An application of mixture distributions in modelization of length of hospital stay

N. Atienza[1, *, †], J. García-Heras[2], J. M. Muñoz-Pichardo[2] and R. Villa[3]

[1]*Departamento de Matemática Aplicada I, Universidad de Sevilla, Spain*
[2]*Departamento de Estadística e I.O, Universidad de Sevilla, Spain*
[3]*Departamento de Análisis Matemático, Universidad de Sevilla, Spain*

SUMMARY

Length of hospital stay (LOS) is an important indicator of the hospital activity and management of health care. The skewness exhibited by this variable poses problems in statistical modeling. The aim of this work is to model the variable LOS within diagnosis-related groups (DRG) through finite mixtures of distributions. A mixture of the union of Gamma, Weibull and Lognormal families is used in the model, instead of a mixture of a unique distribution family. Some theoretical questions regarding the model, such as the identifiability and study of asymptotic properties of ML estimators, are analyzed. The EM algorithm is proposed for performing these estimators. Finally, this new proposed model is illustrated by using data from different DRGs.

KEY WORDS: length of stay; diagnosis-related groups; finite mixture model; Gamma distribution; Weibull distribution; Lognormal distribution

1. INTRODUCTION

Length of stay (LOS) is an easily available indicator of hospital activity. It is used for various purposes, such as management of hospital care, quality control, appropriateness of hospital use and hospital planning [1, 2]. LOS is an indirect estimator of resources consumption and of the efficiency of one of the aspects of hospital patient care: bed management. Lee *et al.* [3] state that 'comprehensive and accurate information about inpatient LOS should be a high priority for

---

health planners and administrators in the strategic planning and deployment of financial, human and physical resources'.

Since the early 1980s, health-care systems in industrial countries have exhibited deep changes in order to reduce overspending on health care, particularly hospital expenditures. Among all the proposed modifications, the hospital stay classification given by Fetter *et al.* [4] received the widest approval for assessing hospital output. In this classification, the main diagnosis established at the end of the stay and the procedures performed during the hospitalization are used to retrospectively classify a patient into a diagnosis-related group (DRG) that determines the amount of payment allocated to the hospital. The U.S.A. Congress decided in October 1983 to implement a Medicare prospective system based on the classification of Fetter *et al.* [4]. Later on, many countries such as Australia or France introduced DRGs to reduce the health-care budget. The DRGs provide a classification system of episodes of hospitalization with clinically recognized definitions, where it is expected that patients in the same class consume similar quantities of resources, as a result of a process of similar hospital care. The mean of the LOS is used as an indicator of the consumption of resources because of its availability and good relation with the raised costs. Hence, we may say that DRGs have been partially created in order to get homogeneous groups with respect to the consumption of services and costs, closely related to the LOS. Moreover, the comparison of the expected LOS by DRG raises conclusions of a given aspect of the management for each specific kind of patients.

The expectation is only a localization measure of the variable LOS. Therefore, a more complete study of the distribution of this variable is convenient. To this aim, in the past years, many authors proposed different technics to analyze the variable LOS. However, the empirical distribution of LOS is well established to be positively skewed, plurimodal, to contain outliers and to significantly vary between DRGs [5, 6], etc. This heterogeneity of LOS poses a problem for statistical analysis, limiting the use of inference techniques based on normality assumptions.

Since a large number of DRGs must be analyzed routinely, automatic procedures are needed for conveniently treating skewness. Different transformations (e.g. the logarithmic one) of LOS have been attempted to attain normality, and subsequently to apply the corresponding tests [5, 7, 8], etc.). However, as Xiao *et al.* [9] show, these approaches rely on the unrealistic homogeneity assumption on the entire sample.

Marazzi *et al.* [6] assessed the adequacy of three conventional parametric models, Lognormal (long-tailed), Weibull and Gamma (short-tailed), for describing the LOS distribution. But, as Lee *et al.* [10] point out, none of them seemed to fit satisfactorily in a wide variety of samples. The main issue is that the assumption of heterogeneous sub-populations would be more appropriate than single DRG populations. Mixture distribution analysis can clarify whether or not a skewed distribution is composed of heterogeneous components [9, 11]. With this new focus, several works model the LOS variable through finite mixture distributions. For example, Quantin *et al.* [12] analyzed the cost and the hospital stay of DRG 589 and 590 (lymphoma and leukemia with or without complications) using mixtures of Weibull distributions. The same model was also applied by Quantin *et al.* [11] to analyze DRG 316 (renal failure). A Gamma mixture was proposed to analyze heterogeneity of maternity LOS in Lee *et al.* [10, 13]. The model was applied to different DRGs related to cesarean delivery (DRG 670, 671, 672 and 687) and to DRGs related to vaginal delivery (DRG 674, 675, 676 and 688). It is then clear that the proposed mixture model depends on which DRG or group of DRGs is considered.

The objective of this work is to analyze the variable LOS unifying the approach given by Marazzi and this new focus, tackling the problem of the hospital stay through a finite mixture of

distributions in the union of Gamma, Weibull and Lognormal families. This new posing can be used to model the variable LOS for most of the DRGs. The results obtained in this work prove the validity of this new approach.

The paper is organized as follows. Section 2 contains a theoretical exposition of the mixture distributions: the identifiability problem, the asymptotic properties of the maximum likelihood (ML) estimators and the adaptation of the EM algorithm, needed for the parametric estimation. In Section 3, a simulation study is presented to analyze the performance of our method. We apply this method in Section 4 to several DRGs with the aim of illustrating the goodness of fit of the proposed model. Finally, we have included a short discussion on the proposed methodology.

2. METHODS

A finite mixture distribution can be considered as a convex combination of distribution functions, which can be applied to model two different practice situations, according to Titterington *et al.* [14]. One in the case of a variable with several underlying categories or with different sources of origin (*direct application*), and the other, when these categories do not exist or are not physically interpretable. In this case, the finite mixture model is used as a statistical tool to provides more flexibility in the fittings and get better results (*indirect application).*

This work deals with the second case. We have obtain a sample of the LOS variables obtained from patients in a given DRG. Moreover, our distributions have specific parametric forms as follows.

Let $\mathscr{F}_L$, $\mathscr{F}_G$ and $\mathscr{F}_W$ be, respectively, the Lognormal, Gamma and Weibull distributions families:

$$\mathscr{F}_L = \left\{ F : F(x; \mu, \sigma) = \int_0^x \frac{1}{\sqrt{2\pi}\sigma u} \exp\left[-\frac{1}{2}\left(\frac{\log u - \mu}{\sigma}\right)^2 \right] du; \mu \in \mathbb{R}, \sigma > 0, x > 0 \right\}$$

$$\mathscr{F}_G = \left\{ F : F(x; a, b) = \int_0^x \frac{b^{-a}}{\Gamma(a)} u^{a-1} \exp\left(-\frac{u}{b}\right) du; a, b > 0, x > 0 \right\}$$

$$\mathscr{F}_W = \left\{ F : F(x; c, d) = \int_0^x \frac{c}{d^c} u^{c-1} \exp\left(-\frac{u^c}{d^c}\right) du; c, d > 0, x > 0 \right\}$$

We propose to model the LOS variable with a threefold mixture distribution:

$$Q(x; \Psi) = \pi_1 F_1(x; \mu, \sigma) + \pi_2 F_2(x; a, b) + \pi_3 F_3(x; c, d)$$

where $\pi_1, \pi_2, \pi_3 > 0$, $\pi_1 + \pi_2 + \pi_3 = 1$, $F_1 \in \mathscr{F}_L$, $F_2 \in \mathscr{F}_G$, $F_3 \in \mathscr{F}_W$ and $\Psi = (\pi_1, \pi_2, \mu, \sigma, a, b, c, d)$ is the parametric vector determinate by the given distributions. Let us denote this class of mixture distributions by $\mathscr{C}$. Obviously, the main problem is the estimation of the parameters. However, before solving this problem, it is necessary to tackle the problem of identifiability, the question on the unique representation of the class of models being considered.

Generally, the problem of estimation of parameters is undertaken through the ML method, because of its desirable statistical properties, such as efficiency, consistency and asymptotic normality under some uniform integrability assumptions on the mixture and its derivatives.

However, despite the good properties of the ML estimators, in our case it is not possible to obtain explicit solutions. Hence, numerical methods must be applied. The EM algorithm, proposed by Dempster *et al.* [15], is the more used method.

Next, we study these three aspects in the finite mixtures considered: identifiability, asymptotic properties and the EM algorithm.

2.1. Identifiability

Identifiability problems concerning finite and countable mixtures have been studied widely. Teicher [16] gave a sufficient condition for a finite mixture to be identifiable. This condition is based on the existence of a linear and one-to-one mapping defined on the distribution family and a total ordering on this family. As an application, he proved the identifiability of finite mixtures of normal distributions and of Gamma distributions by using the bilateral Laplace transform and the lexicographic order. This sufficient condition was subsequently modified by Chandra [17] using the moment-generating function of $\log X$. This result was used by Khalaf [18] to prove the identifiability of finite mixtures of Weibull, Lognormal, Chi-squared and Pareto distributions. Different modifications with application to specific families of distributions have been proposed by several authors (see Brandorff-Nielsen [19] or Henna [20]).

Atienza *et al.* [21] provide a sufficient condition for the identifiability of finite mixtures, which is applied to the class of all finite mixtures generated by the union of Lognormal, Gamma and Weibull distributions, where Teicher's and Henna's conditions are not applicable. Thus, if we denote by $\mathcal{U} = \mathcal{F}_L \cup \mathcal{F}_G \cup \mathcal{F}_W$ the family obtained by the union of these three families, then the class $\mathcal{H}_{\mathcal{U}}$ of all finite mixtures of distributions from $\mathcal{U}$ is identifiable. Obviously, since $\mathcal{C} \subset \mathcal{H}_{\mathcal{U}}$, the class $\mathcal{C}$ is identifiable.

2.2. Asymptotic properties of ML estimators

In finite mixture models, ML estimators have good properties (efficiency, consistency and asymptotic normality) under some uniform integrability assumptions on the mixture and its derivatives up to the third order.

The problem of consistency has been the center of interest for many authors in the study of solutions for both general and specific families. The works of Chanda [22] and Redner and Walker [23] should be cited because of their importance and influence in subsequent works. In particular, these works establish, for a sample of size n, under two conditions, that there exists in any sufficiently small neighborhood of the value of the parameter Ψ^* a unique strongly consistent solution Ψ^n of the likelihood equations and that this solution at least locally maximizes the log-likelihood function and is asymptotically normally distributed. The first condition establishes the uniform integrability for the partial derivatives up to the third order of the mixture. The second condition deals with the positive-definite character of the Fisher information matrix. More precisely, the Fisher information matrix $I(\Psi)$ given by

$$I(\Psi) = \int [\nabla_\Psi \log q(\mathbf{x}; \Psi)][\nabla_\Psi \log q(\mathbf{x}; \Psi)]^t q(\mathbf{x}; \Psi)\, d\mathbf{x}$$

has to be well defined and positive definite at Ψ^*, where ∇_Ψ denotes the gradient of the first partial derivatives with respect to the components of Ψ.

Atienza *et al.* [24] prove that under another two conditions related to the integrability of the components of the mixtures, for any compact subset Ω' of the parametric space Ω which contains

Ψ^*, with probability 1, Ψ^n is a ML estimator in Ω' for sufficiently large n. Furthermore, this ML estimator is unique in Ω', that is, there is no other ML estimator besides Ψ^n which leads to a limiting density different from $q(\mathbf{x}, \Psi^*)$.

The integrability conditions are proved in Atienza *et al.* [25] are to be verified for the union of the so-called $\mathscr{W}$-type families. Since the distribution families $\mathscr{F}_L$, $\mathscr{F}_G$ and $\mathscr{F}_W$ are $\mathscr{W}$-type families, the authors establish that the class $\mathscr{H}_{\mathscr{U}}$ of all finite mixtures of distributions from $\mathscr{U}$ verifies the integrability conditions. The inclusion $\mathscr{C} \subset \mathscr{H}_{\mathscr{U}}$ clearly implies that the class $\mathscr{C}$ also verifies these conditions.

It remains to prove that the second condition is also verified, that is, $I(\Psi)$ is well defined and positive definite at Ψ^*. Applying Proposition 1 (see Appendix), it is easily proved that the information matrix of any element from the class $\mathscr{C}$ is well defined. The definite positive character is obtained by application of Proposition 2. The density function of the mixture distribution can be expressed in the form

$$q(x; \Psi) = \pi_1 f_1(x; \Psi) + \pi_2 f_2(x; \Psi) + (1 - \pi_1 - \pi_2) f_3(x; \Psi)$$

where

$$f_1(x; \Psi) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{1}{2} \left(\frac{\log x - \mu}{\sigma} \right)^2 \right]$$

$$f_2(x; \Psi) = \frac{b^{-a}}{\Gamma(a)} x^{a-1} \exp\left(-\frac{x}{b} \right)$$

and

$$f_3(x; \Psi) = \frac{c}{d^c} x^{c-1} \exp\left(-\frac{x^c}{d^c} \right)$$

with $x > 0$, $\sigma, a, b, c, d > 0$ and $\mu \in \mathbb{R}$. The set of partial derivatives with respect to the parameter is provided in Table I. In order to prove its linear independence character, the following partial order is defined on that set:

$$g_1 \prec g_2 \iff \lim_{x \to \infty} \frac{g_2(x)}{g_1(x)} = 0 \quad \text{and} \quad g_1 \sim g_2 \iff \lim_{x \to \infty} \frac{g_2(x)}{g_1(x)} \in \mathbb{R} \setminus \{0\}$$

The arrangement of the set of partial derivatives $\{w_i(x; \Psi), i = 1, \ldots, 8\}$ is provided in Table II.

Let us consider any linear combination $\sum_{i=1}^{8} \gamma_i w_i(x; \Psi)$ of this set. The linear independence of the set is proved by showing that the equality $\sum_{i=1}^{8} \gamma_i w_i(x; \Psi) = 0$ holds provided that all coefficients γ_i's are zero. Dividing by $w_4(x; \Psi)$ and taking limit $x \to \infty$ give $\gamma_4 = 0$. We may continue in this fashion for cases (1), (2) and (4) in Table II to obtain that the rest of the coefficients are null.

In case (3), an analogous procedure yields $\gamma_4 = \gamma_3 = \gamma_1 = \gamma_7 = 0$. Dividing by $w_2(x; \Psi)$ and taking limit now give

$$\gamma_2 + \gamma_8 \lim_{x \to \infty} \frac{w_8(x; \Psi)}{w_2(x; \Psi)} = 0$$

Table I. Partial derivatives of $q(x; \Psi)$.

Parameter	Partial derivatives of $q(x; \Psi)$
π_1	$w_1(x; \Psi) = f_1(x; \Psi) - f_3(x; \Psi)$
π_2	$w_2(x; \Psi) = f_2(x; \Psi) - f_3(x; \Psi)$
μ	$w_3(x; \Psi) = \pi_1 \left(\frac{\log x - \mu}{\sigma^2} \right) f_1(x; \Psi)$
σ	$w_4(x; \Psi) = \pi_1 \left(\frac{(\log x - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right) f_1(x; \Psi)$
a	$w_5(x; \Psi) = \pi_2 [\log x - \log b - \Gamma'(a)/\Gamma(a)] f_2(x; \Psi)$
b	$w_6(x; \Psi) = \pi_2 \frac{x}{b^2} f_2(x; \Psi)$
c	$w_7(x; \Psi) = (1 - \pi_1 - \pi_2) \left[\frac{1}{c} - \log \frac{x}{d} - \left(\frac{x}{d}\right)^c \log \frac{x}{d} \right] f_3(x; \Psi)$
d	$w_8(x; \Psi) = (1 - \pi_1 - \pi_2) \left(\frac{x^{c-1}}{d} - \frac{c}{xd} \right) f_3(x; \Psi)$

Table II. Partial ordering depending on the values of the parameters.

Case	Arrangement
(1) $c < 1$	$w_4 \prec w_3 \prec w_1 \prec w_7 \prec w_2 \prec w_8 \prec w_6 \prec w_5$
(2) $(c > 1)$ or $(c = 1, b > d)$ or $(c = 1, b = d, a > 2)$	$w_4 \prec w_3 \prec w_1 \prec w_6 \prec w_5 \prec w_2 \prec w_7 \prec w_8$
(3) $(c = 1, b < d)$	$w_4 \prec w_3 \prec w_1 \prec w_7 \prec w_2 \prec w_8 \prec w_6 \prec w_5$
(4) $(c = 1, b = d, 1 < a < 2)$	$w_4 \prec w_3 \prec w_1 \prec w_6 \prec w_7 \prec w_5 \prec w_2 \prec w_8$
(5) $(c = 1, b = d, a < 1)$	$w_4 \prec w_3 \prec w_1 \prec w_7 \prec w_6 \prec w_2 \prec w_8 \prec w_5$
(6) $(c = 1, b = d, a = 2)$	$w_4 \prec w_3 \prec w_1 \prec w_6 \prec w_5 \prec w_7 \prec w_2 \prec w_8$

or, equivalently, $\gamma_8 = (b/(1 - \pi_1 - \pi_2))\gamma_2$. Consequently,

$$\gamma_2 \left[w_2(x; \Psi) + \frac{b}{1 - \pi_1 - \pi_2} w_8(x; \Psi) \right] + \gamma_6 w_6(x; \Psi) + \gamma_5 w_5(x; \Psi) = 0$$

According to the defined order,

$$\left[w_2 + \frac{b}{1 - \pi_1 - \pi_2} w_8 \right] \prec w_6 \prec w_5$$

Dividing by $w_2 + (b/(1 - \pi_1 - \pi_2))w_8$ and proceeding analogously, we obtain the nullity of the rest of the coefficients. A similar reasoning can be applied to cases (5) and (6).

Consequently, the information matrix is positive definite and, as a conclusion, we can assure that ML estimators of finite mixture models of $\mathscr{C}$ have good asymptotic properties.

2.3. Estimation through EM algorithm

In the previous subsections, we have analyzed the problem of identifiability and the properties of ML estimators. The aim of this section is to study their calculation. Mixtures of distributions under our study constitute a parametric family depending on eight parameters, and the calculation

of ML estimators needs to deal with a quite complex system of nonlinear equations. According to the recommendations of Mclachlan and Krishnan [26] we have tackled this problem by using the EM algorithm. We will just sketch the adaptation of this algorithm to our case.

The first problem is the election of the initial values for Ψ ($\Psi^{(0)}$), in order to assure the convergence with an acceptable speed to the biggest local maximum. It consists in the election of the initial weights of the component functions ($\pi_1^{(0)}$ and $\pi_2^{(0)}$) and their parameters ($\mu^{(0)}$, $\sigma^{(0)}$, $a^{(0)}$, $b^{(0)}$, $c^{(0)}$ and $d^{(0)}$). For $\pi_1^{(0)}$ and $\pi_2^{(0)}$, we propose the strategy suggested by Karlis and Xekalaki [27]: we take several initial values running over the range of possible values, $\pi_1, \pi_2 \in (0, 1)$, and choose those that reach the biggest likelihood. This strategy prevents falling in regions where the likelihood is nearly constant, and detect, if they exist, several local maxima and choose, consequently, the global maximum.

For the initial values of the parameters of the component functions, we consider independently the ML estimators of each family.

- Lognormal estimators:

$$\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \log x_i, \quad \sigma^{(0)} = \left[\frac{1}{n}\sum_{i=1}^{n}(\log x_i - \mu^{(0)})^2\right]^{1/2}$$

- Gamma estimators (following Greenwood and Durand [28]):

$$a^{(0)} = s\left(\log\frac{\overline{x}}{\widetilde{x}}\right) \quad \text{and} \quad b^{(0)} = \frac{1}{na^{(0)}}\sum_{i=1}^{n} x_i$$

where $\overline{x}$ and $\widetilde{x}$ are, respectively, the arithmetic and geometric means, and

$$s(t) = \begin{cases} \dfrac{0.5000876 + 0.1648852t - 0.0544276t^2}{t}, & 0 \leqslant t < 0.5772 \\[2mm] \dfrac{8.898919 + 9.05995t + 0.9775373t^2}{t(17.79728 + 11.968477t + t^2)}, & 0.5772 \leqslant t < 17 \\[2mm] \dfrac{1}{t}, & 17 \leqslant t \end{cases}$$

- Weibull estimators (applying Bain and Antle's method [29]):

$$c^{(0)} = \prod_{i=1}^{n} x_{(i)}^{1/n} \bigg/ \prod_{i=1}^{n} \varepsilon_i^{1/nd^{(0)}} \quad \text{and} \quad d^{(0)} = \frac{\sum_{i=1}^{n}\log\varepsilon_i\log x_{(i)} - \dfrac{1}{n}\sum_{i=1}^{n}\log\varepsilon_i\sum_{i=1}^{n}\log x_{(i)}}{\sum_{i=1}^{n}\log^2 x_{(i)} - \dfrac{1}{n}(\sum_{i=1}^{n}\log x_{(i)})^2}$$

where $\varepsilon_i = \sum_{j=1}^{i} 1/(n - j + 1)$ and $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is the ordered sample.

The second problem is to improve the low speed of convergency of the algorithm. We propose to use the conjugate gradient acceleration method [30].

Finally, the stop criterion is determined by the parameter variation between two consecutive steps, the variation measured by the sum of the absolute values of each variation parameter. Following Karlis and Xekalaki [27], the value 10^{-5} gives an appropriate stop criterion.

3. SIMULATION

This section presents a simulation study to analyze the performance of the proposed method. Two different strategies have been used:

- *Simulation A*: For a set of 128 combinations of values of the parameters, 128 samples have been generated (with size $n = 100$) and used to apply the proposed methodology.
- *Simulation B*: Fifteen combinations have been selected from among the 128 previous ones, and 100 samples of size $n = 500$ have been generated and used to apply the proposed methodology to each individual combination.

The combinations of values of the parameters have been selected by fitting several samples of the length of hospital stay variable to each component function separately.

The goal of Simulation A is to get conclusions about the adequacy of the proposed methodology in a wide range of possible values of the eight parameters of the model, so that the possible values related to the variable LOS are covered. Simulation B is carried out to analyze the performance of the proposed EM algorithm.

3.1. Simulation A

The considered mixture model has eight parameters, $\Psi = (\mu, \sigma, a, b, c, d, \pi_1, \pi_2)$. We have considered two different values for each of them, obtaining 64 possible combinations:

$$\mu = 2, 3; \quad \sigma = 0.5, 0.7; \quad a = 7, 10; \quad b = 2, 3; \quad c = 1, 2; \quad d = 15, 25$$

For each of these 64 combinations, we have randomly generated two pairs of values for the weights π_1 and π_2.

For a given parameter $\Psi^g = (\pi_1^g, \pi_2^g, \mu^g, \sigma^g, a^g, b^g, c^g, d^g)$, the associated sample has been generated using the following procedure.

Consider the intervals $I_1 = [0, \pi_1^g]$, $I_2 = (\pi_1^g, \pi_1^g + \pi_2^g]$ and $I_3 = (\pi_1^g + \pi_2^g, 1]$; generate n random values $\{u_i : i = 1, \ldots, n\}$ from a uniform distribution on the interval $[0, 1]$, and for each value u_i:

- if $u_i \in I_1$ then generate a random value x_i from a Lognormal distribution with parameters μ^g and σ^g;
- if $u_i \in I_2$ then generate a random value x_i from a Gamma distribution with parameters a^g and b^g;
- if $u_i \in I_3$ then generate a random value x_i from a Weibull distribution with parameters a^g and b^g.

As a discrepancy measure between the original $F(x, \Psi^g)$ and the estimated distribution function $F(x, \widehat{\Psi})$, the uniform measure [31], also called Kolmogorov measure [32], is proposed:

$$d(\Psi^g, \widehat{\Psi}) = \sup_{x \in \mathbb{R}} |F(x, \Psi^g) - F(x, \widehat{\Psi})|$$
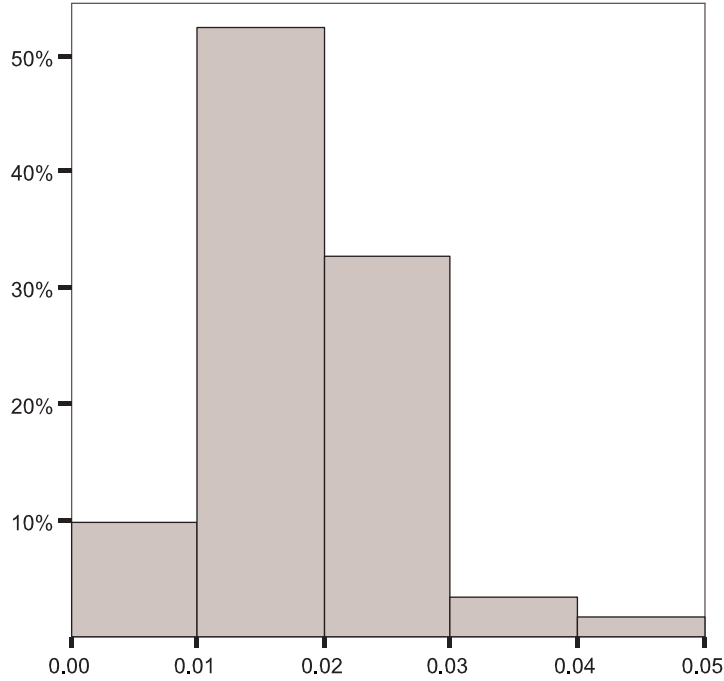
Figure 1. Discrepancy measure: histogram (Simulation A).

Figure 1 represents the obtained results for the 128 generated samples, which should be considered satisfactory, since for 93.8 per cent of the samples the discrepancy measure is less than 0.03, and for none of them is this value bigger than 0.05.

3.2. *Simulation B*

Fifteen among the 128 combinations generated in Simulation A are selected. For each of them:

- the expected value (η) and the variance (v^2) of the associated distribution mixture are calculated. Both values are functions of the parameters $\eta = g_1(\Psi)$ and $v^2 = g_2(\Psi)$.
- $B = 100$ samples of size $n = 500$ are generated. The following statistics are calculated for each sample:

  - The ML estimator of Ψ using the method proposed in this paper, $\{\widehat{\Psi}_k : k = 1, \ldots, B\}$.
  - The expected value and the variance of the associated mixture distribution, $\{(\widehat{\eta}_k, \widehat{v}_k^2) : k = 1, \ldots, B\}$, where $\widehat{\eta}_k = g_1(\widehat{\Psi}_k)$ and $\widehat{v}_k^2 = g_2(\widehat{\Psi}_k)$.
  - The absolute deviation $b(\widehat{\eta}_k) = |\widehat{\eta}_k - \eta|$ and the variance ratio $r(\widehat{v}_k^2) = \widehat{v}_k^2 / v^2$.
  - The discrepancy measure $d(\widehat{\Psi}_k) = d(\widehat{\Psi}_k, \Psi)$.

- The following values have been calculated: estimated absolute bias (EAB), maximum deviation (MD), mean variance ratio (MVR), minimum variance ratio (MINVR), maximum variance

Table III. Fitting measures for Simulation B.

	η	EAB	MD	v^2	MVR	MAXVR	MINVR	d	$d_{\max}$
1	141 357	0.3559	1.1502	906 499	0.9968	1.1583	0.8043	0.0195	0.0639
2	200 132	0.4877	1.6479	1 813 262	1.0025	1.3038	0.7529	0.0190	0.0307
3	254 135	0.6867	2.1972	3 114 499	1.0066	1.2636	0.6177	0.0191	0.0622
4	114 138	0.3153	1.2740	794 162	1.0175	1.3899	0.6226	0.0241	0.0695
5	158 305	0.4212	1.4823	1 071 434	1.0028	1.2873	0.4346	0.0206	0.0825
6	180 735	0.3916	1.6767	1 185 501	0.9540	1.2497	0.4195	0.0276	0.1072
7	197 650	0.3916	1.6767	946 450	0.9540	1.2497	0.4195	0.0231	0.1257
8	213 800	0.8734	2.2910	1 215 532	1.1327	1.4935	0.7263	0.0197	0.0340
9	204 744	0.4720	1.7014	1 788 800	1.0033	1.2520	0.7001	0.0257	0.1181
10	221 625	0.3851	1.5176	1 019 930	1.0152	1.2946	0.7644	0.0200	0.0695
11	240 987	0.5436	1.7358	1 742 483	0.9802	1.3489	0.7041	0.0216	0.0394
12	244 558	0.7408	2.1343	3 843 060	1.0487	1.7494	0.5714	0.0235	0.0701
13	238 942	0.5934	2.0430	3 342 024	1.0643	1.5241	0.6602	0.0193	0.0366
14	224 377	0.4332	1.8552	1 607 892	1.0510	1.6321	0.6654	0.0236	0.0492
15	270 890	0.7565	2.5476	3 300 545	1.0367	1.5315	0.5284	0.0204	0.0549

ratio (MAXVR), mean discrepancy ($\overline{d}$) and maximum discrepancy ($d_{\max}$), where

$$\text{EAB} = \frac{1}{B} \sum_{k=1}^{B} b(\widehat{\eta}_k), \quad \text{MD} = \max_{k=1,\ldots,B} b(\widehat{\eta}_k)$$

$$\text{MVR} = \frac{1}{B} \sum_{k=1}^{B} r(\widehat{v}_k^2), \quad \text{MINVR} = \min_{k=1,\ldots,B} r(\widehat{v}_k^2), \quad \text{MAXVR} = \max_{k=1,\ldots,B} r(\widehat{v}_k^2)$$

$$\overline{d} = \frac{1}{B} \sum_{k=1}^{B} d(\widehat{\Psi}_k), \quad d_{\max} = \max_{k=1,\ldots,B} d(\widehat{\Psi}_k)$$

The obtained results are presented in Table III and drawn in Figure 2 in order to graphically illustrate the adequacy of this methodology. We should stress the following aspects:

1. In order to study the fitting achieved for the distribution, three interesting features are considered: the expected value, as a measure of localization of the distribution, the variance, as a dispersion measure; and the cumulative distribution function. The considered fitting measures are, respectively, the absolute deviation, the variance ratio and the discrepancy measure.

2. The EAB for the expected value of the mixture distribution is less than 1 in all cases. Obviously, EAB grows with the expected value. This growth is, however, not very significant.

3. With respect to the estimation of the variance of the distribution, the MVR is always less than 10 per cent, except in one case (13.27 per cent).

4. The discrepancy among the cumulative distribution functions is almost insignificant. In all cases, the mean discrepancy is less than 3 per cent.

These results confirm the goodness of fit of the proposed procedure for the fitting of a mixture distribution in the class of mixtures considered in this paper.
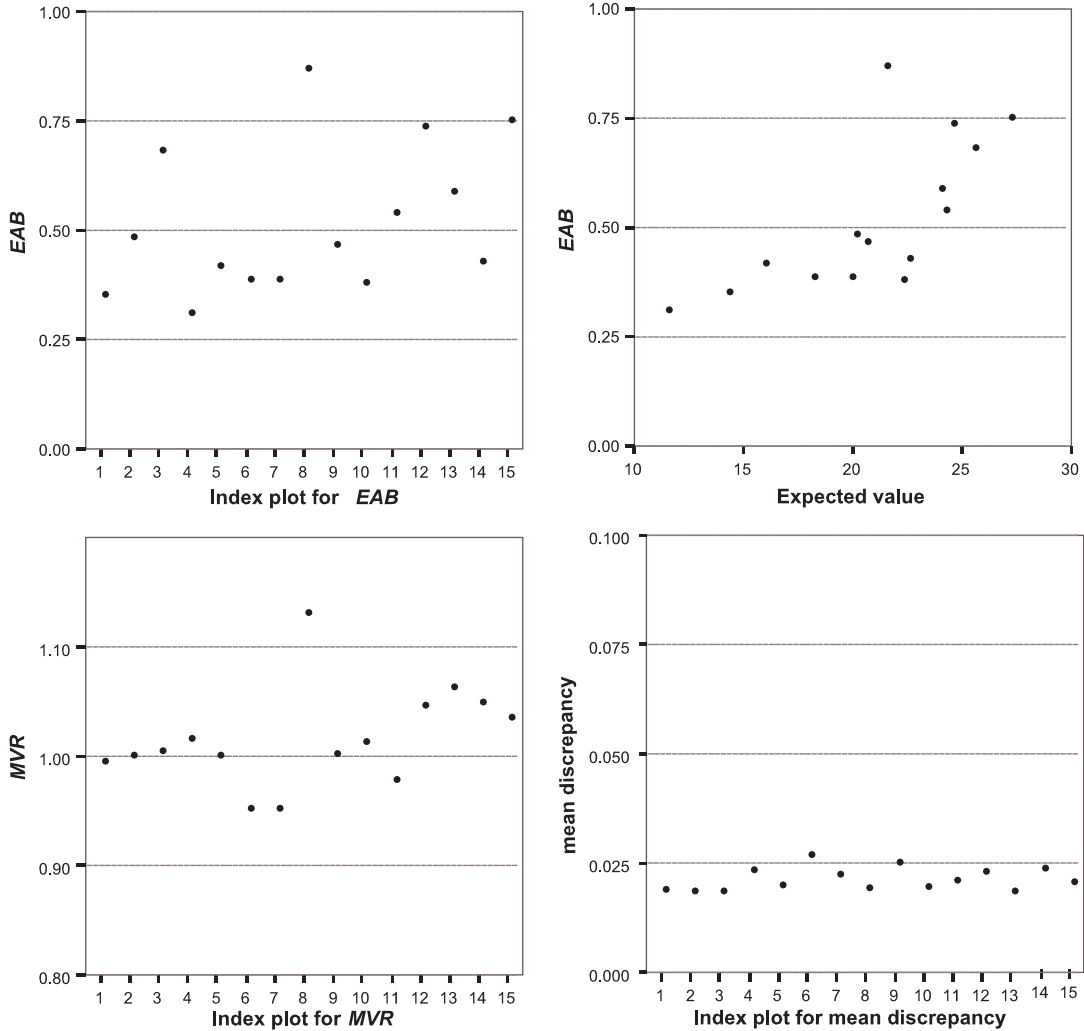
Figure 2. Plots of fitting measures for Simulation B.

4. APPLICATION OF LENGTH OF HOSPITAL STAY

To illustrate the adequacy of this mixture model to the hospital stay, we present the results obtained by applying it to different sets of data of hospital stay obtained from the University Virgen Macarena of Sevilla Hospital (Spain), from different DRGs during fixed periods of time.

The considered DRGs are: DRG 14, 'Specific cerebrovascular disorders except transient ischemic attack', from 1999 to 2002; DRG 88, 'Chronic obstructive pulmonary disease', from 1995 to 1998; DRG 122, 'Circulatory disorders with acute myocardial infarction without cardiovascular complications, discharged alive', from 1995 to 1998; DRG 127, 'Heart failure and shock', from

Table IV. Modelization of LOS variable for several DRGs: MLE and discrepancy measure with empirical distribution.

DRG	Model	n	d_3	μ	σ	a	b	c	d	π_1	π_2
14		1579									
	Lognormal		0.105	2.51	0.76						
	Gamma		0.097			8.35	1.56				
	Weibull		0.096					1.64	17.4		
	Mixture		**0.038**	2.29	1.08	4.96	4.23	3.33	13.5	0.343	0.244
88		1156									
	Lognormal		0.069	2.15	0.78						
	Gamma		0.076			6.10	1.51				
	Weibull		0.085					1.60	12.2		
	Mixture		**0.045**	3.41	0.49	2.61	3.52	35.6	18.7	0.086	0.899
122		1086									
	Lognormal		0.101	2.41	0.56						
	Gamma		0.101			3.89	3.27				
	Weibull		0.124					2.22	14.3		
	Mixture		**0.052**	2.27	1.01	6.22	2.39	4.22	10.6	0.156	0.443
127		1639									
	Lognormal		0.076	2.13	0.706						
	Gamma		0.058			4.59	2.05				
	Weibull		0.075					1.77	11.7		
	Mixture		**0.039**	1.99	0.80	4.40	2.66	7.71	9.64	0.378	0.374
541		1528									
	Lognormal		0.068	2.4	0.73						
	Gamma		0.083			6.85	1.73				
	Weibull		0.096					1.71	15.3		
	Mixture		**0.038**	3.08	0.15	1.92	13.3	2.25	11.1	0.064	0.233

1999 to 2002; DRG 541, 'Respiratory diseases except infection, bronchitis and asthma', from 1999 to 2002.

In Table IV, the values of the uniform measure d_s between the distribution determined by the estimated parameters and the empirical distribution of the sample are shown. Furthermore, the estimated parameters for each individual model (Lognormal, Gamma and Weibull) and for the mixture model are also given.

It can be observed that the discrepancy measure d_s is significantly smaller for the mixture model than for each of the individual models. Besides, as pointed out by Marazzi *et al.* [6], the adequate individual model depends on each DRG. For example, a Weibull distribution seems to be the most adequate model for DRG 14, a Lognormal distribution for DRGs 88 and 541, and a Gamma model for DRGs 122 and 127, whereas the mixture of these three distributions gives us the best result in all cases.

The histogram and density function of the estimated mixture model are presented in Figure 3 which shows an adequate level of fitting. In Table V we summarize a comparative analysis of the proposed mixture and finite mixtures of one of the three considered families (Lognormal,
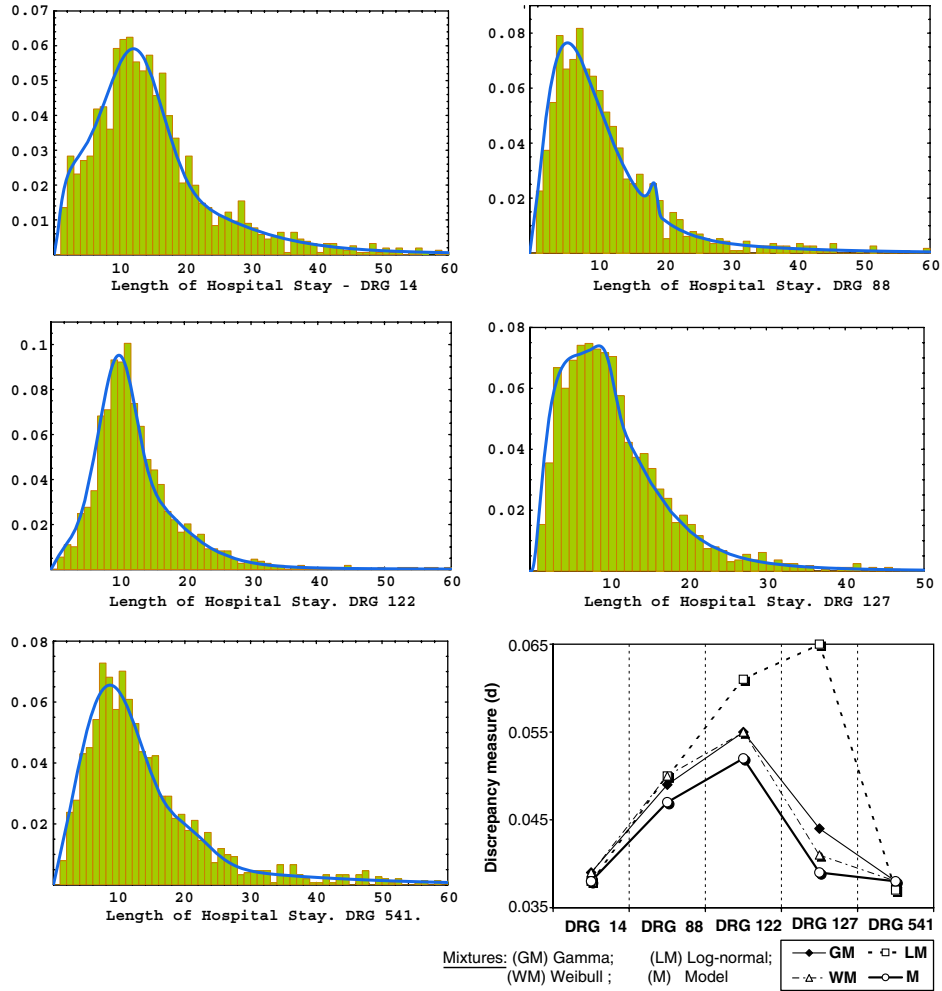
Figure 3. Graphics on the modelization of LOS variable for several DRGs.

Table V. Measure d_s between the estimated distribution and the empirical distribution.

Mixture model	DRG 14	DRG 88	DRG 122	DRG 127	DRG 541
Lognormal	0.038	0.050	0.061	0.065	0.037
Gamma	0.039	0.049	0.056	0.044	0.038
Weibull	0.039	0.050	0.055	0.041	0.038
Proposed model	0.038	0.045	0.049	0.038	0.038

Gamma and Weibull). More specifically, for each family of distributions, mixtures of two or three components have been considered. In Table V and Figure 3, the discrepancy measures for mixtures of three components are shown.

A comparative graphic of these results is presented in Figure 3 which shows that the proposed model is, in general, the most optimal one. Furthermore, from Figure 3, and Table V, we can infer several aspects:

- In the proposed model, the discrepancy measure remains below 0.05.
- The adequate individual mixture model depends on each DRG (Lognormal mixtures for DRGs 14 and 541, Gamma mixture for DRG 88, Weibull mixtures for DRGs 122 and 127).
- The proposed mixture model improves the results obtained for the considered DRGs.

Consequently, our model (three components and eight parameters) improves the results obtained with mixtures of three distributions from the same family (with the same number of components and parameters as in the proposed model).

It can be seen that the following three distributions:

- the distribution (A) that determines the most adequate individual model,
- the distribution (B) that determines the most adequate individual mixture model,
- the most prevalent distribution (C) in the proposed mixture model,

do not coincide.

For example, for DRG 127, these distributions are, respectively, Gamma, Weibull and Lognormal. This situation can be explained as follows: distribution A solves an optimization problem in all individual distributions; distribution B solves an optimization problem in the individual mixture distributions; and distribution C is the one with the biggest weight in the result of the optimization problem in a different family of mixture distributions. Therefore, each distribution is the result of three different optimization problems. Consequently, this situation should not be considered as contradictory or atypical.

5. DISCUSSION

LOS is an easily available indicator of hospital activity. It is an indirect estimator of resources consumption and of the efficiency of a Hospital, so that the mean of the LOS is used as an indicator of the consumption of resources. With respect to this, Marazzi *et al.* state textually: 'unfortunately, LOS distributions are skewed and contain outliers, making the use of this simple statistic questionable'. Therefore, the fitting of an asymmetric distribution is a more global and optimal solution. This solution allows one to obtain estimations of expected value, dispersion measures, median, quartiles and other percentiles. Accordingly, we consider this approach as a good method for the statistical analysis of this kind of data, where the main goal is to determine measures of resource consumption of groups of patients.

In the scientific literature, some families of distributions have been proposed to model this variable. The most used asymmetric distributions are Lognormal, Gamma and Weibull. But as Lee *et al.* [10] point out, it is more convenient to tackle the problem by means of mixture distributions. However, the family determined by the most adequate mixture distribution depends on the given DRG. Consequently, a useful general model for all DRGs is not available.

The objective of this work is to analyze the variable LOS by means of a general model: a finite mixture of distributions in the union of Lognormal, Gamma and Weibull families. This approach can be used to model the variable LOS for most of the DRGs.

This paper has firstly analyzed the theoretical study of the model, tackling the classic problems in mixture distributions: identifiability, asymptotic properties of the ML estimators and the adaptation of the EM algorithm. The adequacy of the algorithm has been analyzed through two simulation procedures. Finally, the obtained results for different DRGs, improving upon other procedures, show the validity of the proposed methodology.

Another aspect that must be discussed is that the proposed model is more complicated than any individual one. However, being a scientific research, we have to make a decision between 'the complexity of the model and the computation procedure' and 'the model fitting'. We decide to take the second option, since the computational effort and the complexity should not be considered as excessive. Furthermore, this methodology seems more adequate for most of the DRGs. The model based on a unique family of distributions is less complex but has two other aspects. First, the necessity of previously selecting the family among the most usual asymmetric distributions. Second, the results provided are less optimal. Therefore, the study of a simpler model does not imply either a significative reduction of the computational effort or better results.

In conclusion, we believe that this work contributes to the development of the statistical analysis of LOS distributions and other consumption variables in health services. Also this approach can be applied to other asymmetric data (for instance, the length of wait for surgical procedures or for medical attention).

APPENDIX A

Proposition 1

Let $\mathscr{F} = \{f(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^t \in \Theta\}$ be a family of density functions for which there exist the partial derivatives with respect to each θ_i and they are integrable with respect to x for all $i \in \{1, \ldots, d\}$. If

$$\int \frac{\left(\dfrac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_i}\right)^2}{f(x; \boldsymbol{\theta})} \, \mathrm{d}x < +\infty$$

then the Fisher information matrix of any finite mixture $q(x; \Psi) = \sum_{h=1}^{K} \pi_h f(x; \boldsymbol{\theta_h})$ of elements from $\mathscr{F}$ is well defined.

Proof

The information matrix is well defined provided that there exist the partial derivatives of $\log q(x; \Psi)$ and the products

$$\left(\frac{\partial \log q(x; \Psi)}{\partial \psi_i}\right)\left(\frac{\partial \log q(x; \Psi)}{\partial \psi_j}\right) q(x; \Psi) = \frac{1}{q(x; \Psi)}\left(\frac{\partial q(x; \Psi)}{\partial \psi_i}\right)\left(\frac{\partial q(x; \Psi)}{\partial \psi_j}\right)$$

are integrable with respect to x for every pair i, j. The previous quotient can be expressed as a sum of functions of one of the forms:

$$\frac{f_h(x, \boldsymbol{\theta}_h) f_l(x, \boldsymbol{\theta}_l)}{q(x; \Psi)}, \quad \frac{f_h(x, \boldsymbol{\theta}_h)\pi_l\left(\dfrac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}}\right)}{q(x; \Psi)}, \quad \frac{\pi_h \pi_l\left(\dfrac{\partial f_h(x, \boldsymbol{\theta}_h)}{\partial \theta_{hs}}\right)\left(\dfrac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}}\right)}{q(x; \Psi)}$$

Using the bound

$$q(x; \Psi) \geqslant \pi_h f_h(x, \boldsymbol{\theta}_h), \quad h = 1, \ldots, K \tag{A1}$$

yields

$$\left| \frac{f_h(x, \boldsymbol{\theta}_h) f_l(x, \boldsymbol{\theta}_l)}{q(x; \Psi)} \right| \leqslant \frac{1}{\pi_h} f_l(x, \boldsymbol{\theta}_l)$$

$$\left| \frac{f_h(x, \boldsymbol{\theta}_h) \pi_l}{q(x; \Psi)} \right| \left| \frac{\partial f_k(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}} \right| \leqslant \frac{\pi_l}{\pi_h} \left| \frac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}} \right|$$

and both of the right-hand functions are integrable. Finally, using the inequality $ab \leqslant (a^2 + b^2)/2$, $a, b \in \mathbb{R}$, we obtain

$$\frac{\pi_h \pi_l \left(\frac{\partial f_h(x, \boldsymbol{\theta}_h)}{\partial \theta_{hs}} \right) \left(\frac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}} \right)}{q(x; \Psi)} \leqslant \frac{\pi_h \pi_l}{2 q(x; \Psi)} \left[\left(\frac{\partial f_h(x, \boldsymbol{\theta}_h)}{\partial \theta_{hs}} \right)^2 + \left(\frac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}} \right)^2 \right]$$

applying inequality (A1) again, we can bound it by

$$\frac{1}{2} \left[\frac{\left(\frac{\partial f_h(x, \boldsymbol{\theta}_h)}{\partial \theta_{hs}} \right)^2}{f_h(x, \boldsymbol{\theta}_h)} + \frac{\left(\frac{\partial f_l(x, \boldsymbol{\theta}_l)}{\partial \theta_{lr}} \right)^2}{f_l(x, \boldsymbol{\theta}_l)} \right]$$

which is integrable by hypothesis. $\qquad\square$

Proposition 2
Let $f(x; \boldsymbol{\theta})$ be a density function. Its Fisher information matrix in $\boldsymbol{\theta}^0 \in \mathbb{R}^d$, $I(\boldsymbol{\theta}^0)$, is positive definite if and only if the set of functions

$$\{w_1(x, \boldsymbol{\theta}^0), \ldots, w_d(x, \boldsymbol{\theta}^0)\}$$

is linear independent a.e. in the set $S_f = \{x : f(x; \boldsymbol{\theta}) \neq 0\}$, where $w_i(x, \boldsymbol{\theta}^0) = (\partial/\partial \theta_i) f(x; \boldsymbol{\theta})|_{\boldsymbol{\theta} = \boldsymbol{\theta}^0}$.

Proof
Let $M(x; \boldsymbol{\theta})$ denote the matrix $[\nabla_{\boldsymbol{\theta}} \log f(x; \boldsymbol{\theta})][\nabla_{\boldsymbol{\theta}} \log f(x; \boldsymbol{\theta})]^t$, with components

$$m_{ij}(x; \boldsymbol{\theta}) = \left(\frac{\partial \log f(x, \boldsymbol{\theta})}{\partial \theta_i} \right) \left(\frac{\partial \log f(x, \boldsymbol{\theta})}{\partial \theta_j} \right)$$

Then, $\boldsymbol{\zeta}^t I(\boldsymbol{\theta})\boldsymbol{\zeta} = 0$ if and only if $\sum_{i,j} \zeta_i \zeta_j m_{ij}(x; \boldsymbol{\theta}) f(x; \boldsymbol{\theta}) = 0$ for almost all $x \in \mathbb{R}^p$. It means that there exists a null set $Z \subset \mathbb{R}^p$ such that for $x \notin Z$, $(\boldsymbol{\zeta}^t M(x; \boldsymbol{\theta})\boldsymbol{\zeta}) f(x; \boldsymbol{\theta}) = 0$. The latter is equivalent to $\boldsymbol{\zeta}^t M(x; \boldsymbol{\theta})\boldsymbol{\zeta} = 0$ for $x \in S_f \backslash Z$. Now, this equality is equivalent to $((1/f(x; \boldsymbol{\theta}))\boldsymbol{\zeta}^t [\nabla_{\boldsymbol{\theta}} f(x; \boldsymbol{\theta})])^2 = 0$, which finally is $\sum_i \zeta_i w_i(x; \boldsymbol{\theta}) = 0$. Taking into account that $\boldsymbol{\zeta}^t I(\boldsymbol{\theta})\boldsymbol{\zeta} \geqslant 0$, the conclusion of the proposition follows. $\qquad\square$

ACKNOWLEDGEMENTS

REFERENCES

1. Fernonw LC, Mccoll I, Mackie C. Firm, patient, and process variables associated with length of stay in four diseases. *British Medical Journal* 1978; **1**:556–559.
2. Lagoe RJ. A community-based analysis of regional differences in hospital stay by DRGs. *Inquiry* 1986; **23**(2): 183–190.
3. Lee AH, Gracey M, Wang K, Kelvin KW. A robustified modelling approach to analyze pediatric length of stay. *Annals of Epidemiology* 2005; **15**(9):637–677.
4. Fetter RB, Youngsoo S, Freeman JL, Averill RF, Thomson JD. Case-mix; definition by diagnosis relates groups. *Medical Care* 1980; **18**(Suppl. 2):1–53.
5. Shachtman RH, Snapinn SM, Quade D, Freund DA, Kronhaus AK. A method for constructing case-mix indexes, with application to hospital length of stay. *Health Services Research* 1986; **20**:737–762.
6. Marazzi A, Paccaud F, Ruffieux C, Beguin C. Fitting the distributions of length of stay by parametric models. *Medical Care* 1998; **36**(6):915–927.
7. Silberbach M, Shumaker D, Menashe V, Cobanoglu A, Morris C. Predicting hospital charge and length of stay for congenital heart disease surgery. *American Journal of Cardiology* 1993; **72**:958–963.
8. Wolfe MW, Roubin GS, Schweiger M. Length of hospital stay and complications after percutaneous transluminal coronary angioplasty. *Circulation* 1995; **92**:311–319.
9. Xiao J, Lee AH, Vemuri SR. Mixture distribution analysis of length of hospital stay for efficient funding. *Socio-economic Planning Sciences* 1999; **33**(1):39–59.
10. Lee AH, Ng ASK, Yau KKW. Determinants of maternity length of stay: a gamma mixture risk-adjusted model. *Health Care Management Science* 2001; **4**:249–255.
11. Quantin C, Sauleau E, Bolard P, Mousson C, Kerkri M, Brunet Lecomte P, Moreau T, Duserre L. Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology* 1999; **52**(3):251–258.
12. Quantin C, Entezam F, Brunet-Lecomte P, Lepage E, Guy H, Duserre L. High cost factors for leukaemia and lymphoma patients: a new analysis of costs within these diagnosis related groups. *Journal of Epidemiology and Community Health* 1999; **53**:24–31.
13. Lee AH, Xiao J, Codde JP, Ng ASK. Public versus private hospital maternity length of stay: a gamma mixture modelling approach. *Health Services Management Research* 2002; **15**(1):46–54.
14. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York, 1985.
15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* (*Statistical Methodology*) 1977; **39**:1–38.
16. Teicher H. Identifiability of finite mixtures. *The Annals of Mathematical Statistics* 1963; **38**(4):1300–1302.
17. Chandra S. On the mixtures of probability distributions. *Scandinavian Journal of Statistics* 1977; **4**:105–112.
18. Khalaf EA. Identifiability of finite mixtures using a new transform. *Annals of the Institute of Statistical Mathematics* 1988; **40**:261–265.
19. Brandorff-Nielsen O. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications* 1965; **12**(1):115–121.
20. Henna J. Examples of identifiable mixture. *Journal of the Japan Statistical Society* 1994; **24**:193–200.
21. Atienza N, García-Heras J, Muñoz-Pichardo JM. A new condition for identifiability of finite mixture distributions. *Metrika* 2005; **63**(2):215–221.
22. Chanda KC. A Note on the consistency and maxima of the roots of likelihood equations. *Biometrika* 1954; **41**:56–61.
23. Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 1984; **26**:195–239.
24. Atienza N, García-Heras J, Muñoz-Pichardo JM, Villa R. On the consistency of MLE in finite mixture models of exponential families. *Journal of Inference Statistics Planning* 2006; **137**(2):496–505. DOI: 10.1016/j.jspi.2005.12.014.

25. Atienza N, García-Heras J, Muñoz-Pichardo JM. Consistency of maximum likelihood estimators in finite mixture models of the union of W-type families. *Communications in Statistics*, *Theory and Methods* 2005; **34**:1471–1485.
26. Mclachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. Wiley: New York, 1997.
27. Karlis D, Xekalaki E. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis* 2002; **41**:577–590.
28. Greenwood JA, Durand D. Aids for fitting the Gamma distribution by maximum likelihood. *Technometrics* 1960; **2**:55–56.
29. Bain LJ, Antle CE. Estimation of parameters in the Weibull distribution. *Technometrics* 1967; **9**:621–627.
30. Jamshidian M, Jenrich RI. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* 1993; **88**:221–228.
31. Zolotarev VM. Probability metrics. *Theory of Probability and its Applications* 1983; **28**:278–302.
32. Kolmogorov AN. Sulla determinazione empririca di una legge di distribuzione. *Giornalle dell'Instituto Italiano degli Attuari* 1933; **4**:83–91.