (provenientes do *European Social Survey*). São feitas análises comparativas quanto ao tempo de computação, à qualidade do agrupamento obtido e à robustez, relativamente a diferentes processos de inicialização.

## 2 Metodologia

Em muitos dos trabalhos propostos a escolha do número de grupos é feita *a posteriori*. Por exemplo, nos métodos hierárquicos, a escolha do número de grupos é feita após o agrupamento, recorrendo aos correspondentes dendrogramas. Os critérios baseados na verosimilhança, habitualmente combinados com a estimação de modelos de mistura finita, também necessitam que o agrupamento seja feito previamente. Entre estes critérios, são comuns os seguintes: *Bayesian Information Criterion* (BIC) [17], *Akaike Information Criterion* (AIC) [2] e suas variantes [5, 6] e *Integrated Complete Likelihood* (ICL)[4]. No uso destes critérios, o agrupamento é feito para dife-

rentes números de grupos e escolhe-se a solução que corresponde ao melhor valor do critério usado. A metodologia que se propõe incorpora a determinação do número de grupos na estimação do modelo de mistura.

## 2.1 Modelos de mistura finita

Os modelos de mistura finita têm uma longa tradição em agrupamento; e.g., Wedel e Kamakura [22] referem o seu uso no âmbito de aplicações em marketing. A sua natureza probabilística/estatística tem várias vantagens importantes. Nomeadamente, a possibilidade de se modelar dados de diferentes naturezas e de se abordar formalmente a estimação do número de grupos.
Seja $\mathbf{Y} = \{\underline{y}_i,\ i = 1, \ldots, n\}$ uma amostra aleatória de $n$ observações independentes de $\underline{Y} = [Y_1, \ldots, Y_D]'$. A ideia base dos modelos de mistura finita é considerar a distribuição conjunta para as variáveis base de segmentação $\underline{Y}$ como sendo uma soma ponderada de distribuições intra-segmentos,

$$f(\underline{y}|\Theta) = \sum_{k=1}^{K} \alpha_k f(\underline{y}|\underline{\theta}_k),$$

onde $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_K, \alpha_1, \ldots, \alpha_K\}$ é o conjunto de todos os parâmetros do modelo, $K$ o número total de segmentos e $\underline{\theta}_k$ representa o conjunto dos parâmetros distribucionais do $k$-ésimo segmento (componente de mistura). Os pesos $\alpha_1$,...,$\alpha_K$ são as probabilidades de cada segmento, pelo que $\alpha_k \geq 0$, para $k = 1, \ldots, K$ e $\sum_{k=1}^{K} \alpha_k = 1$. Em agrupamento, a componente de mistura de onde provém cada uma das observações é desconhecida, por isso, pode dizer-se que os dados observados, $\mathbf{Y}$, são dados incompletos. Essa informação em falta é usualmente designada por $\mathbf{Z}$: $\mathbf{Z} = \{\underline{z}_1, ..., \underline{z}_n\}$ onde $\underline{z}_i = [z_{i1}, ..., z_{iK}]'$ e $z_{ik}$ é um indicador binário que toma o valor 1 se a observação $\underline{y}_i$ foi gerada pela $k$-ésima componente e 0 caso contrário. É habitual assumir-se que $\{\underline{z}_i,\ i = 1, \ldots, n\}$ são i.i.d. e

que seguem uma distribuição multinomial com $K$ categorias e probabilidades $\{\alpha_1, \ldots, \alpha_K\}$. Assim, o logaritmo da verosimilhança dos dados completos,($\mathbf{Y}$,$\mathbf{Z}$), é dado por

$$\log f(\mathbf{Y},\mathbf{Z}|\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \left[\alpha_k f(\underline{y}_i|\underline{\theta}_k)\right].$$

Neste trabalho pretendemos agrupar/segmentar dados mistos, ou seja, de natureza categorial e métrica. Consideremos que $\underline{Y}$ tem M variáveis categoriais que serão modeladas por distribuições multinomiais e G variáveis métricas que serão modeladas por distribuições gaussianas, tal que $M+G = D$. Assumindo que as variáveis são condicionalmente independentes, o logaritmo da verosimilhança para os dados completos é dado por:

$$\log f(\mathbf{Y},\mathbf{Z}|\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \left[\alpha_k \prod_{m=1}^{M} f(\underline{y}_{im}|\underline{\theta}_{km}) \prod_{g=1}^{G} f(\underline{y}_{ig}|\underline{\theta}_{kg})\right].$$

Para se obter as estimativas de máxima verosimilhança é habitual recorrer-se ao algoritmo *Expectation Maximization* (EM) [9].

## 2.2 O algoritmo EM

O algoritmo EM é um algoritmo iterativo que é frequentemente usado quando se pretende obter as estimativas de máxima verosimilhança (ML) ou o máximo *a posteriori* (MAP) na presença de dados incompletos. Um problema bem conhecido deste algoritmo é conduzir a um máximo local (não necessariamente ao global) da função de verosimilhança. Uma forma habitual de ultrapassar este problema consiste em calcular várias estimativas obtidas com condições iniciais diferentes, escolhendo-se para solução final aquela que apresentar valor mais elevado da função de verosimilhança.
O algoritmo EM alterna entre dois passos:

**Passo E:** Calcula o valor esperado do logaritmo da verosimilhança completa, condicional aos dados observados

$$E\left[\log f(\mathbf{Y},\mathbf{Z}|\Theta)\middle|\mathbf{Y},\widehat{\Theta}^{(t)}\right] \equiv \log f(\mathbf{Y},\bar{\mathbf{Z}}^{(t)}|\Theta),$$

onde a igualdade é justificada pelo facto de $\log p(\mathbf{Y},\mathbf{Z}|\Theta)$ ser uma função linear de $\mathbf{Z}$ e onde cada elemento $\bar{z}_{ik}^{(t)}$ de $\bar{\mathbf{Z}}^{(t)}$ é dado por

$$\bar{z}_{ik}^{(t)} = E\left[Z_{ik}\middle|\mathbf{Y},\widehat{\Theta}^{(t)}\right] = P\left[Z_{ik}=1\middle|\underline{y}_i,\widehat{\Theta}^{(t)}\right] = \frac{\alpha_k\ f(\underline{y}_i|\underline{\theta}_k^{(t)})}{\sum_{k=1}^{K}\alpha_k\ f(\underline{y}_i|\underline{\theta}_k^{(t)})},$$

e $t$ indica a iteração que está a ser executada.

**Passo M:** Calcula as estimativas dos parâmetros mediante a maximização do valor esperado do logaritmo da verosimilhança completa obtida no passo **E**

$$\widehat{\Theta}^{(t+1)} = \arg\max_{\Theta}\ \log p(\mathbf{Y},\bar{\mathbf{Z}}^{(t)}\big|\Theta) + \log p(\Theta), \qquad (1)$$

onde $p(\Theta)$ é a probabilidade *a priori* considerada quando se pretende obter estimativas MAP; quando se pretende encontrar os estimadores de ML, a parcela $\log p(\Theta)$ é omitida.

# 3 O algoritmo proposto: EM-MML-mix

Para estimar os parâmetros da mistura de multinomiais e gaussianas e simultaneamente o número de componentes, propomos uma variante do algoritmo EM, designado EM-MML-mix. Tomámos por base dois trabalhos [11, 18] que usam um critério MML e desenvolveram uma variante do algoritmo EM para estimação de mistura de gaussianas e multinomiais, respectivamente.
O critério MML privilegia um modelo estatístico que descreva os dados de forma sucinta, no sentido da teoria da informação. Assim,

para uma v.a. $Y$ com f.(d.)p. $p(y|\Theta)$, o comprimento de codificação óptimo (em bits) de uma observação $y$ que é dado por $l(y,\theta) = -\log_2 p(y|\Theta)$ (eventualmente adicionada de $\log_2 p(\Theta)$ quando os parâmetros $\Theta$ são desconhecidos) deverá ser o menor possível. Para misturas, esta função (cujo desenvolvimento encontra-se em [3]) é

$$l(y,\Theta) = -\log p(\Theta) - \log p(y|\Theta) + \frac{1}{2}\log|I(\Theta)| + \frac{(K-1)KN}{2}\left(1-\log(12)\right),$$

onde $|I(\Theta)|$ é o determinante do valor esperado da matriz de Fisher, $I(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\Theta^2}\log p(Y|\Theta)\right]$, e $N$ é o número de parâmetros a ser estimado em cada componente de mistura. No contexto de agrupamento com modelos de mistura, para ultapassar alguns problemas de cálculo, em [11] o valor esperado da matriz de Fisher foi calculado considerando os dados completos, ou seja, $I_c(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\theta^2}\log p(Y,Z|\Theta)\right]$. Os autores também usaram *priors* independentes de Jeffreys para os parâmetros de mistura, chegando assim à seguinte função *message length*

$$l(y,\Theta) = \frac{N}{2}\sum_{k:\,\alpha_k>0}\log\left(\frac{n\,\alpha_k}{12}\right) + \frac{k_{nz}}{2}\log\frac{n}{12} + \frac{k_{nz}(N+1)}{2} - \log p(y,\Theta)$$

onde $k_{nz}$ o número de componentes com probabilidade diferente de zero.

Para estimar os parâmetros do modelo de mistura de multinomiais e gaussianas, propomos uma variante do algoritmo EM. Este algoritmo, EM-MML-mix permite estimar, simultaneamente, o número de segmentos e os parâmetros distribucionais associados às variáveis base de segmentação.

**Algoritmo 3.1**

**Passo E***: O passo E é igual ao do algoritmo EM*

$$\bar{z}_{ik}^{(t)} = \frac{\alpha_k\; f(\underline{y}_i|\underline{\theta}_k^{(t)})}{\sum_{j=1}^{K}\alpha_j\; f(\underline{y}_i|\underline{\theta}_j^{(t)})},$$

*para* $i = 1, \ldots, n$ *e* $k = 1, \ldots, K$*;*
*onde* $f(\underline{y}_i|\underline{\theta}_k^{(t)}) = \prod_{m=1}^{M} f(\underline{y}_{im}|\underline{\theta}_{km}^{(t)}) \prod_{g=1}^{G} f(\underline{y}_{ig}|\underline{\theta}_{kg}^{(t)})$

**Passo M***: Atualiza as estimativas dos parâmetros do modelo de mistura:*

- *as probabilidades de mistura*

$$\widehat{\alpha}_k^{(t+1)} = \frac{\max\left\{0, \sum_{i=1}^{n} \bar{z}_{ik}^{(t)} - \frac{N}{2}\right\}}{\sum_{j=1}^{K} \max\left\{0, \sum_{i=1}^{n} \bar{z}_{ij}^{(t)} - \frac{N}{2}\right\}},$$

*para* $k = 1, \ldots, K$ *e onde* $N$ *é o número de parâmetros a estimar em cada componente de mistura.*

*Repare-se que no caso de algum valor* $\widehat{\alpha}_k^{(t+1)}$ *ser zero a k-ésima componente da mistura é eliminada. Os parâmetros das componentes da mistura com* $\widehat{\alpha}_k^{(t+1)} = 0$ *não precisam ser calculados uma vez que não contribuem para a verosimilhança, pelo que depois de calculados os valores de* $\widehat{\alpha}_k^{(t+1)}$ *só se calculam os parâmetros das componentes cuja probabilidade de mistura é diferente de zero,* $\widehat{\alpha}_k^{(t+1)} > 0$*.*

- *os parâmetros da multinomial*

$$\widehat{\theta}_{kmc}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)} y_{imc}}{n_m \sum_{i=1}^{n} \bar{z}_{ik}^{(t)}},$$

*para* $k = 1, \ldots, K, m = 1, \ldots, M$ *e* $c = 1, \ldots, C_m$*.*

- *os parâmetros da gaussiana*

$$\widehat{\mu}_{kg}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)} y_{ig}}{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)}}$$

$$\widehat{\sigma}_{kg}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)} (y_{ig} - \widehat{\mu}_{kg}^{(t+1)})^2}{\sum_{i=1}^{n} \bar{z}_{ik}^{(t)}}$$

*para* $k = 1, \ldots, K$ *e* $g = 1, \ldots, G$.

# 4 Agrupamento usando o EM-MML-mix

## 4.1 Agrupamento de dados sintéticos

Para avaliar o desempenho do algoritmo. começámos por aplicá-lo a dados sintéticos. Considerámos 2 conjuntos de dados gerados a partir de duas componentes de mistura. Num dos conjuntos considerámos apenas uma variável métrica e uma variável categorial e no outro considerámos mais uma variável categorial. Para estes 2 conjuntos, geraram-se dados de dimensões diferentes (250 e 1000) e em ambos os casos consideraram-se componentes de mistura equilibradas (120 vs 130 e 450 vs 550) e componentes de mistura não equilibradas (50 vs 200 e 800 vs 200), perfazendo um total de 8 amostras de dados sintéticos. Em cada um dos casos geraram-se 10 réplicas, tendo-se corrido o EM-MML-mix e escolhido a solução que apresentava o menor message length.
O algoritmo EM-MML-mix recupera os dois segmentos. No caso da ou das variáveis categoriais, as probabilidades associadas a cada uma das categorias são exatamente iguais, se arredondadas a uma

casa decimal. Quanto à variável métrica, os resultados não são tão bons, uma vez que se obtêm estimativas próximas para as médias, sendo as dos desvios padrão superiores aos valores originais.

## 4.2 Agrupamento das regiões do European Social Survey

### Dados do ESS

Os dados reais analisados são provenientes do European Social Survey (ESS). Este é um inquérito transnacional dirigido aos cidadãos europeus que se realiza de dois em dois anos, desde 2001. O objetivo da análise é agrupar as 250 regiões (de 21 países) do European Social Survey (round 7, 2014), atendendo a indicadores relacionados com o trabalho, nomeadamente uma variável binária $Y_1-$ *É responsável por supervisionar outros no trabalho?* e uma métrica $Y_2-$ *Número de horas contratadas por semana no trabalho.* Os dados referidos às regiões são obtidos mediante soma ponderada de respostas "sim" e "não" à questão $Y_1$ (codificadas com 1 e 0 respetivamente) e mediante média ponderada referida ao número de horas contratadas. A ponderação atende à dimensão da população e ao peso pós-estratificação. Sendo assim trabalha-se com as variáveis *Número de supervisores* e *Número médio de horas de trabalho*, por região.

### Resultados Comparativos

Para uma análise comparativa dos resultados da variante EM-MML-mix é usada, como alternativa, a tradicional metodologia de estimação baseada no critério da máxima verosimilhança (ML) seguida de critérios de teoria da informação para a determinação do número de segmentos. São usados os critérios BIC, ICL, AIC e variantes. Estes critérios adicionam à função de verosimilhança (que se pretende maximizar) uma penalização da complexidade do modelo que dependerá da dimensão da amostra e/ou do número de parâmetros a estimar (favorecendo um modelo mais parcimonioso). Em resul-

tado da análise são obtidos dois segmentos, independentemente do critério adotado. Há, no entanto, diferenças entre os dois agrupamentos e constata-se que os indicadores de coesão-separação utilizados – índice Silhueta [14] e Calinski e Harabasz [7] – apontam para uma melhor solução produzida pela estimação de modelo de mistura usando o EM tradicional (Tabela 1). O tempo de computação favorece claramente a metodologia que se propõe.

Tabela 1: Qualidade dos agrupamentos obtidos e tempo de computação associado

| Critério | BIC, AIC, CAIC, AIC3, ICL | EM-MML |
|---|---|---|
| Número de grupos | 2 | 2 |
| Índice Silhueta | 0.541 | 0.53 |
| Calinski e Harabasz | 339.613 | 323.493 |
| Tempo de computação | 26,521 s | 6,278 s |

**Os segmentos obtidos**

Em resultado da análise efetuada e atendendo aos indicadores de qualidade de agrupamento, as regiões agrupam-se nos dois segmentos propostos pela aplicação da metodologia de estimação por ML seguida de critérios habituais da teoria de informação para a determinação do número de grupos. No primeiro grupo encontram-se 62 regiões e no segundo 188. Na Figura 1 representa-se a localização geográfica das regiões dos dois segmentos.

De acordo com o seu perfil, o segmento 1 caracteriza-se por ter, em média, mais trabalhadores em funções de supervisão (37%) e por uma média de horas de trabalho que ronda as 32h semanais. No segmento 2 a média de horas de trabalho sobe para 39h e apenas 26% são responsáveis por supervisionar outros no trabalho.

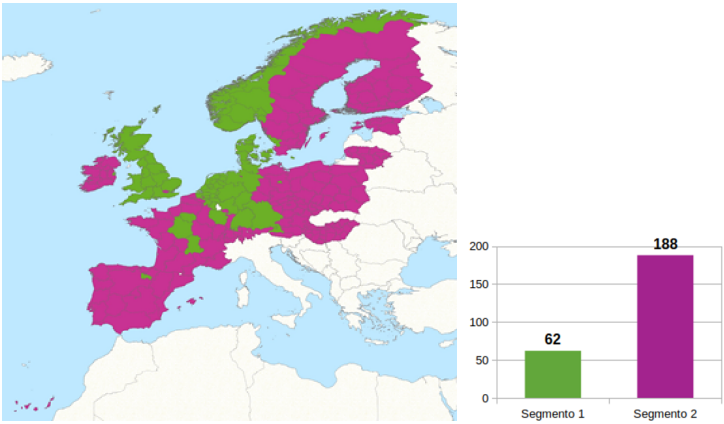


Figura 1: Distribuição geográfica dos segmentos constituídos

# 5 Conclusões

Neste trabalho, propusemos uma variante do algoritmo EM, o EM-MML-mix, para agrupar dados agregados mistos. O algoritmo proposto permite estimar simultaneamente os parâmetros de uma mistura finita de multinomiais e gaussianas, assim como o número de componentes da mistura (número de segmentos), com base no critério *Minimum Message Length*. Quando lidamos apenas com dados categoriais o EM-MML apresenta resultados melhores que o ICL e semelhantes aos obtidos usando BIC, AIC, CAIC e AIC3 [18]. Na presença de dados mistos e no conjunto limitado de testes por agora efetuados sobre dados gerados e um conjunto de dados reais, a metodologia proposta destaca-se somente pelo seu reduzido tempo de computação. Esta será uma vantagem relevante ao trabalhar com um grande volume de dados. Sobre os dados sintéticos, observa-se, contudo, alguma imprecisão na recuperação da estrutura de dados gerados. Por isso, em trabalhos futuros, o EM-MML-mix deverá ser

melhorado e testado em conjuntos de dados com mais variáveis e de maior dimensão, de forma a reavaliar esta vantagem e a obter uma melhor compreensão do seu desempenho.

# Referências

[1] Ahmad, A., Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2), 503–527.

[2] Akaike, H.(1973). Maximum Likelihood Identification of Gaussian Autorregressive Moving Average Models. *Biometrika*, 60, 255–265.

[3] Baxter, R. A. e Olivier, J. J. (2000). Finding overlapping components with MML. *Statistics and Computing*,10(1), 5–16.

[4] Biernacki, C.,Celeux, G., Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 22, 719–25.

[5] Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psycometrika*, 52, 345–370.

[6] Bozdogan, H. (1994). Mixture-Model Cluster Analysis using Model Selection criteria and a new Informational Measure of Complexity. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Approach*, 69–113.

[7] Calinski, R. B., Harabasz, J. (1974) A dendrit method for cluster analysis. *Communications in Statistics*, 3(1), 1–27.

[8] Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In Provost, R., Srikant, R., (eds.): *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 263–268.

[9] Dempster, A., Laird, N., Rubin, D. (1997). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society*, 39, 1–38, Series B.

[10] Everitt, B. S. (1988).A finite mixture model for the clustering of mixed-mode data. *Statistics and probability letters*, 6(5), 305–309.

[11] Figueiredo, M. A. T., Jain, A. K. (2002).Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 381-396.

[12] Fonseca, J. R., Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(2), 155–173.

[13] Hunt, L., Jorgensen, M. (1999). Theory and Methods: Mixture model clustering using the MULTIMIX program. *Australian and New Zealand Journal of Statistics*, 41(2), 154–171.

[14] Kaufman, L., Rousseeuw, P. J. (1990). *Finding groups in data: an Introduction to cluster analysis*. Wiley, NY.

[15] Marbac, M., Sedki, M. (2017). Model-based clustering of Gaussian copulas for mixed data. *Communications in Statistics - Theory and Methods*, 43(26), 11635–11656.

[16] Saha, A., Das, S. (2018). Clustering of fuzzy data and simultaneous feature selection: A model selection approach. *Fuzzy Sets and Systems*, 340, 1–37.

[17] Schwarz, G. (1978). Estimating the Dimension of a Model *The Annals of Statistics*,6, 461–464.

[18] Silvestre, C. (2015). *Clustering with Discrete Mixture Models - An integrated approach for model selection.* Tese de Doutoramento. ISCTE - IUL.

[19] Silvestre, C., Cardoso, M. G. M. S. and Figueiredo, M. (2015). Feature selection for clustering categorical data with an embedded modelling approach. *Expert Systems*, 32, 444–453.

[20] Vermunt, J., Magidson, J. (2002). *Applied latent class analysis*. JA Hagenaars and AL McCutcheon, Cambridge: Cambridge University Press.

[21] Wallace, C. S., Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, 11(2), 195–209.

[22] Wedel, M., Kamakura, W. (2002). *Market Segmentation- Conceptual and Methodological Foundations*. Vol. 8. Edições Springer Science & Business Media, New York.