ISSN: 1696-2281
SORT 29 (2) July-December 139-292 (2005)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 29 (2), July-December 2005

Formerly Qüestiió

Contents

SORT 29 (2) July-December 2005, 141-164

Likelihood for random-effect models

Youngjo Lee[1] and John A. Nelder[2]

[1]*Seoul National University,* [2]*Imperial College London, U.K.*

Abstract

For inferences from random-effect models Lee and Nelder (1996) proposed to use hierarchical likelihood (h-likelihood). It allows inference from models that may include both fixed and random parameters. Because of the presence of unobserved random variables h-likelihood is not a likelihood in the Fisherian sense. The Fisher likelihood framework has advantages such as generality of application, statistical and computational efficiency. We introduce an extended likelihood framework and discuss why it is a proper extension, maintaining the advantages of the original likelihood framework. The new framework allows likelihood inferences to be drawn for a much wider class of models.

MSC: 62F10 62F15 62F30

Keywords: generalized linear models, hierarchical models, h-likelihood.

1 Introduction

Ever since Fisher introduced the concept of likelihood in 1921, the likelihood function has played an important part in the development of both the theory and the practice of statistics. The likelihood framework has advantages such as generality of application, algorithmic *wiseness* (Efron, 2003), consistency and asymptotic efficiency, which can be summarized as computational and statistical efficiency. Savage (1976) states "The most fruitful, and for Fisher, the usual definition of the likelihood associated with an observation is the probability or density of observation as a function of the parameter, modulo a multiplicative constant." Edwards (1972, pp. 12) similarly defines "the likelihood $L(H|R)$ of the hypothesis H given data R, and a specific model, to

Address for correspondence: Y. Lee, Department of Statistics, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea. E-mail: youngjo@plaza.snu.ac.kr
Received: October 2005

be proportional to $P(R|H)$, the constant of proportionality being arbitrary. However, when the problem does not fit into the usual parametric framework, the definition of the likelihood function is not immediately obvious.

Suppose that a model class consists of three types of object, observable random variables (data), unobservable (or unobserved) random variables and unknown fixed parameters. Special cases are subject-specific inferences for random-effect models, prediction of unobserved future observations, missing data problems etc. Consider the simplest example of a 2-level hierarchy with the model

$$y_{ij} = \beta + u_i + e_{ij},$$

where $u_i \backsim N(0, \lambda)$ and $e_{ij} \backsim N(0, \phi)$ with u_i and e_{ij} uncorrelated. This model leads to a specific multivariate distribution. From one point of view the parameters of the model are β, λ and ϕ, and it is straightforward to write down the likelihood from the multivariate normal distribution and to obtain estimates by maximizing it. However, although the u_i are thought of initially as having been obtained by sampling from a population, once a particular sample has been obtained they are fixed quantities and estimates of them will often be of interest (Searle et al 1992). The likelihood based upon the multivariate normal distribution provides no information on these quantities.

There have been several attempts to extend likelihood beyond its use in parametric inference to more general models that include unobserved random variables; see, for example, Henderson (1975) and Lee and Nelder (1996) for random-effect models, and Yates (1933) and Box *et al.* (1970) for missing data problems. Except for Lee and Nelder, these extensions have been successful only for inference of location parameters in limited classes of models. Pearson (1920) pointed out the limitation of Fisher likelihood inferences in prediction. As a likelihood solution various predictive likelihoods have been proposed (Bjørnstad, 1990): see Barndorff-Nielsen and Cox (1996) for interval estimates. Interpretation of the profile predictive likelihood approach of Mathiasen (1979) in the h-likelihood perspective is in Pawitan (2001, Chapter 16). In this paper we concentrate on likelihood inferences for random effects.

Bayarri *et al.* (1988) considered the following example: There is a single fixed parameter θ, a single unobervable random quantity U and a single observable quantity Y. The unobserved random variable U has a probability function

$$f_\theta(u) = \theta \exp(-\theta u) \quad \text{for } u > 0, \quad \theta > 0,$$

and an observable random variable Y has conditional probability function

$$f_\theta(y|u) = f(y|u) = u \exp(-uy) \quad \text{for } y > 0, \quad u > 0,$$

free of θ. Throughout this paper we use $f_\theta()$ as probability functions of random variables with fixed parameters θ; the arguments within the brackets can be either conditional or

unconditional. Thus, $f_\theta(y|u)$ and $f_\theta(u|y)$ have different functional forms even though we use the same $f_\theta()$ to mean probability functions with parameters θ.

Starting from a basic definition that the likelihood function is proportional to $L(r|\theta)$ where r and θ denote two derived classes, Bayarri *et al.* (1988) argue that there is no unique way of deciding which of the three classes should be regarded as part of r and which part of θ and, furthermore, that there is no unique way of deciding which random variables and parameters should explicitly enter the likelihood function. They consider three possibilities for an extended-likelihood for three objects:

$$\begin{aligned} L(y|\theta) &\equiv f_\theta(y) = \int f(y|u)\, f_\theta(u) du = \theta/(\theta+y)^2, \\ L(y|u,\theta) &\equiv f(y|u) = u\exp(-uy), \\ L(u;y|\theta) &\equiv f(y|u)\, f_\theta(u) = u\theta\exp\{-u(\theta+y)\}. \end{aligned}$$

Here "$L(y|\theta) \equiv f_\theta(y)$" means that $L(y|\theta) = f_\theta(y)c(y)$ for some $c(y) > 0$. The *marginal* likelihood $L(y|\theta)$ gives the (marginal) maximum-likelihood (ML) estimator $\hat\theta = y$, but is totally uninformative about the unknown value of u of U. The *conditional* likelihood in the form $L(y|u,\theta)$, which may be regarded as

$$L(\text{observed}|\text{unobserved}),$$

is uninformative about θ and loses the relationship between u and θ reflected in $f_\theta(u)$. Finally, the *joint* likelihood $L(u;y|\theta)$, which may be regarded as

$$L(\text{random variables}|\text{parameters}),$$

yields, if maximized with respect to θ and u, the useless estimators $\hat\theta = \infty$ and $\hat u = 0$. Bayarri *et al.* (1988) therefore concluded that none is useful as a likelihood for more general inferences.

In extended likelihood dividing the three types of object into two derived classes would be confusing. For example the empirical Bayes method uses $f_\theta(u|y)$, which seems to belong to $L(\text{observed}|\text{unobserved})$. If so it cannot be distinguished from $L(y|u,\theta) \equiv f_\theta(y|u)$. In this paper $L(a;b)$ denotes the likelihood for the argument a using the probability function $f_\theta(b)$. Here $L(\theta,u;u|y) \equiv f_\theta(u|y)$ and $L(\theta,u;y|u) \equiv f_\theta(y|u)$. We use capital letters such as L for likelihood and lowercase letters such as $l = \log L$ for log likelihood which we shall abbreviate to *loglihood* (a useful contraction which we owe to Michael Healy).

In this paper we resolve Bayarri *et al.*'s (1988) problem by showing that it is possible to make unambiguous likelihood inferences about a wide class of models having unobserved random variables. If in this example, instead of the joint likelihood $L(\theta,u;y,u)$, we use a particular form of it, the so-called h-likelihood

$$L(\theta, u; y, \log u) \equiv f(y|\log u) f_\theta(\log u) = u^2 \theta \exp\{-u(\theta + y)\}$$

maximization gives the ML estimator $\hat{\theta} = y$, and the random effect estimator $\hat{u} = 1/y$.

When θ is known, the best predictor (BP; Searle et al 1992, pp. 261) of u is defined by

$$\hat{u} = E(u|y) = 2/(\theta + y).$$

When θ is unknown, the h-likelihood gives the empirical BP of u

$$\hat{u} = \widehat{E(u|y)} = 2/(\theta + y)|_{\theta=\hat{\theta}} = 1/y.$$

The word *predictor* has often been used for random-effect estimates. For the prediction of unobserved future observations we believe that it is the right one to use. However, for inference about unknown random effects the word *estimate* seems more appropriate because we are estimating unknown u, fixed once the data y are known, though possibly changing in future samples.

Our goal is to establish an extended likelihood framework by showing that the h-likelihood is a proper extension of the Fisher likelihood to random-effect models, maintaining the original likelihood framework for parametric inferences. In Section 2 we define the h-likelihood for hierarchical generalized linear models (HGLMs). In Section 3, we show why the h-likelihood, among joint likelihoods, should be used, and in Section 4 we illustrate why we need to distinguish two classes of parameters, fixed effects (for location) and dispersions. In Section 5, we describe the extended likelihood framework, and in Section 6 we illustrate extended likelihood inferences using Bayarri *et al.*'s (1988) example. We also explain how our method differs from Breslow and Clayton's (1993) penalized-quasi-likelihood (PQL) method and illustrate why the latter suffers from severe bias. In Section 7 we discuss how the extended framework preserves the advantages of the original likelihood framework, giving our conclusions in Section 8.

2 HGLMS

HGLMs are generalized linear models (GLMs) in which the linear predictor contains both fixed and random parameters: They take the form

$$\mu = E(y|u) \text{ and } var(y|u) = \phi V(\mu)$$

with a linear predictor

$$\eta = g(\mu) = X\beta + Zv, \tag{1}$$

where $g()$ is a GLM link function, X and Z are model matrices for fixed and random parameters (effects) respectively, and $v_i = v(u_i)$ are random effects after some transformation $v()$. For simplicity we consider the case of just one random vector u.

Here the joint density of the responses y and the random effects u can be used to define a *joint likelihood*

$$L(\theta, u; y, u) \equiv f_{\beta,\phi}(y|u) f_\lambda(u), \tag{2}$$

where $\theta = (\beta, \phi, \lambda)$. In (2) $f_{\beta,\phi}(y|u)$ is a density with a distribution from a one-parameter exponential family, while the second term $f_\lambda(u)$ is the density function of the random effects u with parameter λ.

2.1 H-likelihoods

We call $L(\theta, u; y, u)$ a joint likelihood, a phrase first used by Henderson (1975) in the context of linear mixed models; in our notation these are normal-normal HGLMs, in which the first element refers to the distribution of $y|u$ and the second to that for u. A joint likelihood is not a likelihood in the Fisherian sense because of the presence of unobservables, namely the random effects. Bjørnstad (1996) showed that the joint likelihood satisfies the likelihood principle that the likelihood of the form $L(\theta, u; y, u)$ carries all the (relevant experimental) information in the data about the unobserved quantities u and θ (Edwards 1972, pp.30 and Berger 1985, pp.28). However, the likelihood principle does not provide any obvious suggestions on how to use this likelihood for statistical analysis. It tells us only that some joint likelihood should serve as the basis for such an analysis. For example, the use of $L(\theta, u; y, u)$ in Bayarri *et al.*'s (1988) example results in useless inferences about the random parameter.

A joint likelihood $L(\theta, u; y, k(u))$ is not in general invariant with respect to the choice of parametrization $k(u)$ of the random parameter u, because a change in this parametrization involves a Jacobian term for u. Lee and Nelder (1996) proposed to use the joint density of y and the random effects $v = v(u)$ on the particular scale as shown in (1) to form a subclass of joint likelihoods,

$$L(\theta, v; y, v) \equiv f_{\beta,\phi}(y|v) f_\lambda(v). \tag{3}$$

These were called h-likelihoods by Lee and Nelder (1996), who used them as extended likelihoods for HGLMs. Even though $f_{\beta,\phi}(y|v(u)) \equiv f_{\beta,\phi}(y|u)$ mathematically, we write the conditional density as $f_{\beta,\phi}(y|v(u))$ to stress that the function $v(u)$ defines the scale on which the random effects are assumed to combine additively with the fixed effects β in the linear predictor.

3 Why h-likelihoods among joint likelihoods?

Given that some joint likelihood should serve as the basis for statistical inferences of a general nature, we want find a particular one whose maximization gives meaningful estimators of the random parameters. Maintaining invariance of inferences from the joint likelihood for trivial re-expressions of the underlying model leads to a unique definition of the h-likelihood. For further development we need the following property of joint likelihoods.

Property. The joint likelihoods $L(\theta, u; y, u)$ and $L(\theta, u; y, k(u))$ give identical inferences about the random effects if $k(u)$ is a linear parametrization of u.

This property of joint likelihoods is meaningful because the BP property can be preserved only under linear transformation, i.e. $E\{k(u)|y\} = k(E(u|y))$ only if $k()$ is linear.

Consider a simple normal-normal HGLM of the form: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$ with $N = mn$

$$y_{ij} = \beta + v_i + e_{ij}, \tag{4}$$

where $v_i \backsim i.i.d.\ N(0, \lambda)$ and $e_{ij} \backsim i.i.d.\ N(0, 1)$. Consider a linear transformation $v_i = \sigma v_i^*$ where $\sigma = \lambda^{1/2}$ and $v_i^* \backsim i.i.d.\ N(0, 1)$. The joint loglihoods $l(\theta, v; y, v)$ and $l(\theta, v^*; y, v^*)$ give the same inference for v_i and v_i^*. In (3) the first term $\log f_{\beta,\phi}(y|v)$ is invariant with respect to reparametrizations; in fact $f_{\beta,\phi}(y|v) = f_{\beta,\phi}(y|u)$ functionally for one-to-one parametrization $v = v(u)$. Let $\hat{v}_i$ and $\hat{v}_i^*$ maximize $l(\theta, v; y; , v)$ and $l(\theta, v^*; y, v^*)$, respectively. Then, we have invariant estimates $\hat{v}_i = \sigma \hat{v}_i^*$ because

$$-2 \log f_\lambda(v) = m \log(2\pi\sigma^2) + \sum v_i^2/\sigma^2 = -2 \log f_\lambda(v^*) + m \log(\sigma^2),$$

these loglihoods differ only by a constant.

Consider now model (4), but with a parametrization

$$y_{ij} = \beta + \log u_i + e_{ij}, \tag{5}$$

where $\log(u_i) \backsim i.i.d.\ N(0, \lambda)$. Let $\log(u_i) = \sigma \log u_i^*$ and $\log(u_i^*) \backsim i.i.d.\ N(0, 1)$. Here we have

$$\begin{aligned} -2 \log f_\lambda(u) &= m \log(2\pi\lambda) + \sum (\log u_i)^2/\lambda + 2 \sum \log u_i \\ &= -2 \log f_\lambda(u^*) + m \log(\lambda) + 2 \sum \log(u_i/u_i^*). \end{aligned}$$

Let $\hat{u}_i$ and $\hat{u}_i^*$ maximize $l(\theta, u; y; , u)$ and $l(\theta, u^*; y, u^*)$, respectively. Then, $\log \hat{u}_i \neq \sigma \log \hat{u}_i^*$ because $\log u_i = \sigma \log u_i^*$, i.e. $u_i = u_i^{*\sigma}$, is no longer a linear transformation.

Clearly the two models (4) and (5) are equivalent, so that if h-likelihood is to be a

useful notion we need their corresponding h-loglihoods be equivalent as well. In fact the h-likelihood for model (5) is

$$L(\theta, v; y, v) \equiv f_{\beta,\phi}(y|\log u) f_{\lambda}(\log u)$$

in accordance with the rule that the random effect appears linearly in the linear predictor on the scale $v = \log u$, giving

$$\eta_{ij} = \mu_{ij} = \beta + v_i \quad \text{with} \quad \mu_{ij} = E(y_{ij}|v_i).$$

To maintain invariance of inference with respect to equivalent modellings, we must define the h-likelihood on the particular scale $v(u)$ on which the random effects combine additively with the fixed effects β in the linear predictor.

For simplicity of argument, let $\lambda = 1$, so that there is no dispersion parameter, but only a location parameter β. The h-loglihood $l(\theta, v; y, v)$ is given by

$$-2h = -2l(\theta, v; y, v) \equiv \{N \log(2\pi) + \sum_{ij}(y_{ij} - \beta - v_i)^2\} + \{m \log(2\pi) + \sum_{i} v_i^2\}.$$

This has its maximum at the BP

$$\hat{v}_i = E(v_i|y) = \frac{n}{n+1}(\bar{y}_{i.} - \beta).$$

Suppose that we estimate β and v by joint maximization of h. The solution is

$$\hat{\beta} = \bar{y}_{..} = \sum_{ij} y_{ij}/N \quad \text{and} \quad \hat{v}_i = \frac{n}{n+1}(\bar{y}_{i.} - \bar{y}_{..}) = \sum_{j}(y_{ij} - \bar{y}_{..})/(n+1).$$

Now $\hat{\beta}$ is the ML estimator and $\hat{v}_i$ is the empirical BP defined by

$$\hat{v}_i = \widehat{E(v_i|y)},$$

and can be also justified as the best linear unbiased predictor (BLUP; Searle *et al.* 1992, pp. 269).

The joint loglihood $L(\beta, u; y, u)$ gives

$$-2L(\beta, u; y, u) \equiv \{N \log(2\pi) + \sum(y_{ij} - \beta - \log u_i)^2\} + \{m \log(2\pi) + \sum(\log u_i)^2 + 2\sum(\log u_i)\} \tag{6}$$

with an estimate

$$\hat{v}_i = \log \hat{u}_i = \frac{n}{n+1}(\bar{y}_{i.} - \beta) - 1/(n+1).$$

The joint maximization of $L(\beta, u; y, u)$ leads to

$$\hat{\beta} = \bar{y}_{..} + 1 \text{ and } \hat{v}_i = \frac{n}{n+1}(\bar{y}_{i.} - \bar{y}_{..}) - 1.$$

Thus, in this example joint maximization of the h-loglihood provides satisfactory estimates of both the location and random parameters for either parameterization, while that of a joint loglihood may not.

4 Is joint maximization valid for estimation of dispersion parameters?

Lee and Nelder (1996, 2001*a*) distinguished two types of parameters, fixed effects (location parameters) β and dispersion parameters (ϕ, λ). The use of restricted maximum likelihood (REML) shows that different functions must be maximized to estimate location and dispersion parameters. Our generalization of REML shows that an appropriate adjusted profile h-likelihood (APHL) should be used for estimation of dispersion parameters (Lee and Nelder, 1996, 2001).

Consider the following two equivalent non-normal models: for $i = 1, \ldots, m$

$$y_i|u_i \backsim Poisson(\delta u_i) \quad \text{and} \quad u_i \backsim \exp(1), \tag{7}$$

and

$$y_i|w_i \backsim Poisson(w_i) \quad \text{and} \quad w_i \backsim \exp(1/\delta), \tag{8}$$

where $w_i = \delta u_i$; so we have $E(u_i) = 1$ and $E(w_i) = \delta$.

Note that while fixed effects β appear only in $f_{\beta,\phi}(y|v)$, dispersion parameters (ϕ, λ) can appear in both $f_{\beta,\phi}(y|v)$ and $f_\lambda(v)$. In model (7), use of the log link, on which the fixed and random effects are additive, leads to

$$\log \mu_i = \beta + v_i,$$

where $\mu_i = E(y_i|u_i) = E(y_i|w_i)$, $\beta = \log \delta$, and $v_i = \log u_i$. Now β is a fixed effect, so that $\delta = \exp(\beta)$ is a location parameter in the HGLM context.

In model (8) there is only one random component and no fixed effect, so that the choice of link function, and therefore of $v(u)$, is arbitrary. With an identity link, δ is no longer a fixed effect but becomes the dispersion parameter λ appearing in $f_\lambda(w)$, for which the maximized h-likelihood maintains invariance only with respect to translations. Lee and Nelder (1996, 2001*a*) proposed a different estimation scheme for such parameters as we shall see in the next Section.

We now show that the joint maximization of the h-loglihoods cannot be used for estimation of the dispersion parameters. Suppose that we have an identity link in model

(8). Then h-loglihoods are $L(\theta, u; y, u)$ and $L(\theta, w; y, w)$ for the linear transformation $w = \delta u$. Then,

$$\log f(u) = -\sum u_i = -\sum w_i/\delta = \log f_\delta(w) + m \log \delta$$

so that given δ, random-effect predictions are given by $\hat{u}_i = \hat{w}_i/\delta = y_i/(\delta + 1)$. However, for δ the maximization of $L(\theta, u; y, u)$ yields an estimating equation $\sum y_i = \delta \sum \hat{u}_i = \delta \sum y_i/(\delta + 1)$ with a solution $\hat{\delta} = \infty$ and the use of $L(\theta, w; y, w)$ yields an estimating equation $\delta = \sum \hat{w}_i/m = \delta \sum y_i/\{m(\delta + 1)\}$ with a solution $\hat{\delta} = \bar{y}_\cdot - 1$ where $\bar{y}_\cdot = \sum y_i/m$. Thus, different estimates for the dispersion parameter δ are obtained by jointly maximizing the h-loglihoods $L(\theta, u; y, u)$ and $L(\theta, w; y, w)$ from the same model (8) but with a different parametrization $w = \delta u$.

In model (7), use of the log link leads to the h-loglihoods $L(\theta, u; y, \log u)$ and $L(\theta, w; y, \log w)$ for linear transformation $z = \beta + v$ where $v = \log u$ and $z = \log w$. Here, we have

$$\log f(v) = \sum(-u_i + v_i) = \sum(-w_i/\delta + z_i - \log \delta) = \log f_\delta(z). \tag{9}$$

The joint maximization of this h-loglihood $L(\theta, u; y, \log u)$ with respect to δ and v_i gives

$$\hat{u}_i = (y_i + 1)/(\hat{\delta} + 1) = \widehat{E(u_i|y_i)}$$

because $E(u_i|y_i) = (y_i + 1)/(\delta + 1)$, and the marginal ML estimator $\hat{\delta} = \bar{y}$. Similarly, joint maximization of $L(\theta, w; y, \log w)$ with respect to δ and z_i gives

$$\hat{w}_i = \hat{\delta}(y_i + 1)/(\hat{\delta} + 1) = \widehat{E(w_i|y_i)}$$

because $E(w_i|y_i) = \delta(y_i + 1)/(\delta + 1)$, so also $\hat{z}_i = \hat{v}_i + \log \hat{\delta}$, and again the marginal ML estimator is given by $\hat{\delta} = \bar{y}_\cdot$ Thus, identical estimates for the location parameter δ are obtained by jointly maximizing the h-loglihoods $L(\theta, u; y, \log u)$ and $L(\theta, w; y, \log w)$.

In multiplicative models such as (7) because

$$E(y_i|u_i) = \delta u_i = (c\delta)(u_i/c) \text{ for any } c > 0$$

we may put constraints on either the random effects or the fixed effects. Lee and Nelder (1996) proposed to put constraints on the random effects; for example in model (7) we put $E(u_i) = 1$ which is convenient when there is more than one random component. This strategy converts model (8) to an equivalent model (7), where the log link is an obvious choice in forming the h-loglihood, giving invariant inference for the fixed effect δ. Thus, putting constraints on random effects enlarges the set of fixed effects $\beta = \log \delta$ which can be estimated by a direct maximization of h. We recommend following this strategy in defining h-loglihoods, though it is not compulsory.

In the multiplicative model above neither u_i nor β is separately identifiable because they depend upon an arbitrary constraint. In an additive model such as (4) β is identifiable as $E(y_{ij})$ because of constraints $E(v_i) = 0$ and $E(e_{ij}) = 0$. The model (4) assumes $E(v_i) = 0$ and the model (7) assumes $E(u_i) = 1$, so that care is necessary in comparing parameter estimates from different models. Lee and Nelder (2004) showed that differences in the behaviour of parameters between random-effect models and GEE models are caused by assuming different constraints and therefore are based on a failure to compare like with like.

5 Extended likelihood framework

The original Fisher likelihood framework had two types of object and two kinds of inference:
Data Generation: Generate an instance of the data y from a probability function with given fixed parameters θ

$$f_\theta(y).$$

Parameter Estimation: Given the data y, make an inference about an unknown fixed θ in the stochastic model by using the likelihood

$$L(\theta; y).$$

The connection between these two processes is given by

$$L(\theta; y) \equiv f_\theta(y),$$

where L and f are algebraically identical, but on the left-hand side y is fixed while θ varies and on the right-hand side θ is fixed while y varies.

The extended likelihood framework for three types of object can be described as follows:
Data Generation: (i) Generate an instance of the random effects v from a probability function $f_\theta(v)$ and then with v fixed, (ii) generate an instance of the data y from a probability function $f_\theta(y|v)$. The combined stochastic model is given by the product of the two probability functions

$$f_\theta(v)\, f_\theta(y|v). \tag{10}$$

Parameter Estimation: Given the data y, we can (i) make inferences about θ by using the marginal likelihood $L(\theta; y) \equiv f_\theta(y)$, and (ii) given θ, make inferences about v by using

the conditional likelihood in the form

$$L(\theta, v; v|y) \equiv f_\theta(v|y).$$

Given the data y, the extended likelihood for the joint unknowns (v, θ) is given by

$$L(\theta, v; y, v) = L(\theta; y)L(\theta, v; v|y) \equiv f_\theta(y)\, f_\theta(v|y). \tag{11}$$

The connection between these two processes is given by

$$f_\theta(y)\, f_\theta(v|y) \equiv L(\theta, v; y, v) \equiv f_\theta(v, y) = f_\theta(v)\, f_\theta(y|v). \tag{12}$$

On the left-hand side y is fixed while (v, θ) vary, while on the right-hand side θ is fixed while (v, y) vary. In the extended likelihood framework the v appear in data generation as random instances, but in parameter estimation as unknowns.

The combined stochastic model $f_\theta(y|v)\, f_\theta(v)$ in (10) for data generation is often easily available in an explicit form. However, the practical difficulties in extended likelihood inference stem from the fact that the two components,

$$f_\theta(y) = \int f_\theta(y|v)\, f_\theta(v)dv \quad \text{and} \quad f_\theta(v|y) = f_\theta(y|v)\, f_\theta(v) / \int f_\theta(y|v)\, f_\theta(v)dv,$$

are generally hard to obtain, except for some conjugate families, because of the integration involved. However, the h-likelihood can exploit the connection (12) to give likelihood inferences of a general nature. In the next Section we show how to implement inferential procedures without explicitly computing the two components $f_\theta(y)$ and $f_\theta(v|y)$.

Let

$$h = m + \log f_\theta(v|y) = \log f_\theta(v) + \log f_\theta(y|v), \tag{13}$$

where m is the marginal loglihood $m = \log L(\theta; y)$. This is the h-loglihood, which plays the same role as the loglihood in Fisher's likelihood inference.

5.1 Inference for random parameters

The relative simplicity of h-likelihood methods of inference becomes apparent when we compare them with other methods. If the conditional density $f_\theta(v|y)$ follows the normal distribution, it is immediate that given θ, the maximum h-likelihood estimator for v is a BP, i.e. $\hat{v} = E(v|y)$. If there exists a transformation $k()$ such that $L(\theta, v; v|y)$ $(\equiv f_\theta(k(v)|y))$ takes the form of a normal distribution, the h-likelihood gives a BP for $\widehat{k(v)} = k(\hat{v}) = E\{k(v)|y\}$. However, the h-likelihood gives more than this.

Consider a mixed linear model

$$Y = X\beta + Zv + e,$$

where $v \backsim MVN(0, \Lambda)$ and $e \backsim MVN(0, \Sigma)$ and MVN stands for a multivariate normal distribution. Henderson (1975) showed that the joint maximization of his joint loglihood (which, for the normal-normal model, is the h-loglihood) leads to the estimating equations

$$\begin{pmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}Z \\ Z^T\Sigma^{-1}X & Z^T\Sigma^{-1}Z + \Lambda^{-1} \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^T\Sigma^{-1}Y \\ Z^T\Sigma^{-1}Y \end{pmatrix}.$$

Set H to be the square matrix of the left hand side, $V = Z\Lambda Z^T + \Sigma$, and $D = Z^T\Sigma^{-1}Z + \Lambda^{-1}$. Then, given Λ and Σ, the solution for β gives the ML estimator, satisfying

$$X^T V^{-1} X\hat{\beta} = X^T V^{-1} Y,$$

and the solution for v gives the empirical BPs

$$\hat{v} = \widehat{E(v|Y)} = E(v|Y)|_{\beta=\hat{\beta}} = \Lambda Z^T V^{-1}(Y - X\hat{\beta}) = D^{-1}Z^T\Sigma^{-1}(Y - X\hat{\beta}).$$

Furthermore, H^{-1} gives estimates of

$$E\left\{\begin{pmatrix} \hat{\beta} - \beta \\ \hat{v} - v \end{pmatrix}\begin{pmatrix} \hat{\beta} - \beta \\ \hat{v} - v \end{pmatrix}^T\right\}.$$

This yields $(X^T V^{-1} X)^{-1}$ as a variance estimate for β, which coincides with that for the ML estimates. Now we see that H^{-1} also gives the correct estimate for $E\left\{(\hat{v}-v)\,(\hat{v}-v)^t\right\}$.

When β is known we use the BP

$$\tilde{v} = E(v|Y).$$

Then we have

$$var(\tilde{v} - v) = E\left\{(\tilde{v} - v)(\tilde{v} - v)^T\right\} = E\left\{var(v|Y)\right\}.$$

Note here that

$$var(v|Y) = \Lambda - \Lambda Z^T V^{-1} Z\Lambda = D^{-1}.$$

When β is known D^{-1} gives a proper estimate of the variance of $\tilde{v} - v$.

The extended likelihood principle of Bjørnstad (1996) is that the joint likelihood of the form $L(\theta, v; y, v)$ carries all the information in the data about the unobserved

quantities v and θ. Because $f_\theta(y)$ in (11) does not involve v, $L(\theta, v; v|y) \equiv f_\theta(v|y)$ seems to carry all the information in the data about the random parameters. This leads to the empirical Bayes (EB) method for inference about v, which uses the estimated posterior

$$f_{\hat{\theta}}(v|y),$$

where $\hat{\theta}$ are usually the marginal ML estimators (Carlin and Louis, 2000). Thus, maximization of the h-likelihood yields EB-mode estimators, and they can be obtained without computing $f_\theta(v|y)$. However, when β is unknown the estimated posterior $f_{\hat{\beta}}(v|y)$ for the EB procedure gives $D^{-1}|_{\beta=\hat{\beta}}$ as a naive estimate for $var(\hat{v} - v)$, and this does not properly account for the uncertainty caused by estimating β. Various complicated remedies have been suggested for the EB interval estimate (Carlin and Louis, 2000).

By contrast, the h-loglihood gives a straightforward correction. Here we have

$$var(\hat{v} - v) = E\{var(v|Y)\} + E\left\{(\hat{v} - \tilde{v})(\hat{v} - \tilde{v})^T\right\},$$

where the second term shows the variance inflation caused by estimating the unknown β. As an estimate for $var(\hat{v} - v)$ the appropriate component of H^{-1} gives

$$\{D^{-1} + D^{-1}Z^T\Sigma^{-1}X(X^TV^{-1}X)^{-1}X^T\Sigma^{-1}ZD^{-1}\}|_{\beta=\hat{\beta}}.$$

Because $\hat{v} - \tilde{v} = -\Lambda Z^T V^{-1} X(\hat{\beta} - \beta)$ we can show that

$$E\left\{(\hat{v} - \tilde{v})(\hat{v} - \tilde{v})^T\right\} = D^{-1}Z^T\Sigma^{-1}X(X^TV^{-1}X)^{-1}X^T\Sigma^{-1}ZD^{-1}.$$

Thus, the h-loglihood handles correctly the variance inflation caused by estimating fixed effects. From this we can construct confidence bounds for unknown v, fixed once the data are observed. Lee and Nelder (1996) extended the results of this section to general HGLMs under some regularity conditions. Later, we illustrate this further by using Bayarri *et al.*'s (1988) example as a non-normal model.

We see that inferences about random effects cannot be made by using solely $f_\theta(v|y)$, as the EB method does. Because $f_\theta(v|y)$ involves the fixed parameters θ we should use the whole h-likelihood to reflect the uncertainty about θ; it is the other component $f_\theta(y)$ which carries the information about this. The notation $L(\theta, v; v|y)$ shows that the EB problem is caused by the nuisance fixed parameters θ.

5.2 Inference for fixed parameters for both location and dispersion

The likelihood principle of Birnbaum (1962) is that the marginal likelihood $L(\theta; y)$ carries all the (relevant experimental) information in the data about the fixed parameters θ, so that $L(\theta; y)$ should be used for inferences about θ: see also Berger and Wolpert

(1984). For inferences about fixed parameters θ we can use $f_\theta(y)$ alone because $L(\theta; y)$ does not involve nuisance random parameters at all. However, in general the marginal likelihood requires intractable integration. One method of obtaining the marginal ML estimators for θ is the expectation-maximization (EM) algorithm of Dempster *et al.* (1977). This exploits the property (13) of joint loglihoods using the result that under appropriate regularity conditions

$$E(\partial h/\partial\theta|y) = \partial m/\partial\theta + E(\partial \log f_\theta(v|y)/\partial\theta|y) = \partial m/\partial\theta.$$

The last equality is immediate from the fact that

$$\int f_\theta(v|y)dv = 1.$$

The EM algorithm is often numerically slow to converge and it is analytically hard to evaluate the conditional expectation $E(h|y)$. Alternatively, simulation methods, such as Monte Carlo EM (Vaida and Meng, 2004) and Gibbs sampling (Karim and Zeger, 1992), can be used to evaluate the conditional expectation, but these methods are computationally intensive. Instead, numerical integration using Gauss-Hermite quadrature (Crouch and Spiegelman, 1990) could be directly applied to obtain the ML estimators, but this also becomes computationally heavier as the number of random components increases.

By contrast, we can obtain estimators for θ by directly maximizing appropriate quantities derived from the h-loglihood, and compute their standard error estimates from the second derivatives. In our framework we do not need to evaluate an analytically difficult expectation step nor use a computationally intensive method, such as Monte Carlo EM or numerical integration. Instead we maximize adjusted profile h-likelihoods (APHLs) to obtain ML and REML estimators. Ha and Lee (2005*a*) showed how h-likelihood gives straightforward estimators of β for mixed linear models with censoring, whereas the ordinary EM method has difficulty.

In our framework there are two useful adjusted profile loglihoods for inferences about fixed parameters. The marginal loglihood m can be obtained from the h-loglihood by integrating out the random parameters,

$$m \equiv \log \int f_\theta(v, y)dv = \log \int f_\theta(y|v)\, f_\theta(v)dv. \tag{14}$$

In mixed linear models the conditional density,

$$f_{\phi,\lambda}(y|\tilde{\beta}) = f_{\beta,\theta}(y)/\, f_{\beta,\theta}(\tilde{\beta}),$$

where $\tilde{\beta}$ are ML estimators given (ϕ, λ), is free of β (Smyth , 2002), so that the restricted loglihood of Patterson and Thompson (1971) can be written as

$$r = \log L(\phi, \lambda; y|\tilde{\beta}) \equiv \log f_{\phi,\lambda}(y|\tilde{\beta}).$$

This has been proposed for inference about the dispersion parameters (ϕ, λ) to reduce bias, especially in finite samples: see also Harville (1977).

In our framework the marginal loglihood m is an adjusted profile loglihood for the fixed parameters θ, after eliminating random parameters v by integration from the h-loglihood h, and the restricted loglihood r is that for the dispersion parameters (ϕ, λ), after eliminating fixed parameters β by conditioning from the marginal loglihood m. However, in general they hard to obtain because they use the marginal loglihood m. Let l be a loglihood, either a marginal loglihood m or an h-loglihood h, with nuisance parameter α, random or fixed. Lee and Nelder (2001a) considered a set of functions $p_\alpha(l)$, defined by

$$p_\alpha(l) = [l - \frac{1}{2}\log\det\{D(l,\alpha)/(2\pi)\}]|_{\alpha=\tilde{\alpha}} \tag{15}$$

where $D(l,\alpha) = -\partial^2 l/\partial\alpha^2$ and $\tilde{\alpha}$ solves $\partial l/\partial\alpha = 0$. For fixed effects β the use of $p_\beta(m)$ is equivalent to conditioning on $\tilde{\beta}$, i.e. $p_\beta(m) \simeq r = l(\phi, \lambda; y|\tilde{\beta}) \equiv \log f_{\phi,\lambda}(y|\tilde{\beta})$ to the first order (Cox and Reid, 1987), while for random effects v the use of $p_v(h)$ is equivalent to integrating them out using the first-order Laplace approximation, i.e. $p_v(h) \simeq m$ (Lee and Nelder 2001a). The set of functions $p_*()$ may be regarded as derived loglihoods for various subsets of parameters.

In mixed linear models

$$m \equiv p_v(h) \text{ and } p_\beta(m) \equiv p_{\beta,v}(h).$$

Thus, here the marginal ML estimators for β and their standard error estimates can be obtained by directly maximizing the adjusted profile h-loglihood $p_v(h)$, instead of the joint maximization of the previous Section. The restricted loglihood r has been used only in mixed linear models. Its natural extension is $p_\beta(m)$ (Cox and Reid, 1987). To avoid intractable integration, instead of $p_\beta(m)$ we may use $p_{\beta,v}(h)$ as the restricted likelihood for dispersion parameters. The use of $p_{\beta,v}(h)$ for estimating the dispersion parameters (ϕ, λ) means that we can eliminate both random and fixed effects simultaneously from the h-likelihood. Lee and Nelder (2001a) showed that in general $p_{\beta,v}(h)$ is approximately $p_\beta(p_v(h))$ and that numerically $p_{\beta,v}(h)$ provides good dispersion estimators for HGLMs. This reduces bias of the ML estimator greatly in frailty models with nonparametric baseline hazards where the number of nuisance β increases with sample (Ha and Lee, 2005b) .

In principle we should use the h-loglihood h for inferences about v, the marginal-loglihood m for β and the restricted loglihood $p_\beta(m)$ for the dispersion parameters.

When m is numerically hard to obtain, we propose to use APHLs $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to m and $p_\beta(m)$; $p_{\beta,v}(h)$ gives approximate restricted ML estimators for the dispersion parameters and $p_v(h)$ approximate ML estimators for the location parameters. Higher-order approximations can be useful for improved accuracy (Lee and Nelder 2001*a*). Although in general a joint maximization of the h-loglihood does not provide marginal ML estimators for β the deviance differences constructed from h and $p_v(h)$ are often very similar, so that we propose to use h for estimating β unless it yields non-ignorable biases. For example, in HGLMs for binary data $p_v(h)$ should be used for estimating β (Noh and Lee 2004).

Using the formula

$$p_v(h) = [h - \frac{1}{2}\log\det\{D(h,v)/(2\pi)\}]|_{v=\tilde{v}},$$

model (4) gives $D(h,v) = diag(d_i)$ where $d_i = n + 1$, and model (7) gives $d_i = y_i + 1$, i.e. for both models $D(h,v)$ is independent of the fixed effects β, depending only upon dispersion parameters if these exist, so that the maximization of h also provides the ML estimators for β. This is true for three models, the normal-normal, Poisson-gamma and gamma-inverse gamma with log link; for these explicit forms for m are available (Lee and Nelder 1996).

Breslow and Clayton (1993) proposed the use of the REML estimating equations for normal mixed linear models to estimate dispersion parameters in GLMMs. In mixed linear models the adjustment term $D(h,\delta)$ with $\delta = (v,\beta)$ does not involve δ. However, this is not so in general, so that Breslow and Clayton's (1993) method suffers from severe bias because it ignores derivative terms $\partial\tilde{\delta}/\partial\lambda$ and $\partial\tilde{\delta}/\partial\phi$. Furthermore, Lin and Breslow's (1996) correction of the Breslow and Clayton's method still suffers from the non-ignorable bias caused by ignoring these important terms, while the h-likelihood procedure does not (Noh and Lee, 2004).

We now illustrate the h-likelihood approach for random-effect models using Bayarri *et al.*'s (1988) example.

6 Bayarri *et al.*'s example revisited

Let us return to Bayarri et al's (1988) example:

$$y|u \backsim \exp(u) \quad \text{and} \quad u \backsim \exp(\theta), \tag{16}$$

and equivalently

$$y|w \backsim \exp(w/\theta) \quad \text{and} \quad w \backsim \exp(1) \tag{17}$$

where $E(w) = 1$ and $E(u) = 1/\theta$. Here we have the marginal loglihood

$$m = \log L(\theta; y) = \log\theta - 2\log(\theta + y).$$

This gives the marginal ML estimator $\hat{\theta} = y$ and its variance estimator

$$\widehat{var}(\hat{\theta}) = -\{\partial^2 m/\partial\theta^2|_{\theta=\hat{\theta}}\}^{-1} = 2y^2.$$

Following the strategy of putting the constraints on random effects $E(w) = 1$ let us consider the model (17) first. Here because

$$\mu = E(y|w) = \theta/w,$$

the log link achieves additivity

$$\eta = \log\mu = \beta + v,$$

where $\beta = \log\theta$ and $v = -\log w$. This leads to the h-loglihood

$$h = l(\theta, v; y, v) \equiv \log f_\theta(y|v) + \log f(v) = -v - \log\theta - wy/\theta - w - v.$$

Suppose that θ and therefore β is known. The maximization $\partial h/\partial v = -2 + (y/\theta + 1)w = 0$ gives the BP

$$\hat{w} = 2\theta/(y + \theta) = E(w|y).$$

Here the BP is on the w scale. Then, the corresponding Hessian $-\partial^2 h/\partial w^2|_{w=\hat{w}} = 2/\hat{w}^2 = (y + \theta)^2/(2\theta^2)$ gives as an estimate for $var(\hat{w} - w)$

$$var(w|y) = 2\theta^2/(y + \theta)^2.$$

Now suppose that θ and therefore β is unknown. The joint maximization

$$\partial h/\partial v = -2 + (y/\theta + 1)w = 0 \text{ and } \partial h/\partial\theta = -1/\theta + yw/\theta^2 = 0$$

gives the ML estimator $\hat{\theta} = y$ and the empirical BP $\hat{w} = 2\hat{\theta}/(y + \hat{\theta}) = \widehat{E(w|y)} = \theta\widehat{E(u|y)} = 1$. Because there is only one random effect in the model the constraint $E(w) = 1$ makes $\hat{w} = 1 = E(w)$: for more discussion see Lee and Nelder (1996, 2004). Because

$$-\partial^2 h/\partial w^2|_{\theta=\hat{\theta},w=\hat{w}} = 2, \quad -\partial^2 h/\partial\theta^2|_{\theta=\hat{\theta},w=\hat{w}} = 1/y^2, \quad -\partial^2 h/\partial\theta\partial w|_{\theta=\hat{\theta},w=\hat{w}} = -1/y$$

we have an estimator

$$\widehat{var}(\hat{\theta}) = 2y^2,$$

which is the same as that from the marginal loglihood. Now we have

$$\widehat{var}(\hat{w} - w) = 1 = var(w),$$

which reflects the variance increase caused by estimating θ; note that

$$\widehat{var}(w|y) = 2\theta^2/(y+\theta)^2|_{\theta=\hat{\theta}} = 1/2.$$

Here $-\partial^2 h/\partial v^2|_{w=\hat{w}} = 2$, so that h and $p_v(h)$ are proportional, showing that the joint maximization is a convenient tool to compute an exact ML estimator and its standard error estimates.

Suppose that we use the model (16) with an identity link. Now θ is a dispersion parameter appearing in $f_\theta(u)$ and the h-loglihood is given by

$$h = L(\theta, u; y, u) \equiv \log f(y|u) + \log f_\theta(u) = \log u + \log\theta - u(\theta + y).$$

Then, the equation $\partial h/\partial u = 1/u - (\theta + y) = 0$ gives $\tilde{u} = 1/(\theta + y)$. From this we get

$$p_u(h) \equiv \log\tilde{u} + \log\theta - \tilde{u}(\theta+y) - \frac{1}{2}\log\{1/(2\pi\tilde{u}^2)\} = \log\theta - 2\log(\theta+y) - 1 + \frac{1}{2}\log 2\pi,$$

which is proportional to the marginal loglihood m, and so yields the same inference for θ. Here $-\partial^2 h/\partial u^2|_{u=\tilde{u}} = 1/\tilde{u}^2 = (\theta+y)^2$ and thus h and $p_v(h)$ are no longer proportional, so that the joint maximization cannot give an exact ML estimator for dispersion parameters.

Schall's (1991) method is the same as Breslow and Clayton's (1993) PQL method for GLMMs. They are the same as the h-likelihood method, but ignore $\partial\tilde{u}/\partial\theta$ in the dispersion estimation (Lee and Nelder 2001*a*). Now suppose that the $\partial\tilde{u}/\partial\theta$ term is ignored in maximizing $p_u(h)$. Then we have the estimating equation

$$1 = \theta\tilde{u} = \theta/(\theta+y), \quad \text{for } y > 0$$

which gives an estimator $\hat{\theta} = \infty$. Thus, the term $\partial\tilde{u}/\partial\theta$ should not be ignored; if it is, it can result in a severe bias in estimation and a distortion of the standard error estimate; here, for example,

$$\widehat{var}(\hat{\theta}) = \hat{\theta}^2 = \infty.$$

Similarly, in models (5) and (7) the joint maximization of $l(y, u|\lambda)$ does not provide a

valid inference for β as we saw, but $p_u(l(y,u|\lambda))$ does provide the ML estimator of β for both models. Thus, for model (8) even though we chose $l(y,w|\lambda)$ as the h-loglihood our inferential procedure provides equivalent inference to that from a model with a log-link. Thus, with proper use of h-likelihood it is possible to have meaningful inferences about both the random and fixed parameters.

For some non-linear random-effect models, such as those occurring in pharmacokinetics, the definition of h-likelihood is less clear, but we may still use $p_u(l(y,u))$ for inference about non-random parameters. Lee and Nelder (1996) noted that $p_u(l(y,u))$ is invariant, i.e. gives invariant inference, with respect to an arbitrary linear transformation of u. Even though the simplicity of the h-likelihood algorithm is lost, $p_u(l(y,u))$ provides a good inferential criterion because the Laplace approximation is often very accurate.

6.1 Discussion

There have been many alleged counterexamples similar to that of Bayarri *et al.* (1988), purporting to show that an extension of the Fisher likelihood to three objects is not possible. An important criticism is that we may get qualitatively different (i.e. non-invariant) inferences for trivial re-expressions of the underlying model. We see that defining the h-likelihood on the right scale avoids such difficulties. These complaints are caused by a misunderstanding of the h-likelihood framework and the wrong use of joint maximization to obtain all the parameter estimates. Another criticism has been about the statistical efficiency of the h-likelihood procedures (for example, Little and Rubin, 2002). An appropriate adjusted profile h-likelihood (APHL) should be used for estimation of dispersion parameters (Lee and Nelder, 1996, 2001). The h-likelihood is a natural way of making inferences about unobservables v. The definition of h in the proper scale of v and the use of APHLs give valid inferences. All the alleged counterexamples for missing data problems in Little and Rubin (2002; chapter 6.3) can be refuted similarly as in Section 6: for detailed discussion see Yun *et al.* (2005).

Hinkley (1979) and Butler (1986) introduced predictive likelihood for inferences of unobserved future observations. In missing data problemes and random-effect models, the APHL $p_v(h)$, eliminating random parameters v, is used for inferences about fixed θ. However, in prediction problems for an unobserved future observation v, various predictive likelihoods, eliminating fixed parameters θ, have been proposed for inferences about random v (Bjørnstad, 1990). Davison (1986) proposed to use the APHL $p_\theta(h)$, derived as an approximate predicted likelihood under a non-informative prior, which Butler (1990) called the modified profile predictive loglihood. Following Barndorff-Nielsen (1983) Bjørnstad (1990) suggested an approximation

$$\log f_\theta(z|y) = p_\theta(h) + \log\det(\partial\hat{\theta}/\partial\hat{\theta}_v),$$

where $\hat{\theta}$ is the ML estimator based upon y and $\hat{\theta}_v$ is the maximum h-logihood estimator.

If the number of predictands remains fixed as the number of y grows we have $\log\{\det(\partial\hat{\theta}/\partial\hat{\theta}_v)\} = O_p(1/n)$, so that we can also derive the predictive likelihood $p_\theta(h)$ as the approximate conditional likelihood, following Cox and Reid (1987). So, with h-likelihood perspectives, the predictive likelihood is an attempt to derive the APHL for predictions, eliminating all fixed parameters.

7 Advantages of extended likelihood framework

The difficulty in obtaining the two components $f_\theta(y)$ and $f_\theta(v|y)$ has limited the class of stochastic models that allow likelihood inference. Except for the limited conjugate family there are few models which allow explicit forms for these two components. The use of h-likelihood makes such limitations unnecessary, so that likelihood inference can be drawn from a much wider class of models. Furthermore, the extended likelihood framework preserves the advantages of the original framework.

7.1 Generality of application

HGLMs have become increasingly popular since the initial synthesis of GLMs, random-effect models, and structured-dispersion models was found to be extendable to include models for temporal and spatial correlations (Lee and Nelder 2001*a*, 2001*b*). Heterogeneity of means between clusters (the so-called between-cluster variation) can be modelled by introducing random effects in the mean. In HGLMs both fixed and random effects are allowed for the mean but only fixed effects for the dispersion. We have introduced double HGLMs (Lee and Nelder 2005), which allow both fixed and random effects not only for the mean but also for the dispersion. This means that heterogeneity of dispersion between clusters can be similarly modelled by introducing random effects in the dispersion. We now have a systematic way of generating heavy-tailed distributions for various types of data such as counts and proportions. This class will, among other things, enable models of types widely used in the analysis of financial data to be explored, and should give rise to new extended classes of models. The h-likelihood plays a key role in the synthesis of the inferential tools needed for these models.

7.2 Statistical efficiency of h-likelihood method

HGLMs have received increasing attention due to their wide applicability and ease of interpretation. However, the computation of the ML estimation of the parameters is a complex task. The marginal loglihood m, obtained by integrating out the random

effects, is in general analytically intractable. The computational problems are magnified when the random effects have a crossed design, where the data cannot be reduced to small independent clusters. For example, in the Salamander data marginal likelihood inference, based upon numerical integration using Gauss-Hermite quadrature is not feasible since a 120-dimensional integral is required. Thus, various approximate methods have been proposed by Schall (1991), Breslow and Clayton (1993), Drum and McCullagh (1993), Shun and McCullagh (1995), Lee and Nelder (1996, 2001), Lin and Breslow (1996) and Shun (1997). For binary data Noh and Lee (2004) showed numerically that the h-likelihood estimator has less bias than the other methods including MCMC-type methods: see also the simulation studies of Poisson and binomial models (Lee and Nelder 2001*a*), of frailty models (Ha *et al.*, 2001) and of mixed linear models with censoring (Ha *et al.*, 2002). We have not seen any method which outperforms the h-likelihood procedure, though we do not say that the current h-likelihood procedure is incapable of improvement.

7.3 Computational efficiency of h-likelihood method

The h-likelihood (13) gives a new definition of conjugate families (Lee and Nelder 2001*a*), showing that the likelihood for conjugate family for $\log f_\theta(v)$ takes the form of a GLM. It is sum of component likelihoods, $\log f_\theta(v)$ and $\log f_\theta(y|v)$, both representable as GLM likelihoods. This means that an extended class of models can be decomposed into component GLMs (Lee and Nelder 2001a, 2005) and these extended models can be fitted as an interconnected set of component GLMs. This greatly facilitates the development of model-checking techniques for the whole class (Lee and Nelder 2001*a*). A single algorithm, iterative weighted least squares, can be used throughout all these extended classes of models and requires neither prior distributions of parameters nor multi-dimensional quadrature. The h-likelihood plays a key role in the synthesis of the computational algorithms needed for this extended class of models.

This formulation means that a great variety of models can be fitted by a single algorithm and compared using extensions of standard GLM procedures. Thus we can change the link function, allow various types of term in the linear predictor and use model-selection methods for adding or deleting terms. Furthermore various model assumptions can be checked by applying GLM model-checking procedures to the component GLM. This establishes, we believe, algorithmic *wiseness* in the sense of Efron (2003).

8 Conclusion

In general the computation of the ML and/or REML estimation of the parameters is a complex task due to the intractable integration to obtain the marginal loglihood. With the use of h-likelihood we can obtain ML and REML estimators by maximizing APHLs. However, still many believe that the marginal loglihood, without involving random effects, is the default loglihood. However, its use has always left a problem of inference about unobservable random variables (subject-specific inferences, Zeger *et al.*, 1988) and has restricted stochastic models to those having an explicit marginal likelihood. Thus, Bayesian methods have been extensively used for models without an explicit marginal likelihood, while likelihood inference is relatively less well developed, because the definition of likelihood for such inferences is not agreed.

We do not object to the use of marginal likelihood for inferences about fixed parameters, which in our approach appears as an APHL similar to the restricted likelihood. The restricted likelihood cannot allow inferences about fixed effects because they are eliminated. Similarly, the marginal likelihood cannot allow inferences about individuals, so that some other method must be used for this. As we use the marginal likelihood for inferences about β in the REML procedure it would be natural to use the h-likelihood for inferences about random effects. Thus, it is the h-likelihood that is fundamental, giving both marginal inference for fixed parameters and subject-specific inference for random or combined fixed and random parameters.

It is perhaps unfortunate that Bayesians, from Lindley and Smith (1972) onwards, seem to have made a take-over bid for all hierarchical models, implying that one has to be a Bayesian to deal with them. The availability of Markov-chain Monte Carlo, making models without an explicit marginal likelihood seem more easily handled via Bayesian computations, has appeared to justify this. By using h-likelihood, we may deal with models with random effects directly in a likelihood framework because there is an explicit analytic form of the likelihood. Furthermore inferences about random effects are possible without resorting to an empirical Bayesian framework. There seems to be no evidence that MCMC-type methods give better estimators than the h-likelihood method at least with binary data (Noh and Lee, 2004).

H-likelihood, as an extended likelihood, gives a powerful and practical framework for statistical inference; being a natural extension of Fisher likelihood to models with random parameters, it will become, we believe, widely used for inference for unobserved random variables. Nevertheless, it remains to be seen if any further generalizations can be made.

Acknowledgments We thank Sir David Cox, G. Casella, M. Crowder, M. Healy, J. Lawless, J. Lee, Y. Pawitan and S. Senn for their constructive comments, and P. McCullagh and C. McCulloch for providing stimulating examples and discussions. This

research was supported by a grant from Korea Research Foundation Grant (KRF-2003-070-C00008).

References

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343-365.

Barndorff-Nielsen, O. E. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli*, 2, 319-340.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1988). What is the likelihood function? (with discussion). *Statistical Decision Theory and Related Topics IV. Vol. 1*, eds S.S. Gupta and J. O. Berger, New York: Springer.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Berger, J. O., and Wolpert, R. (1984). *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics Monograph Series.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Am. Statist. Ass.*, 57, 269-306.

Bjørnstad, J. F. (1990). Predictive likelihood principle: a review (with discussion). *Statist. Sci*., 5, 242-265.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and likelihood principle. *J. Am. Statist. Ass.*, 91, 791-806.

Box, M. J., Draper, N. R. and Hunter, W. G. (1970). Missing values in multi-response nonlinear data fitting. *Technometrics*, 12, 613-620.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, 88, 9-25.

Butler, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. R. Statist. Soc.* B, 48, 1-38.

Butler, R. W. (1990). Comment on "Predictive likelihood inference with applications" by Bjørnstad. *Statist. Sci*., 5, 255-259.

Carlin, B. P. and Louis, T. A. (2000). *Bayesian and Empirical Bayesian Methods for Data Analysis*. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, 32, 1-18.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models. *J. Am. Statist. Ass.*, 85, 464-469.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, 73, 323-332.

Dempster, A. P. N., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B, 39, 1-38.

Drum, M. L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, 49, 677-689.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Efron, B. (2003). A conversation with good friends. *Statist. Sci.*, 18, 268-281.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3-32.

Ha, I. D. and Lee, Y. (2005*a*). Multilevel mixed linear models for survival data. *Lifetime Data Analysis*, 11, 131-142.

— (2005*b*). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models, to appear in *Biometrika*, 42, 717-723.

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, 88, 233-243.

— (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, 8, 163-176.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection models. *Biometrics*, 31, 423-447.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.*, 7, 718-728. Corr 8, 694.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, 48, 681-694.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, 58, 619-678.

— (2001*a*). Hierarchical generalized linear models: A synthesis of generalised linear models, random-effect model and structured dispersion. *Biometrika*, 88, 987-1006.

— (2001*b*). Modelling and analysing correlated non-normal data, *Statistical Modelling*, 1, 3-16.

— (2004). Conditional and marginal models: another view (with discussion). *Statist. Sci.*, 19, 219-238.

— (2005). Double hierarchical generalized linear models (with discussion). to appear at *Appl. Statist.*

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Statist. Ass.*, 91, 1007-1016.

Lindley, D. V. and Smith, A. F. M. (1972). Bayesian estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, 34, 1-41.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley.

Mathiasen, P. E. (1979). Prediction functions. *Scand. J. Statist*., 6, 1-21.

Noh, M. and Lee, Y. (2004). REML estimation for binary data in GLMMs. Manuscript prepared for publication.

Patterson, H. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Clarendon Press.

Pearson, K. (1920). The fundamental problems of practical ststistics. *Biometrika*, 13, 1-16.

Savage, L. J. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist.*, 4, 441-500.

Schall, R. (1991). Estimation in generalised linear models with random effects. *Biometrika*, 78, 719-727.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley and Sons.

Shun, Z. (1997). Another look at the salamander mating data: a modified Laplace approximation approach. *J. Am. Statist. Ass.*, 92, 341-349.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high-dimensional integrals. *J. R. Statist. Soc.* B, 57, 749-760.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *J.Comp. Graph. Statist.*, 11, 1-12.

Vaida, F. and Meng, X. L. (2004). Mixed linears models and the EM algorithm in *Applied Bayesian and Causal Inference from an Incomplete Data Perspective*. Gelman, A. and Meng, X. L. (editors): John Wiley and Sons.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.*, 1, 129-142.

Yun, S., Lee, Y. and Kenward, M. G. (2005). Using h-likelihood for missing observations. manuscript prepared for publication.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Discussion of “likelihood for random-effect models” by Lee and Nelder

James Henry Roger

Research Statistics Unit, GlaxoSmithKline, United Kingdom.
Email: james.h.roger@gsk.com

The authors are congratulated in summarizing material from their extensive work over the last few years into a paper which is readable by any well qualified postgraduate statistician. The first example from Bayarri *et al.* (1988) is a very useful way to demonstrate the definition of joint likelihood and its relationship to the marginal likelihood L(y| θ)and the conditional likelihood L(y|u,θ). I will describe my points in terms of this example.

If the joint likelihood was written as $L(\theta, u) = u\theta \exp\{-u(\theta + y)\}\, dy\, du$ then there would be no ambiguity about how the joint likelihood changes when a transformation is made to the u space since on the new scale it simply becomes $L(\theta, \log u) = u^2\theta \exp\{-u(\theta + y)\}\, dy\, d\log u$. The problem occurs when one decides to choose a best combination of θ and u by maximizing this function over both θ and u ignoring the du or $d\log u$. For different parametrizations one is ignoring a different multiplier. So one derives different estimators depending upon the choice of the parameterization for the random effect u. For joint estimation, one might at first sight think of optimizing in the θ direction by maximization while using some other method in the u direction. The median is an obvious choice for unidimensional u, but does not generalize easily, and is awkward to solve. The h-likelihood approach selects a specific natural choice for the parameterization of u and then uses maximization in both the θ and u directions. In contrast, when solving for the fixed effect θ, one simply integrates over the probability part of the joint likelihood to obtain the marginal likelihood, which is clearly a valid procedure. The point of including du in the definition of the joint likelihood $L(\theta, u)$ is that in the u direction it has to be interpreted as a probability density rather than a likelihood. So if we maximize in this direction then this has to interpreted in terms of the scale du.

This is a parallel problem to that of handling multiple nuisance parameters in a model with only fixed effects. The profile likelihood is easy to use but is not a proper likelihood. Conditional or marginal likelihoods can sometimes be found which remove the nuisance parameters, while being proper likelihoods. The modified profile likelihood (Barndorff-Nielsen, 1983) is an approximation to such a conditional or marginal likelihood. It includes a Jacobean term that maintains invariance under

transformation of the parameter space. Cox and Reid (1987) indicate that this Jacobean term is often intractable and limits the application of the method. One approach is to choose an appropriate scale and proceed without the Jacobean term hoping that it is almost constant. However, Pawitan (Example 10.8, 2001) points out that for apparently automatic parameterization, dropping the Jacobean can lead to the wrong form of analysis. Another way to adjust the profile likelihood, when a suitable conditional or marginal likelihood is not readily available, is to split the log-likelihood into the sum of independent components and then profile each component using an estimate of the nuisance parameters based on the other components ignoring the component itself, in a jackknife-like procedure. The difference here, like the complete modified profile likelihood, is that no choice needs to be made for an appropriate parameterization. This problem of removing nuisance parameters can alternatively be handled by assuming that the nuisance parameters have been sampled from a population with some rather flat distribution and making the problem hierarchical. This allows one to effectively integrate out the nuisance parameters within a likelihood framework. But here again some choice needs to be made about an appropriate scale on which the probability should be *flat*. Lastly, the nuisance parameters can be handled by a full Bayesian approach, but here again a flat prior is described on some specific parameterization.

The relevance to this paper is that the authors are effectively working in the opposite direction. They are taking the joint likelihood function which is partly defined in terms of a probability structure, choosing a specific parameterization, and then allowing themselves to maximize over all the parameters. This is in contrast to doing something more appropriate such as integration in the probability direction, arriving at a marginal likelihood and then maximizing across the likelihood space. The choice of a scale for the random effects u is equivalent to choosing an noninformative prior. But the process is going in the opposite direction, attempting to go from probability to likelihood rather than the other direction. The motivation is to have a simple procedure for choosing a best estimate - maximization. This procedure of attempting to move from probability space to likelihood space might be classed as *anti-Bayesian*. However for location and scale models, including random effects in this way, it may be an effective pragmatic way forward when the marginal likelihood is intractable. Certainly many real statistical problems fall into this framework.

Why should we want to estimate random effects? In a standard linear mixed model the BLUP estimates are seen as best estimates for the individual. But the set of estimates as a whole are not representative of the distribution from which they are drawn. It is well known that they are shrunk. It is important to see that any estimate for a random effect needs to be made in the light of what it is going to be used for. Then one can identify whether an estimator delivers good estimates or not. One might argue that *good* involves the estimates of the random effects together being representative of the underlying distribution, in the way that one might want residuals to collectively represent the error

distribution in a regression problem. Equally, they might be required for the basis of some bootstrap procedure. However, when estimation of the random effects is purely done to handle the random part from the model, as suggested in this paper, it may be sufficient to ignore this aspect of the estimation process.

The paper provides insight into methods which are certainly going to become routinely available and routinely practiced. Hopefully it will help users to understand both the possible limitations and the pragmatic advantages of proceeding in this way.

References

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70, 343-365.

Bayarri, M.J., DeGroot, M.H. and Kadane, J.B. (1988). What is the likelihood function (with discussion). *Statistical Decision Theory and Related Topics IV. Vol. 1*, eds S.S. Gupta and J.O. Berger, New York: Springer.

Cox, D.R. and Reid, N. (1987). Parametric orthogonality and approximate conditional inference. *J. R. Staist. Soc.* B, 32, 1-18.

Pawitan, Y. (2001). *In All Likelihood; Statistical Modelling and Inference using Likelihood.* Oxford: Clarendon Press.

Il Do Ha

Department of Asset Management, Daegu Haany University, Gyeongsan, 712-715, Korea. E-mail: idha@dhu.ac.kr

This is an important paper that systematically establishes an h-likelihood framework for the inference of various random-effect models. In general, h-likelihood avoids the integration that is necessary to obtain marginal likelihood and provides a statistically efficient and simple unified framework for such models. However, it is not well known that the h-likelihood can be used in some non-parametric settings. The h-likelihood method, first introduced by Lee and Nelder (1996), has since been well developed and its estimating properties have been extensively studied for HGLMs. Here, I should like to concentrate on h-likelihood inferences of frailty models in multivariate survival analysis. This model is an extension of Cox's proportional-hazards model and has been used for the analysis of correlated survival data in the form of recurrent or multiple event times observed in the same cluster. For models with a gamma frailty distribution the marginal likelihood has a closed form and has been frequently used; see for example Nielsen *et al.* (1992). However, in general the marginal likelihood does not allow such a closed form and requires a numerically intractable integration. As an alternative, Ha *et al.* (2001) and Ha and Lee (2003) have proposed the use of Lee and Nelder's (1996) h-likelihood for the analysis of frailty models with a parametric or nonparametric baseline hazard. Frailty models with parametric baseline hazards (e.g. Weibull) can be directly fitted using a Poisson HGLM technique (Ha and Lee, 2003). Below, I would like to show how the h-likelihood method can give a straightforward estimation for frailty models with a nonparametric baseline hazard.

Given unobserved frailties $U_i = u_i$ ($i = 1, \ldots, q$), the conditional hazard function for the jth ($j = 1, \ldots, n_i$) observation of the ith cluster has the form

$$\lambda_{ij}(t|u_i) = \lambda_0(t) \exp(x_{ij}^T \beta) u_i, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and β is a $p \times 1$ vector of unknown regression parameters for fixed covariates $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$. The frailties U_i are assumed to be independent and identically distributed random variables with frailty parameter σ^2, satisfying $E(U_i) = 1$ and $\text{var}(U_i) = \sigma^2$; for example, gamma, lognormal or inverse Gaussian can be considered as the distribution of U_i. Note that in frailty models the additivity of fixed and random effects can be obtained via the log link,

applied to the conditional hazard

$$\log \lambda_{ij}(t|u_i) = \log \lambda_0(t) + x_{ij}^T\beta + \log u_i.$$

Because in the frailty models (1) the functional form of baseline hazard $\lambda_0(t)$ is unknown, we can follow Breslow's idea, and consider the baseline cumulative hazard function $\Lambda_0(t)$ to be a step function with jumps at the observed death times, $\Lambda_0(t) = \sum_{k:y_{(k)}\le t} \lambda_{0k}$, where $y_{(k)}$ is the kth ($k = 1, \ldots, D$) smallest distinct death time among the observed event times or censored times t_{ij}'s and $\lambda_{0k} = \lambda_0(y_{(k)})$. In this semi-parametric frailty model the number of nuisance parameters λ_{0k} increases with sample size $n = \sum n_i$. The objective of the study is inferences for β and σ^2. Let $\omega = (\omega_1, ..., \omega_D)^T$, where $\omega_k = \log \lambda_{0k}$. Following Lee and Nelder (1996), the h-likelihood for semi-parametric frailty models (1), denoted by h, is defined by

$$h = h(\omega, \beta, \sigma^2) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i},$$

where

$$\sum_{ij} \ell_{1ij} = \sum_k d_{(k)}\omega_k + \sum_{ij} \delta_{ij}\eta_{ij} - \sum_k \exp(\omega_k)\Big\{ \sum_{(i,j)\in R(y_{(k)})} \exp(\eta_{ij})\Big\},$$

$\ell_{1ij} = \ell_{1ij}(\omega, \beta; y_{ij}|u_i)$ is the logarithm of the conditional density function for $y_{ij} = (t_{ij}, \delta_{ij})$ given $U_i = u_i$, $\ell_{2i} = \ell_{2i}(\sigma^2; v_i)$ is the logarithm of the density function for $V_i = \log U_i$ with parameter σ^2, $\eta_{ij} = x_{ij}^T\beta + v_i$ with $v_i = \log u_i$, δ_{ij} is the censoring indicator, being 1 if t_{ij} is observed and 0 if it is censored, $d_{(k)}$ is the number of deaths at $y_{(k)}$ and $R(y_{(k)}) = \{(i, j) : t_{ij} \ge y_{(k)}\}$ is the risk set at $y_{(k)}$. Following Section 5.2 of Lee and Nelder, we can use h for inferences about $\alpha = (\omega^T, \beta^T, v^T)^T$ with $v = (v_1, ..., v_q)^T$ and use an adjusted profile likelihood, $p_\alpha(h)$ for inferences about σ^2. However, fitting frailty models (1) based on h can be difficult because the number of nuisance parameters ω increases with sample size n. Thus, a computationally efficient procedure is necessary to eliminate ω. The development of the adjusted profile likelihood provides a straightforward solution. Ha *et al.* (2001) proposed to use a profile likelihood h^* after eliminating ω, defined by

$$h^* = h|_{\omega=\hat{\omega}},$$

where $\hat{\omega}$ are solutions of the estimating equations $\partial h/\partial\omega = 0$, which become a multinomial likelihood with the ω eliminated (Ha and Lee, 2005*b*). For gamma or log-normal frailty models, h^* also becomes the kernel of the penalized partial likelihood (Ripatti and Palmgren, 2000). Ha *et al.* (2001) further showed that given σ^2 the

estimating equations for $\tau = (\beta^T, v^T)^T$ are obtained by the relationship

$$\partial h^*/\partial \tau = (\partial h/\partial \tau)|_{\omega=\hat{\omega}} .$$

For inference about the frailty parameter σ^2, Ha *et al.* (2001) used the adjusted profile likelihood $p_{\beta,v}(h^*)$, which gives an equivalent inference using $p_{\omega,\beta,v}(h)$. For gamma frailty models, Ha and Lee (2003) confirmed that the second-order Laplace approximation works better. Thus, Ha *et al.*'s (2001) method based on the profile likelihood h^* is a numerically efficient way of eliminating nuisance parameters in the h-likelihood method for semi-parametric frailty models. Furthermore, Ha and Lee (2003, 2005*b*) have demonstrated numerically that the h-likelihood method yields parametric estimators which are less biased than those obtained using maximum marginal likelihood (Nielsen *et al.*, 1992), penalized likelihood (Ripatti and Palmgren, 2000) or best linear unbiased predictor method (Ma *et al.*, 2003).

Various frailty models have been developed; for example, time dependent models (Yau and McGilchrist, 1998), nested models (Yau, 2001), mixture cure models (Yau and Ng, 2001), spatially correlated models (Li and Ryan, 2002) and treatment-by-center interaction models (Glidden and Vittinghoff, 2004). That is, random components in frailty models can be nested or crossed and also temporally and spatially correlated. Now we illustrate how to handle more complex random-effect structures using h-likelihood. The one-component frailty model (1), $\eta = X\beta + Zv$, can be extended to a multi-component model as follows:

$$\eta = X\beta + Z_1 v^{(1)} + Z_2 v^{(2)} + \cdots + Z_k v^{(k)} , \tag{2}$$

where X is the $n \times p$ model matrix corresponding to β, Z_r $(r = 1, \ldots, k)$ are the $n \times q_r$ model matrices corresponding to the $q_r \times 1$ frailties $v^{(r)}$, and $v^{(r)}$ and $v^{(l)}$ are independent for $r \neq l$. Let $Z = (Z_1, \ldots, Z_k)$, $v = (v^{(1)T}, \ldots, v^{(k)T})^T$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)^T$ and $q = \sum_r q_r$. Then, the multi-component model (2) can be written as in (1). Thus, the extension of h-likelihood results from one-component models to multi-component models is straightforward. Furthermore, for the model selection in (2) we may use deviance based on adjusted profile likelihood $p_v(h^*)$ or $p_{\beta,v}(h^*)$ (Ha, Lee and MacKenzie, 2005). In these multi-component models the marginal likelihood is difficult to compute. Moreover, Bayesian approaches using MCMC may be computationally intensive. The h-likelihood approach provides a simple, unifying, framework and a numerically efficient fitting algorithm for inference.

More recently, the h-likelihood methods have been applied to various survival areas:

(i) Random-effect models with non-proportional hazards (MacKenzie, Ha and Lee, 2003),
(ii) Joint modelling of repeated measures and survival data (Ha, Park and Lee, 2003),

(iii) Frailty models with structured dispersion (Noh, Ha and Lee, 2005),
(iv) Multilevel mixed linear models with censoring (Ha and Lee, 2005*a*),
(v) Genetic mixed linear models for twin survival data (Ha, Lee and Pawitan, 2005).

In summary, the h-likelihood gives a systematic extended-likelihood inference for various survival models such as frailty models and mixed linear models, leading to a useful methodology for multivariate survival analysis.

Additional References

Glidden, D.V. and Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 23, 369-388.

Ha, I.D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, 12, 663-681.

Ha, I.D., Lee, Y. and Pawitan (2005). Genetic mixed linear models for twin survival data under LTRC. manuscript prepared for publication.

Ha, I.D., Lee, Y. and MacKenzie (2005). Model selection for multi-component frailty models. *Lifetime Data Analysis*, in press.

Ha, I.D., Park, T. and Lee, Y. (2003). Joint modelling of repeated measures and survival time data. *Biometrical Journal*, 45, 647-658.

Li, Y. and L. Ryan (2002). Modelling spatial survival data using semiparametric frailty models. *Biometrics*, 58, 287-297.

Ma, R., Krewski, D. and Burnett, R.T. (2003). Random effects Cox models: a Poisson modelling approach. *Biometrika*, 90, 157-169.

MacKenzie, G., Ha, I.D. and Lee, Y. (2003). Non-PH multivariate survival models based on the GTDL. *The proceedings of 18th International Workshop on Statistical Modelling*, Leuven, Belgium, 273-277.

Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19, 25-44.

Noh, M., Ha, I.D. and Lee, Y. (2005). Dispersion frailty models and HGLMs. *Statistics in Medicine*, in press.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56, 1016-22.

Yau, K.K.W. (2001). Multilevel models for survival analysis with random effects. *Biometrics*, 57, 96-102.

Yau, K.K.W. and McGilchrist, C.A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine*, 17, 1201-1213.

Yau, K.K.W. and Ng, A.S.K. (2001). Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistics in Medicine*, 20, 1591-1607.

Geert Molenberghs

Center for Statistics, Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium. Email: geert.molenberghs@uhasselt.be

Geert Verbeke

Biostatistical Centre, Catholic University of Leuven, U.Z. St.-Rafaël, Kapucijnenvoer 35, B-3000 Leuven, Belgium. Email: geert.verbeke@med.kuleuven.be

Repeatedly measured outcomes, generalized linear models, and random-effects models are very common in statistical practice. Combining these for repeated non-Gaussian responses has lead to a number of modeling families and inferential approaches. Important work has been done by Breslow and Clayton (1993), Wolfinger and O'Connell (1993), and Lee and Nelder (1996, 2001ab, 2004, 2005). While there is communality between them, there are important differences, both in the model formulation as well as in the inferential route taken (Molenberghs and Verbeke 2005).

To fix ideas, we will focus on the situation of subject $i = 1, \ldots, N$ measured repeatedly over time at a number of occasions $j = 1, \ldots, n$, but our comments apply to general correlated, hierarchical data settings. Denote the outcome for subject i at time j by Y_{ij}. The so-called *generalized linear mixed model* (GLMM) has become very popular, not only due to the work by Breslow and Clayton (1993) and Wolfinger and O'Connell (1993), but also generated by the existence of procedures in standard software packages, such as SAS and MLwiN. The GLMM assumes a subject's random effects to be following a normal distribution, conditional upon which a generalized linear model is followed for the individual measures. As the authors point out, a lot of emphasis as been placed on marginal likelihood, where the marginal distribution, integrating out the random effects, is considered the basis of inference for fixed effects and variance components, perhaps complemented with empirical Bayes estimation for the random effects. Unlike in very specific cases, such as a normally distributed outcome, the marginal likelihood will be cumbersome because the integral mentioned earlier has no analytic solution. This applies, in particular, to the very popular logistic-normal GLMM, where the binary outcomes, given random effects, are modeled by logistic regression. Within the marginal likelihood framework, common approaches to this problem involve (1) approximation to the integrand (e.g., Laplace transforms), (2) approximation to the data (so-called penalized quasi-likelihood, PQL, or marginal quasi-likelihood, MQL),

and (3) approximation of the integral (numerical integration) (Molenberghs and Verbeke 2005). As Lee and Nelder correctly point out, all of these suffer from serious drawbacks. Numerical integration, when done with sufficiently high numerical accuracy, is on one hand appealing since it virtually avoids numerical bias, but the process is computer intensive. On the other hand, such methods as PQL and MQL can be severely biased, especially for short binary sequences, which often is the undoing of the advantage deriving from numerical simplicity.

Therefore, the work by Lee and Nelder is very welcome, since joint likelihood or, more precisely, properly conducted *h*-likelihood, provides a valuable *third way* in between marginal likelihood and Bayesian methodology. The original papers perhaps have not caught on as they should have, but the current paper clarifies a number of issues, giving the methodology its proper place.

In the light of the numerical complexity and bias issues surrounding conventional marginal likelihood, it is important to consider a proper alternative, even if that might be subject, in some cases, to small amounts of bias, and necessitates careful guidelines for its use. Let us expand on each of these in turn. First, as Lee and Nelder show, bias is present in a number of joint likelihood settings, but they equally well show that *h*-likelihood in some cases eliminates and in others drastically reduces bias. Both the theoretical developments are convincing and the examples insightful, even though developments for one or a few more general settings, such as a proper longitudinal setting with replication, would have been welcome. Second, Lee and Nelder provide a set of guidelines/rules to be followed when applying *h*-likelihood. One has to ensure the model is formulated and the method applied at the proper scale, where fixed and random effects add linearly. As the authors state, this is possible in broad classes of frequently encountered models, although not in general. Further, one has to carefully distinguish between inferences for location parameters, dispersion parameters, and random effects. In some cases, the distinction between a location and a dispersion parameter is not clear, and the distinction between both may depend on the parameterizations. These are subtle issues and the user ought to appreciate the need to practice with the method before getting a good feel for it. In this sense, the use of the functions of type similar to (15), may appear somewhat off-putting for the novice user. But, especially in situations where marginal likelihood, or a suitable approximation to it, is either biased or computationally beyond reach, the authors' method is a viable alternative. Moreover, in many other frequently encountered settings, one is confronted with specific and somewhat *ad hoc* looking guidelines as well such as, for example, generalized estimating equations (GEE, Liang and Zeger 1986, Molenberghs and Verbeke 2005). Another setting where care is needed occurs when choosing between ML and REML in linear mixed models. With the latter method, likelihood ratios for variance components are valid, but for fixed effects are not. In the same model, inference depends on whether or not negative variance components are allowed (Verbeke and Molenberghs 2000).

Note also that Henderson's (1975) application of joint likelihood ideas has proven a powerful tool, especially in the early days of the linear mixed model. It is still used in a number of software packages, as an alternative to classical Fisher scoring or Newton Raphson, the latter of which are inspired by marginal likelihood.

The *h*-loglikelihood has a number of appealing side properties, such as correctly handling the variance inflation in *D*, caused by estimating fixed effects.

A tangential issue is that random-effects models, regardless of the inferential path followed, can be applied to cross-sectional data, as long as the random effects and the residual error are not in a linear relationship. This applies to most GLM settings, as well as to survival outcomes with random effects, in that context usually termed frailties. However, a researcher should then reflect very carefully on whether the parameters, resulting from such a model, really have substantive meaning, or merely allow for a slightly more flexible model than the one without random effects. This contrasts with proper longitudinal or clustered data, where a genuine hierarchy applies.

Additional References

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Molenberghs, G. and Verbeke, G. (2005). *Discrete Longitudinal Data.* New York: Springer.

Verbeke and Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48, 233-243.

Authors' rejoinder

We thank the discussants for raising many interesting points, and regret that space restriction does not allow us to reply in detail.

Professors Molenbergs and Verbeke point out that marginal likelihood has been considered as basis of inferences about fixed parameters, complemented with empirical Bayes (EB) estimation for random effects. Numerical integration is often computationally intensive (often not feasible) and other methods, such as PQL and MQL, are severely biased. Thus, our h-likelihood procedure provides a valuable third way between marginal and Bayesian methodology. We believe that our approach completes Fisher likelihood by allowing inferences about random parameters. As we noted, the EB procedure uses information only on a component $\log f_\theta(v|y)$ of the h-loglihood $h = \log f_\theta(v|y) + \log f_\theta(y)$, so that it cannot reflect the uncertainty about θ in the other component $\log f_\theta(y)$. Thus, the h-likelihood approach handles correctly the variance inflation of standard errors of estimates of random parameters caused by estimating extra fixed parameters. As we shall show below it also handles correctly those of fixed parameters caused by estimating extra random parameters. The h-likelihood provides satisfactory estimation in wide class of models, and extensive simulation studies have been done for HGLMs for longitudinal studies by Noh and Lee (2004). There seems to be no other method which performs better than the h-likelihood method, even in the extreme case of binary matched pairs. Furthermore, there can be several alternative random-effect models leading to the same marginal model and the h-likelihood inferences from these are equivalent (Lee and Nelder, 2005). GLMs with random effects (HGLMs) can also be applied to repeated measures of survival outcomes.

Professor Ha explains his experience of using the h-likelihood methods for the analysis of such survival data. He explains how the h-likelihood gives a straightforward procedure for eliminating nuisance parameters associated with the non-parametric baseline hazards, by simply profiling them. Furthermore, he points out that in frailty models the h-likelihood provides estimators less biased than those obtained by using ML, penalized likelihood or the best linear unbiased predictor method. It is also nice that these frailty models,either with parametric or non-parametric baseline hazards, can be fitted by Poisson HGLMs. The use of h-likelihood makes this point clear (Ha and Lee, 2005). Furthermore, the h-likelihood method can be applied to left truncation and genetic studies. He also pointed out that the h-likelihood procedure contributes to model-selection problems for random-effect models, which is not well understood yet. We hope that this paper will improve the understanding of the h-likelihood approach.

Dr Roger points out that the original modified (adjusted in our paper) profile likelihood of Barndorff-Nielsen (1983), to eliminate nuisance fixed parameters such as $p_\beta(m)$, also includes a Jacobian term and it is the cunning parameterization, allowing the

parameter orthogonality of Cox and Reid (1987), that makes the Jacobian term almost constant. Similarly, the h-likelihood procedure overcomes a difficulty associated with intractable Jacobian terms by choosing a proper parameterization of random parameters. If orthogonal parameterization of fixed parameters is not available, the intractable Jacobian term may not be eliminated, which limits the application of the method. However, for models without an apparent parameterization of random parameters the adjusted profile likelihood $p_u(h)$, we show, can still gives satisfactory estimation for fixed parameters. He distinguishes random-effect BLUP estimation from error estimation. We can show that error estimators $\hat{e}$ are also BLUP and that there is shrinkage estimation because $E(\sum \hat{e}_i^2/n) < \sigma^2$. Therefore, there is no clear distinction between random-effect estimation and error estimation. Thus, random-effects estimates can represent the underlying distribution after a proper standardization, and these are useful for model checking, using normal probability plots etc. (Lee and Nelder, 2001a). We thank him also for his insightful comments about joint estimation. Professor Pawitan suggest us an alternative justification of natural parameterization of u for joint estimation as follows.

When joint estimation gives the marginal ML estimation

In this paper we have shown that in HGLMs there is a particular joint likelihood whose maximization gives meaningful estimators of the random parameters. Maintaining invariance of inferences from the joint likelihood for trivial re-expressions of the underlying model leads to a unique definition of the h-likelihood. In some HGLMs simple maximization of the h-likelihood gives the marginal ML estimators for the location parameters. We now give a condition for the joint inference from the h-likelihood to be the same as that from marginal likelihood. Let β_1 and β_2 be an arbitrary pair of values of β. The evidence about these two parameter values is contained in the likelihood ratio

$$\frac{L(\beta_1; y)}{L(\beta_2; y)}$$

Suppose there exists a scale v, such that the likelihood ratio is preserved in the following sense

$$\frac{L(\beta_1, \hat{v}_{\beta_1}; y, v)}{L(\beta_2, \hat{v}_{\beta_2}; y, v)} = \frac{L(\beta_1; y)}{L(\beta_2; y)}, \tag{1}$$

where $\hat{v}_{\beta_1}$ and $\hat{v}_{\beta_2}$ are the MLEs of v when β is known at β_1 and β_2, so that $\hat{v}_\beta$ is *information-neutral* concerning β. Alternatively, (1) is equivalent to

$$\frac{L(\beta_1, \hat{v}_{\beta_1}; v|y)}{L(\beta_2, \hat{v}_{\beta_2}; v|y)} = 1,$$

which means that neither the likelihood component $L(\beta, \hat{v}_\beta; v|y)$ nor $\hat{v}_\beta$ carry any information about β, as is required by the classical likelihood principle.

Let $I_m(\hat{\beta})$ be the observed Fisher information of the MLE $\hat{\beta}$ from the marginal likelihood $L(\beta; y)$ and let the partitioned matrix

$$I_h^{-1}(\hat{\beta}, \hat{v}) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}$$

be the inverse of the observed Fisher information matrix of $(\hat{\beta}, \hat{v})$ from the h-loglihood $h(\beta, v; y, v)$, where I^{11} corresponds to the $\hat{\beta}$ part. If the condition (1) holds, then we can show the following:

(i) The MLE $\hat{\beta}$ from the marginal likelihood $L(\beta; y)$ coincides with the $\hat{\beta}$ from the joint maximizer of the h-likelihood $h(\beta, v; y, v)$.

(ii) The information matrices for $\hat{\beta}$ from the two likelihoods also match, in the sense that

$$I_m^{-1} = I^{11},$$

which means that Wald-based inference on the fixed parameter β can be obtained directly from the h-likelihood framework.

(iii) Furthermore, I^{22} yields an estimate of $\text{var}(\hat{v} - v)$. If $\hat{v} = E(v|y)|_{\beta=\beta}$ this estimates $\text{var}(\hat{v} - v) \geq E\{\text{var}(v|y)\}$, accounting for the inflation of variance caused by estimating β. □

Condition (1) is too restrictive because it does not cover models with dispersion parameters ϕ, so that we may generalize it as follows. Suppose there are two subsets of the fixed parameters (β, ϕ) such that

$$\frac{L(\beta_1, \phi, \hat{v}_{\beta_1,\phi}; y, v)}{L(\beta_2, \phi, \hat{v}_{\beta_2,\phi}; y, v)} = \frac{L(\beta_1, \phi; y)}{L(\beta_2, \phi; y)}, \tag{2}$$

but

$$\frac{L(\beta, \phi_1, \hat{v}_{\beta,\phi_1}; y, v)}{L(\beta, \phi_2, \hat{v}_{\beta,\phi_2}; y, v)} \neq \frac{L(\beta, \phi_1; y)}{L(\beta, \phi_2; y)}.$$

Here v is information-neutral for β not for ϕ. This means that joint inference using the h-likelihood is possible only for (β, v), with ϕ needing a marginal loglihood or $p_v(h)$. In HGLMs the location parameters, we believe, approximately satisfy the condition (1) or its generalized version (2).

Finally, h-likelihood gives indeed a unified and pragmatic way of implementing likelihood inferences for complex models involving unobservables. The simplicity of inferential procedure of h-likelihood is important for inferences about complex models having many components. We believe that h-likelihood will become a major apparatus for statistical inference.

Lee, Y. and Nelder, J. A. (2005). Fitting via alternative random-effect models. to appear at *Statist. Comp*.

SORT 29 (2) July-December 2005, 183-200

Muliere and Scarsini's bivariate Pareto distribution: sums, products, and ratios

Saralees Nadarajah[1], Samuel Kotz[2]

[1]*University of Nebraska,* [2]*The George Washington University*

Abstract

We derive the exact distributions of $R = X + Y$, $P = X\,Y$ and $W = X/(X + Y)$ and the corresponding moment properties when X and Y follow Muliere and Scarsini's bivariate Pareto distribution. The expressions turn out to involve special functions. We also provide extensive tabulations of the percentage points associated with the distributions. These tables –obtained using intensive computing power– will be of use to practitioners of the bivariate Pareto distribution.

MSC: 33C90, 62E99

Keywords: incomplete beta function, Gauss hypergeometric function, Muliere and Scarsini's bivariate Pareto distribution, products of random variables, ratios of random variables, sums of random variables.

1 Introduction

Since the 1930s, the statistics literature has seen many developments in the theory and applications of linear combinations and ratios of random variables. Some of these include:

- Ratios of normal random variables appear as sampling distributions in single equation models, in simultaneous equations models, as posterior distributions for parameters of regression models and as modeling distributions, especially in

Address for correspondence: Saralees Nadarajah. Department of Statistics. University of Nebraska. Lincoln, NE 68583. Samuel Kotz. Department of Engineering Management and Systems Engineering. The George Washington University. Washington, D.C. 20052

Received: December 2004
Accepted: February 2005

economics when demand models involve the indirect utility function (details in Yatchew, 1986).

- Weighted sums of uniform random variables –in addition to the well known application to the generation of random variables– have applications in stochastic processes which in many cases can be modeled by these weighted sums. In computer vision algorithms these weighted sums play a pivotal role (Kamgar-Parsi *et al.*, 1995). An earlier application of the linear combinations of uniform random variables is given in connection with the distribution of errors in *n*th tabular differences Δ^n (Lowan and Laderman, 1939).

- Ratio of linear combinations of chi-squared random variables are part of von Neumann's (1941) test statistics (mean square successive difference divided by the variance). These ratios appear in various two-stage tests (Toyoda and Ohtani, 1986). They are also used in tests on structural coefficients of a multivariate linear functional relationship model (details in Chaubey and Nur Enayet Talukder (1983) and Provost and Rudiuk (1994)).

- Sums of independent gamma random variables have applications in queuing theory problems such as determination of the total waiting time and in civil engineering problems such as determination of the total excess water flow into a dam. They also appear in test statistics used to determine the confidence limits for the coefficient of variation of fiber diameters (Linhart (1965) and Jackson (1969)) and in connection with the inference about the mean of the two-parameter gamma distribution (Grice and Bain, 1980).

- Linear combinations of inverted gamma random variables are used for testing hypotheses and interval estimation based on generalized *p*-values, specifically for the Behrens-Fisher problem and variance components in balanced mixed linear models (Witkovský, 2001).

- As to the Beta distributions their linear combinations occur in calculations of the power of a number of tests in ANOVA (Monti and Sen, 1976) among other applications. More generally, the linear combinations are used for detecting changes in the location of the distribution of a sequence of observations in quality control problems (Lai, 1974). Pham-Gia and Turkkan (1993, 1994, 1998, 2002) and Pham-Gia (2000) provided applications of sums and ratios to availability, Bayesian quality control and reliability.

- Linear combinations of the form $T = a_1 t_{f_1} + a_2 t_{f_2}$, where t_f denotes the Student *t* random variable based on f degrees of freedom, represents the Behrens-Fisher statistic and – as early as the middle of the twentieth century – Stein (1945) and Chapman (1950) developed a two-stage sampling procedure involving the T to test whether the ratio of two normal random variables is equal to a specified constant.

- Weighted sums of the Poisson parameters are used in medical applications for directly standardized mortality rates (Dobson *et al.*, 1991).

In this paper, we consider the distributions of $R = X+Y$, $P = XY$ and $W = X/(X+Y)$ when X and Y are correlated Pareto random variables with the joint survivor function expressed as the mixture of two components:

$$\bar{F}(x,y) = \frac{\lambda_1+\lambda_2}{\lambda_0+\lambda_1+\lambda_2}\bar{F}_a(x,y) + \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2}\bar{F}_s(x,y), \quad (1)$$

where $\bar{F}_a(x,y)$ is the absolutely continuous part with respect to Lebesgue measure given by

$$\bar{F}_a(x,y) = \left(\frac{x}{\beta}\right)^{-\lambda_1}\left(\frac{y}{\beta}\right)^{-\lambda_2}\left\{\max\left(\frac{x}{\beta},\frac{y}{\beta}\right)\right\}^{-\lambda_0} \quad (2)$$

and $\bar{F}_s(x,y)$ is the singular part concentrating on the line $x = y$ given by

$$\bar{F}_s(x,y) = \left\{\max\left(\frac{x}{\beta},\frac{y}{\beta}\right)\right\}^{-(\lambda_0+\lambda_1+\lambda_2)} \quad (3)$$

for $x \geq \beta$, $y \geq \beta$, $\beta > 0$, $\lambda_0 > 0$, $\lambda_1 > 0$ and $\lambda_2 > 0$. The joint density of the absolutely continuous part is:

$$f_a(x,y) = \begin{cases} \dfrac{\lambda_2(\lambda_0+\lambda_1)}{\beta^2}\left(\dfrac{x}{\beta}\right)^{-(1+\lambda_0+\lambda_1)}\left(\dfrac{y}{\beta}\right)^{-(1+\lambda_2)}, & \text{if } x > y \geq \beta, \\ \dfrac{\lambda_1(\lambda_0+\lambda_2)}{\beta^2}\left(\dfrac{y}{\beta}\right)^{-(1+\lambda_0+\lambda_2)}\left(\dfrac{x}{\beta}\right)^{-(1+\lambda_1)}, & \text{if } y > x \geq \beta. \end{cases}$$

This distribution is due to Muliere and Scarsini (1987) and therefore known as Muliere and Scarsini's bivariate Pareto distribution. It has received applications in several areas especially in reliability (see, for example, Kotz *et al* (2000)).

The paper is organized as follows. In Sections 2 and 3, we derive explicit expressions for the pdfs and moments of $R = X + Y$, $P = XY$ and $W = X/(X + Y)$. In Section 4, we provide extensive tabulations of the associated percentage points, obtained by means of intensive computing power. These values will be of use to the practitioners of the bivariate Pareto distribution.

The calculations of this paper involve several special functions, including the incomplete beta function defined by

$$B_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$$

and, the Gauss hypergeometric function defined by

$$G(a,b;c;x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{x^k}{k!},$$

where $(e)_k = e(e+1)\cdots(e+k-1)$ denotes the ascending factorial. We also need the following important lemma.

Lemma 1 *(Equation (3.194.2), Gradshteyn and Ryzhik, 2000) For $\mu > \nu$,*

$$\int_u^{\infty} \frac{x^{\mu-1}}{(1+\beta x)^{\nu}} dx = \frac{u^{\mu-\nu}}{\beta^{\nu}(\nu-\mu)} G\left(\nu, \nu-\mu; \nu-\mu+1; -\frac{1}{\beta u}\right).$$

The properties of the above special functions can be found in Prudnikov *et al.* (1986) and Gradshteyn and Ryzhik (2000).

2 Probability density functions

Theorems 1 to 3 derive the pdfs of $R = X + Y$, $P = XY$ and $W = X/(X+Y)$ when X and Y are distributed according to (1)–(3).

Theorem 1 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned} f_R(r) = {} & \frac{\lambda_0 (2\beta)^{\lambda_0+\lambda_1+\lambda_2}}{r^{1+\lambda_0+\lambda_1+\lambda_2}} + \frac{(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}}{(\lambda_0+\lambda_1+\lambda_2) r^{1+\lambda_0+\lambda_1+\lambda_2}} K_1(r) \\ & + \frac{(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}}{(\lambda_0+\lambda_1+\lambda_2) r^{1+\lambda_0+\lambda_1+\lambda_2}} K_2(r) \end{aligned} \tag{4}$$

for $2\beta \le r < \infty$, where

$$\begin{aligned} K_1(r) = {} & \left(\frac{r}{\beta} - 1\right)^{\lambda_2} G\left(-\lambda_0-\lambda_1-\lambda_2, -\lambda_2; 1-\lambda_2; -\frac{\beta}{r-\beta}\right) \\ & - G(-\lambda_0-\lambda_1-\lambda_2, -\lambda_2; 1-\lambda_2; -1) \end{aligned}$$

and

$$\begin{aligned} K_2(r) = {} & G(-\lambda_0-\lambda_1-\lambda_2, -\lambda_1; 1-\lambda_1; -1) \\ & - \left(\frac{r}{\beta} - 1\right)^{\lambda_1} G\left(-\lambda_0-\lambda_1-\lambda_2, -\lambda_1; 1-\lambda_1; -\frac{\beta}{r-\beta}\right). \end{aligned}$$

Proof. Set $(R, W) = (X + Y, X/R)$ and note that the Jacobian is R for the continuous part and $1/2$ for the singular part. From (1)–(3), the joint pdf of (R, W) can be written as

$$f(r,w) = \frac{\lambda_1+\lambda_2}{\lambda_0+\lambda_1+\lambda_2} f_a(r,w) + \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2} f_s(r,w), \tag{5}$$

where

$$f_a(r,w) = \begin{cases} \lambda_2(\lambda_0+\lambda_1)\beta^{\lambda_0+\lambda_1+\lambda_2} r(rw)^{-(1+\lambda_0+\lambda_1)}(r(1-w))^{-(1+\lambda_2)}, & \text{if } w > 1/2, \\ \lambda_1(\lambda_0+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2} r(rw)^{-(1+\lambda_1)}(r(1-w))^{-(1+\lambda_0+\lambda_2)}, & \text{if } w < 1/2 \end{cases} \tag{6}$$

and

$$f_s(r,w) = (1/2)(\lambda_0+\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}(rw)^{-(1+\lambda_0+\lambda_1+\lambda_2)}. \tag{7}$$

Note that $f_a(r,w)$ is the joint density of the absolutely continuous part and that $f_s(r,w)$ is the density of the singular part along the line $w = 1/2$. Thus, the pdf of R can be written as

$$\begin{aligned} f_R(r) &= \frac{\lambda_0(2\beta)^{\lambda_0+\lambda_1+\lambda_2}}{r^{1+\lambda_0+\lambda_1+\lambda_2}} + \frac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}}{(\lambda_0+\lambda_1+\lambda_2)r^{1+\lambda_0+\lambda_1+\lambda_2}} I_1(r) \\ &\quad + \frac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}}{(\lambda_0+\lambda_1+\lambda_2)r^{1+\lambda_0+\lambda_1+\lambda_2}} I_2(r), \end{aligned} \tag{8}$$

where

$$I_1(r) = \int_{1/2}^{1-\beta/r} w^{-(1+\lambda_0+\lambda_1)}(1-w)^{-(1+\lambda_2)}dw$$

and

$$I_2(r) = \int_{\beta/r}^{1/2} w^{-(1+\lambda_1)}(1-w)^{-(1+\lambda_0+\lambda_2)}dw.$$

These integrals can be calculated by application of Lemma 1. Setting $u = w/(1-w)$, one can calculate

$$\begin{aligned} I_1(r) &= \int_1^{r/\beta-1} u^{-(1+\lambda_0+\lambda_1)}(1+u)^{\lambda_0+\lambda_1+\lambda_2}du \\ &= \lambda_2^{-1}K_1(r), \end{aligned} \tag{9}$$

where the second step follows by application of Lemma 1. Similarly, setting $u =$

$(1-w)/w$, one can show that

$$\begin{aligned} I_2(r) &= \int_{r/\beta-1}^{1} u^{-(1+\lambda_0+\lambda_2)}(1+u)^{\lambda_0+\lambda_1+\lambda_2}du \\ &= \lambda_1^{-1}K_2(r). \end{aligned} \tag{10}$$

The result of the theorem follows by substituting (9) and (10) into (8). ■

Theorem 2 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned} f_P(p) &= \frac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{(\lambda_2-\lambda_1-\lambda_0)(\lambda_0+\lambda_1+\lambda_2)}\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(\lambda_0+\lambda_1+\lambda_2)/2-1}\left\{\beta^{\lambda_0+\lambda_1-\lambda_2}p^{(\lambda_2-\lambda_1-\lambda_0)/2}-1\right\} \\ &\quad +(1/2)\lambda_0\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(\lambda_0+\lambda_1+\lambda_2)/2-1} \\ &\quad +\frac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{(\lambda_0+\lambda_2-\lambda_1)(\lambda_0+\lambda_1+\lambda_2)}\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(1+\lambda_0+\lambda_2)}\left\{p^{(\lambda_0+\lambda_2-\lambda_1)/2}-\beta^{\lambda_0+\lambda_2-\lambda_1}\right\} \end{aligned} \tag{11}$$

for $\beta^2 < p < \infty$.

Proof. Set $(X,P)=(X,XY)$ and note that the Jacobian is $1/X$. From (1)–(3), the joint pdf of (X,P) can be written as

$$f(x,p) = \frac{\lambda_1+\lambda_2}{\lambda_0+\lambda_1+\lambda_2}f_a(x,p)+\frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2}f_s(x,p),$$

where

$$f_a(x,p) = \begin{cases} \lambda_2(\lambda_0+\lambda_1)\beta^{\lambda_0+\lambda_1+\lambda_2}x^{-(2+\lambda_0+\lambda_1)}(p/x)^{-(1+\lambda_2)}, & \text{if } x > \sqrt{p}, \\ \lambda_1(\lambda_0+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}x^{-(2+\lambda_1)}(p/x)^{-(1+\lambda_0+\lambda_2)}, & \text{if } x < \sqrt{p} \end{cases}$$

and

$$f_s(x,p) = (1/2)(\lambda_0+\lambda_1+\lambda_2)\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(\lambda_0+\lambda_1+\lambda_2)/2-1}.$$

Note that $f_a(x,p)$ is the joint density of the absolutely continuous part and that $f_s(x,p)$ is the density of the singular part along the line $x=\sqrt{p}$. Thus, the pdf of P can be written

as

$$\begin{aligned} f_P(p) &= \frac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(1+\lambda_2)}\int_{\sqrt{p}}^{p/\beta}x^{-(1+\lambda_0+\lambda_1-\lambda_2)}dx \\ &\quad +(1/2)\lambda_0\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(\lambda_0+\lambda_1+\lambda_2)/2-1} \\ &\quad +\frac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}p^{-(1+\lambda_0+\lambda_2)}\int_{\beta}^{\sqrt{p}}x^{\lambda_0+\lambda_2-\lambda_1-1}dx. \end{aligned}$$

The result of the theorem follows by elementary integration of the above integrals. ■

Theorem 3 *If X and Y are jointly distributed according to (1)–(3) then*

$$f_W(w) = \begin{cases} \dfrac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{(\lambda_0+\lambda_1+\lambda_2)^2}\dfrac{(1-w)^{\lambda_0+\lambda_1-1}}{w^{\lambda_0+\lambda_1+1}}, & \text{if } w > 1/2, \\ \dfrac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2}, & \text{if } w = 1/2, \\ \dfrac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{(\lambda_0+\lambda_1+\lambda_2)^2}\dfrac{w^{\lambda_0+\lambda_2-1}}{(1-w)^{\lambda_0+\lambda_2+1}}, & \text{if } w < 1/2 \end{cases} \tag{12}$$

for $0 < w < 1$.

Proof. Using (5)–(7), one can write

$$f_W(w) = \begin{cases} \dfrac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}w^{-(1+\lambda_0+\lambda_1)}(1-w)^{-(1+\lambda_2)} \\ \quad \times\displaystyle\int_{\beta/(1-w)}^{\infty}r^{-(1+\lambda_0+\lambda_1+\lambda_2)}dr \qquad \text{if } w > 1/2, \\ (1/2)\lambda_0\beta^{\lambda_0+\lambda_1+\lambda_2}w^{-(1+\lambda_0+\lambda_1+\lambda_2)}\displaystyle\int_{2\beta}^{\infty}r^{-(1+\lambda_0+\lambda_1+\lambda_2)}dr, \\ \qquad\qquad \text{if } w = 1/2, \\ \dfrac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}w^{-(1+\lambda_1)}(1-w)^{-(1+\lambda_0+\lambda_2)} \\ \quad \times\displaystyle\int_{\beta/w}^{\infty}r^{-(1+\lambda_0+\lambda_1+\lambda_2)}dr \qquad \text{if } w < 1/2. \end{cases} \tag{13}$$

The result of the theorem follows by elementary integration of the above integrals. ■

Using special properties of the hypergeometric functions, one can derive elementary forms for the pdf in (4). This is illustrated in the corollaries below.

Corollary 1 *If X and Y are jointly distributed according to (1)–(3) and if $\lambda_1 \geq 1$ is an integer then the pdf of R is given by (4) with*

$$\begin{aligned} K_2(r) &= \sum_{k=0}^{\lambda_1} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_1)_k\,(-1)^k}{(1-\lambda_1)_k\,k!} \\ &\quad -\left(\frac{r}{\beta}-1\right)^{\lambda_1} \sum_{k=0}^{\lambda_1} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_1)_k\,(-1)^k}{(1-\lambda_1)_k\,k!} \left(\frac{\beta}{r-\beta}\right)^k. \end{aligned}$$

Corollary 2 *If X and Y are jointly distributed according to (1)–(3) and if $\lambda_2 \geq 1$ is an integer then the pdf of R is given by (4) with*

$$\begin{aligned} K_1(r) &= \left(\frac{r}{\beta}-1\right)^{\lambda_2} \sum_{k=0}^{\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_2)_k\,(-1)^k}{(1-\lambda_2)_k\,k!} \left(\frac{\beta}{r-\beta}\right)^k \\ &\quad -\sum_{k=0}^{\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_2)_k\,(-1)^k}{(1-\lambda_2)_k\,k!}. \end{aligned}$$

Corollary 3 *If X and Y are jointly distributed according to (1)–(3) and if $\lambda_0+\lambda_1+\lambda_2 \geq 1$ is an integer then the pdf of R is given by (4) with*

$$\begin{aligned} K_1(r) &= \left(\frac{r}{\beta}-1\right)^{\lambda_2} \sum_{k=0}^{\lambda_0+\lambda_1+\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_2)_k\,(-1)^k}{(1-\lambda_2)_k\,k!} \left(\frac{\beta}{r-\beta}\right)^k \\ &\quad -\sum_{k=0}^{\lambda_0+\lambda_1+\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_2)_k\,(-1)^k}{(1-\lambda_2)_k\,k!} \end{aligned}$$

and

$$\begin{aligned} K_2(r) &= \sum_{k=0}^{\lambda_0+\lambda_1+\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_1)_k\,(-1)^k}{(1-\lambda_1)_k\,k!} \\ &\quad -\left(\frac{r}{\beta}-1\right)^{\lambda_1} \sum_{k=0}^{\lambda_0+\lambda_1+\lambda_2} \frac{(-\lambda_0-\lambda_1-\lambda_2)_k\,(-\lambda_1)_k\,(-1)^k}{(1-\lambda_1)_k\,k!} \left(\frac{\beta}{r-\beta}\right)^k. \end{aligned}$$

Figures 1 to 3 illustrate the shape of the pdfs (4), (11) and (12) for selected values of λ_0, λ_1 and λ_2. Each plot contains four curves corresponding to selected values of λ_1 and λ_2. The effect of the parameters is evident.

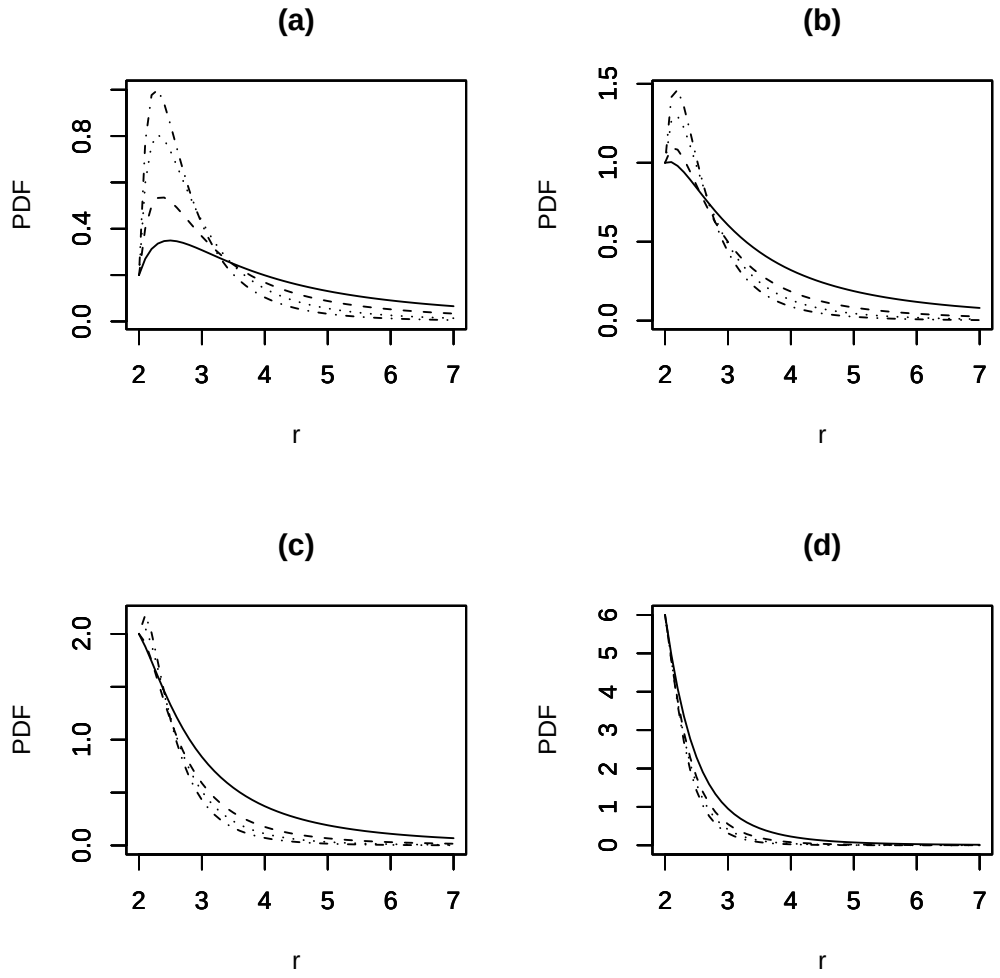

Figure 1: *Plots of the pdf of (4) for* $\beta = 1$ *and (a):* $\lambda_0 = 0.1$*; (b):* $\lambda_0 = 0.5$*; (c):* $\lambda_0 = 1$*; and, (d):* $\lambda_0 = 3$*. The four curves in each plot are: the solid curve (*$\lambda_1 = 1$*,* $\lambda_2 = 1$*), the curve of lines (*$\lambda_1 = 3$*,* $\lambda_2 = 1$*), the curve of dots (*$\lambda_1 = 3$*,* $\lambda_2 = 2$*), and the curve of lines and dots (*$\lambda_1 = 3$*,* $\lambda_2 = 3$*).*

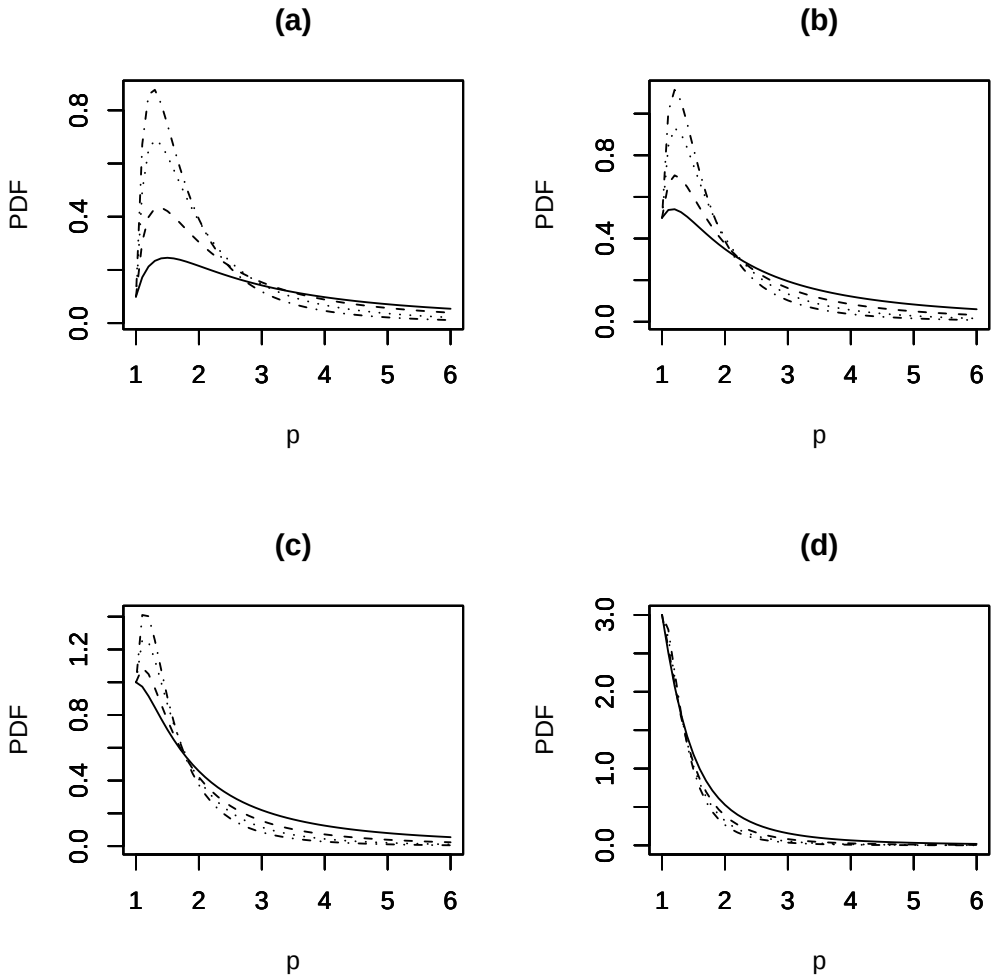

Figure 2: *Plots of the pdf of (11) for* $\beta = 1$ *and (a):* $\lambda_0 = 0.1$*; (b):* $\lambda_0 = 0.5$*; (c):* $\lambda_0 = 1$*; and, (d):* $\lambda_0 = 3$*. The four curves in each plot are: the solid curve (*$\lambda_1 = 1$*,* $\lambda_2 = 1$*), the curve of lines (*$\lambda_1 = 3$*,* $\lambda_2 = 1$*), the curve of dots (*$\lambda_1 = 3$*,* $\lambda_2 = 2$*), and the curve of lines and dots (*$\lambda_1 = 3$*,* $\lambda_2 = 3$*).*

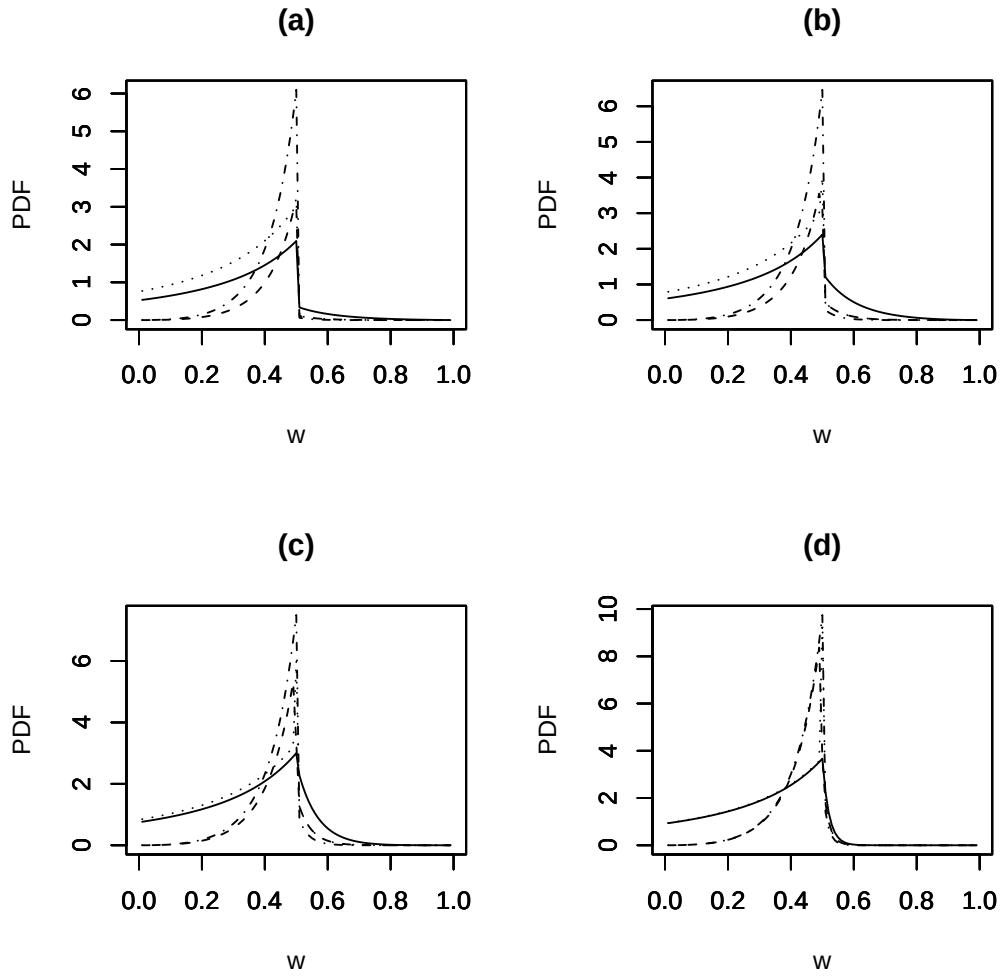

Figure 3: *Plots of the pdf of (12) for* $\beta = 1$ *and (a):* $\lambda_0 = 0.1$*; (b):* $\lambda_0 = 0.5$*; (c):* $\lambda_0 = 2$*; and, (d):* $\lambda_0 = 10$*. The four curves in each plot are: the solid curve (*$\lambda_1 = 1$*,* $\lambda_2 = 1$*), the curve of lines (*$\lambda_1 = 1$*,* $\lambda_2 = 3$*), the curve of dots (*$\lambda_1 = 3$*,* $\lambda_2 = 1$*), and the curve of lines and dots (*$\lambda_1 = 3$*,* $\lambda_2 = 3$*).*

3 Moments

Here, we derive the moments of $R = X + Y$, $P = XY$ and $W = X/(X + Y)$ when X and Y are distributed according to (1)–(3). We need the following lemma.

Lemma 2 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned} E(X^m Y^n) &= \frac{\beta^{m+n}\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{(m-\lambda_0-\lambda_1)(m+n-\lambda_0-\lambda_1-\lambda_2)(\lambda_0+\lambda_1+\lambda_2)} \\ &\quad + \frac{\beta^{m+n}\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{(n-\lambda_0-\lambda_2)(m+n-\lambda_0-\lambda_1-\lambda_2)(\lambda_0+\lambda_1+\lambda_2)} \\ &\quad + \frac{\beta^{m+n}\lambda_0}{m+n+\lambda_0+\lambda_1+\lambda_2} \end{aligned}$$

for $m \geq 1$ and $n \geq 1$.

Proof. One can write

$$\begin{aligned} E(X^m Y^n) &= \frac{\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}\int_\beta^\infty\int_y^\infty x^{m-1-\lambda_0-\lambda_1}y^{n-1-\lambda_2}dxdy \\ &\quad + \lambda_0\beta^{\lambda_0+\lambda_1+\lambda_2}\int_\beta^\infty x^{m+n-1-\lambda_0-\lambda_1-\lambda_2}dx \\ &\quad + \frac{\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{\lambda_0+\lambda_1+\lambda_2}\beta^{\lambda_0+\lambda_1+\lambda_2}\int_\beta^\infty\int_x^\infty y^{-(1+\lambda_0+\lambda_2)}x^{-(1+\lambda_1)}dydx. \end{aligned}$$

The result of the theorem follows by elementary integration of the above integrals. ■

The moments of $R = X + Y$ and $P = XY$ are now simple consequences of this lemma as illustrated in Theorems 4 and 5. The moments of $W = X/(X + Y)$ require a separate treatment as shown by Theorem 6.

Theorem 4 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned} E(R^n) &= \frac{2^n\beta^n\lambda_0}{n+\lambda_0+\lambda_1+\lambda_2} + \sum_{k=0}^{n}\binom{n}{k}\left[\frac{\beta^n\lambda_2(\lambda_0+\lambda_1)(\lambda_1+\lambda_2)}{(n-k-\lambda_0-\lambda_1)(n-\lambda_0-\lambda_1-\lambda_2)(\lambda_0+\lambda_1+\lambda_2)}\right. \\ &\quad \left. + \frac{\beta^n\lambda_1(\lambda_0+\lambda_2)(\lambda_1+\lambda_2)}{(k-\lambda_0-\lambda_2)(n-\lambda_0-\lambda_1-\lambda_2)(\lambda_0+\lambda_1+\lambda_2)}\right] \end{aligned}$$

for $n \geq 1$.

Proof. the result follows by writing

$$E\left((X+Y)^n\right) = \sum_{k=0}^{n}\binom{n}{k}E\left(X^{n-k}Y^k\right)$$

and applying Lemma 2 to each expectation in the sum. ■

Theorem 5 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned}E\left(P^n\right) &= \frac{\beta^{2n}\lambda_2\left(\lambda_0+\lambda_1\right)\left(\lambda_1+\lambda_2\right)}{\left(n-\lambda_0-\lambda_1\right)\left(2n-\lambda_0-\lambda_1-\lambda_2\right)\left(\lambda_0+\lambda_1+\lambda_2\right)}\\ &+\frac{\beta^{2n}\lambda_1\left(\lambda_0+\lambda_2\right)\left(\lambda_1+\lambda_2\right)}{\left(n-\lambda_0-\lambda_2\right)\left(2n-\lambda_0-\lambda_1-\lambda_2\right)\left(\lambda_0+\lambda_1+\lambda_2\right)}\\ &+\frac{\beta^{2n}\lambda_0}{2n+\lambda_0+\lambda_1+\lambda_2}\end{aligned}$$

for $n \geq 1$.

Proof. follows by writing $E(P^n) = E(X^nY^n)$ and applying Lemma 2 with $m = n$. ■

Theorem 6 *If X and Y are jointly distributed according to (1)–(3) then*

$$\begin{aligned}E\left(W^n\right) &= \frac{2^{1-n}\lambda_0}{\lambda_0+\lambda_1+\lambda_2}+\frac{\lambda_2\left(\lambda_0+\lambda_1\right)\left(\lambda_1+\lambda_2\right)}{\left(\lambda_0+\lambda_1+\lambda_2\right)^2}B_{1/2}\left(\lambda_0+\lambda_1, n-\lambda_0-\lambda_1\right)\\ &+\frac{\lambda_1\left(\lambda_0+\lambda_2\right)\left(\lambda_1+\lambda_2\right)}{\left(\lambda_0+\lambda_1+\lambda_2\right)^2}B_{1/2}\left(n+\lambda_0+\lambda_2, -\lambda_0-\lambda_2\right) \qquad (14)\end{aligned}$$

for $n \geq 1$.

Proof. Using (12), one can write

$$\begin{aligned}E\left(W^n\right) &= \frac{\lambda_2\left(\lambda_0+\lambda_1\right)\left(\lambda_1+\lambda_2\right)}{\left(\lambda_0+\lambda_1+\lambda_2\right)^2}\int_{1/2}^{1}\frac{w^n(1-w)^{\lambda_0+\lambda_1-1}}{w^{\lambda_0+\lambda_1+1}}dw+\frac{2^{1-n}\lambda_0}{\lambda_0+\lambda_1+\lambda_2}\\ &+\frac{\lambda_1\left(\lambda_0+\lambda_2\right)\left(\lambda_1+\lambda_2\right)}{\left(\lambda_0+\lambda_1+\lambda_2\right)^2}\int_{0}^{1/2}\frac{w^{n+\lambda_0+\lambda_2-1}}{(1-w)^{\lambda_0+\lambda_2+1}}dw.\end{aligned}$$

The result of the theorem follows by the definition of the incomplete beta function. ■

Using special properties of the incomplete beta function, one can derive elementary forms of (14). This is shown in the corollaries below.

Corollary 4 *If X and Y are jointly distributed according to (1)–(3) and if $\lambda_0 + \lambda_1 \geq 1$ is an integer then $E(W^n)$ is given by (14) with*

$$B_{1/2}(\lambda_0 + \lambda_1, n - \lambda_0 - \lambda_1) = 1 - 2^{\lambda_0+\lambda_1-n} \sum_{k=1}^{\lambda_0+\lambda_1} \frac{\Gamma(n - \lambda_0 - \lambda_1 + k - 1)}{\Gamma(n - \lambda_0 - \lambda_1)\Gamma(k)} 2^{1-k}.$$

Corollary 5 *If X and Y are jointly distributed according to (1)–(3) and if $\lambda_0 + \lambda_2 \geq 1$ is an integer then $E(W^n)$ is given by (14) with*

$$B_{1/2}(n + \lambda_0 + \lambda_2, -\lambda_0 - \lambda_2) = 1 - 2^{\lambda_0+\lambda_2} \sum_{k=1}^{n+\lambda_0+\lambda_2} \frac{\Gamma(n - \lambda_0 - \lambda_2 - 1)}{\Gamma(-\lambda_0 - \lambda_2)\Gamma(k)} 2^{1-k}.$$

Percentage points for $R = X + Y$

λ_0	λ_1	λ_2	α						
			0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	1	1	9.62379	16.49151	29.78508	70.40482	137.2642	662.6351	1259.997
1	1	2	6.38258	9.402676	14.83461	30.03895	55.16546	260.3903	511.3611
1	1	3	5.419237	7.756518	12.03189	24.47247	45.52252	212.9589	429.7302
1	1	4	5.028103	7.113937	10.81814	21.20653	38.43352	183.8694	342.4035
1	2	1	6.390126	9.439726	14.88197	30.05685	54.93929	265.512	551.438
1	2	2	4.709026	5.991161	7.717563	11.07508	14.77118	30.48835	43.03218
1	2	3	4.108536	5.012974	6.194804	8.438637	10.91330	21.23383	29.51412
1	2	4	3.806872	4.581755	5.616983	7.630785	9.834232	19.15476	25.87484
1	3	1	5.435317	7.754701	11.96454	23.94088	44.06230	218.9689	440.4970
1	3	2	4.104989	5.016544	6.21157	8.413445	10.88446	21.34119	29.27734
1	3	3	3.617398	4.213916	4.936576	6.121482	7.261394	11.20831	13.68865
1	3	4	3.369725	3.85984	4.433214	5.385524	6.294604	9.31356	11.19996
1	4	1	5.011635	7.096623	10.78617	21.23060	38.18032	175.2684	338.8366
1	4	2	3.80762	4.58071	5.625067	7.593149	9.829113	18.83798	25.31333
1	4	3	3.372456	3.862036	4.436062	5.373236	6.295974	9.400779	11.25767
1	4	4	3.149176	3.525237	3.951599	4.617037	5.197676	7.02943	7.937064
2	1	1	6.908587	11.80303	21.87108	52.17421	103.0996	499.1869	997.795
2	1	2	5.026371	7.20023	11.28552	23.11933	42.80327	195.6307	377.0428
2	1	3	4.290314	5.792941	8.742232	18.13929	34.57562	164.6154	310.1243
2	1	4	3.932082	5.174204	7.661537	16.03917	29.84287	150.3417	292.0324
2	2	1	5.022213	7.195747	11.29309	23.38030	43.37083	202.7843	408.1873
2	2	2	4.13171	5.197745	6.670142	9.64867	12.99950	27.08907	37.89176
2	2	3	3.688318	4.443407	5.44835	7.400021	9.607725	18.71532	25.11885
2	2	4	3.446081	4.072247	4.913453	6.573777	8.551287	17.07842	23.09799
2	3	1	4.29382	5.793729	8.750932	18.24647	35.13249	172.9237	343.8436
2	3	2	3.695045	4.449435	5.443715	7.428022	9.612759	19.03360	25.71027
2	3	3	3.365524	3.891746	4.532711	5.631077	6.69489	10.43221	12.59107
2	3	4	3.170085	3.599843	4.105821	4.970653	5.800506	8.645942	10.33633
2	4	1	3.938950	5.184407	7.665587	15.86660	30.17476	143.1466	284.7486

2	4	2	3.444092	4.074746	4.925434	6.631013	8.602288	17.32115	23.43268
2	4	3	3.174564	3.601501	4.110391	4.961649	5.79828	8.581511	10.37010
2	4	4	3.003146	3.33961	3.724340	4.33036	4.89138	6.589682	7.558714
3	1	1	5.556553	9.369284	17.22973	41.75966	82.96045	402.7553	777.0871
3	1	2	4.356332	6.135416	9.57303	19.70144	36.90099	170.3009	324.3305
3	1	3	3.798096	5.011188	7.527998	15.77983	29.90835	145.9120	283.4977
3	1	4	3.499727	4.460710	6.534267	13.78972	26.22162	132.2372	284.8490
3	2	1	4.346251	6.141804	9.52928	19.51615	35.56617	174.0592	359.7555
3	2	2	3.795461	4.756487	6.126913	8.92387	12.12327	25.40350	35.55154
3	2	3	3.446532	4.123285	5.055806	6.904248	9.032536	17.89067	24.55036
3	2	4	3.234209	3.78897	4.564601	6.174418	8.080741	16.08926	22.26428
3	3	1	3.797431	5.014095	7.522061	15.69015	29.62894	147.5539	316.1749
3	3	2	3.445716	4.130013	5.054499	6.874383	9.011033	17.91510	24.94480
3	3	3	3.199305	3.686841	4.28561	5.317371	6.38796	10.10062	12.19057
3	3	4	3.035728	3.430755	3.906324	4.714228	5.508655	8.228773	9.973009
3	4	1	3.491053	4.432245	6.497168	13.59319	25.43914	123.7052	239.6344
3	4	2	3.234879	3.797424	4.571231	6.147995	8.042847	16.38848	22.63103
3	4	3	3.036627	3.430917	3.905991	4.727645	5.506118	8.309395	9.892161
3	4	4	2.905312	3.219497	3.58608	4.175939	4.700929	6.402495	7.481372
4	1	1	4.751449	8.002954	14.64670	34.75727	67.59802	329.042	666.0516
4	1	2	3.928277	5.493376	8.457637	17.04472	31.56688	148.9422	289.8442
4	1	3	3.496069	4.542001	6.688109	13.90893	26.0087	119.8621	241.3712
4	1	4	3.244412	4.050198	5.87665	12.53929	24.00743	113.0687	238.5718
4	2	1	3.941387	5.541231	8.624192	17.4132	32.66607	151.8941	300.2353
4	2	2	3.552877	4.440468	5.732454	8.303123	11.28017	24.11745	34.02633
4	2	3	3.273833	3.909206	4.794988	6.536048	8.465026	16.80705	23.01364
4	2	4	3.089486	3.605092	4.332355	5.847722	7.649926	15.64277	21.51948
4	3	1	3.490893	4.541283	6.738373	14.01726	26.40212	125.2505	243.3163
4	3	2	3.269694	3.89497	4.77076	6.543371	8.531305	17.14365	23.55384
4	3	3	3.077487	3.536144	4.113783	5.12258	6.14586	9.683466	11.84106
4	3	4	2.937028	3.308072	3.77329	4.567861	5.368209	8.138209	9.9069
4	4	1	3.246128	4.043495	5.834828	12.46214	23.63761	113.7882	219.6971
4	4	2	3.08685	3.599695	4.323378	5.837564	7.619	15.45620	21.34394
4	4	3	2.935896	3.307082	3.767617	4.569923	5.372604	8.09562	9.654783
4	4	4	2.826395	3.128217	3.478607	4.048215	4.57136	6.206746	7.152812

4 Percentiles

In this section, we provide extensive tabulations of the percentiles of the distribution of R (percentiles for P and W are not given since their pdfs are elementary). These percentiles are computed numerically by solving the equation

$$\int_0^{r_\alpha} f_R(r)dr = \alpha,$$

where $f_R(r)$ is given by (4). Evidently, this involves computation of the hypergeometric functions and routines for this are widely available. We used the function hypergeom $(\cdot)$ in the algebraic manipulation package, MAPLE. The percentiles are given for $\alpha = 0.90$, 0.95, 0.975, 0.99, 0.995, 0.999, 0.9995, $\beta = 1$, $\lambda_0 = 1, 2, 3, 4$, $\lambda_1 = 1, 2, 3, 4$ and $\lambda_2 = 1, 2, 3, 4$.

Similar tabulations could be easily derived for other values of λ_0, λ_1 and λ_2. We hope these numbers will be of use to the practitioners of the bivariate Pareto distribution (see Section 1).

Acknowledgments

The authors would like to thank the referee and the editor for carefully reading the paper and for their great help in improving the paper.

References

Chapman, D.G. (1950). Some two-sample tests. *Annals of Mathematical Statistics*, 21, 601-606.

Chaubey, Y.P. and Talukder, A.B.M. Nur Enayet (1983). Exact moments of a ratio of two positive quadratic forms in normal variables. *Communications in Statistics-Theory and Methods*, 12, 675-679.

Dobson, A.J., Kulasmaa, K. and Scherer, J. (1991). Confidence intervals for weighted sums of Poisson parameters. *Statistics in Medicine*, 10, 457-462.

Gradshteyn, I.S. and Ryzhik, I.M. (2000). *Table of Integrals, Series, and Products* (sixth edition). San Diego: Academic Press.

Grice, J.V. and Bain, L.J. (1980). Inferences concerning the mean of the gamma distribution. *Journal of the American Statistical Association*, 75, 929-933.

Jackson, O.A.Y. (1969). Fitting a gamma or log-normal distribution to fibre-diameter measurements on wool tops. *Applied Statistics*, 18, 70-75.

Kamgar-Parsi, B., Kamgar-Parsi, B. and Brosh, M. (1995). Distribution and moments of weighted sum of uniform random variables with applications in reducing Monte Carlo simulations. *Journal of Statistical Computation and Simulation*, 52, 399-414.

Kimball, C.V. and Scheibner, D.J. (1998). Error bars for sonic slowness measurements. *Geophysics*, 63, 345-353.

Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000). *Continuous Multivariate Distributions, volume 1: Models and Applications* (second edition). New York: John Wiley and Sons.

Lai, T.L. (1974). Control charts based on weighted sums. *Annals of Statistics*, 2, 134-147.

Linhart, H. (1965). Approximate confidence limits for the coefficient of variation of gamma distributions. *Biometrics*, 21, 733-738.

Lowan, A.N. and Laderman, J. (1939). On the distribution of errors in *n*th tabular differences. *Annals of Mathematical Statistics*, 10, 360-364.

Malik, H.J. (1970). The distribution of the product of two noncentral beta variates. *Naval Research Logistics Quarterly*, 17, 327-330.

Mikhail, N.N. and Tracy, D.S. (1975). The exact non-null distribution of Wilk's Λ criterion in the bivariate collinear case. *Canadian Mathematical Bulletin*, 17, 757-758.

Monti, K.L. and Sen, P.K. (1976). The locally optimal combination of independent test statistics. *Journal of the American Statistical Association*, 71, 903-911.

Muliere, P. and Scarsini, M. (1987). Characterization of a Marshall-Olkin type class of distributions. *Annals of the Institute of Statistical Mathematics*, 39, 429-441.

von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics*, 12, 367-395.

Pham-Gia, T. (2000). Distributions of the ratios of independent beta variables and applications. *Communications in Statistics-Theory and Methods*, 29, 2693-2715.

Pham-Gia, T. and Turkkan, N. (1993). Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*, 22, 1755-1771.

Pham-Gia, T. and Turkkan, N. (1994). Reliability of a standby system with beta component lifelength. *IEEE Transactions on Reliability*, 71-75.

Pham-Gia, T. and Turkkan, N. (1998). Distribution of the linear combination of two general beta variables and applications. *Communications in Statistics-Theory and Methods*, 27, 1851-1869.

Pham-Gia, T. and Turkkan, N. (2002). The product and quotient of general beta distributions. *Statistical Papers*, 43, 537-550.

Provost, S.B. and Rudiuk, E.M. (1994). The exact density function of the ratio of two dependent linear combinations of chi-square variables. *Annals of the Institute of Statistical Mathematics*, 46, 557-571.

Prudnikov, A.P., Brychkov, Y.A. and Marichev, O.I. (1986). *Integrals and Series* (volumes 1, 2 and 3). Amsterdam: Gordon and Breach Science Publishers.

Rousseau, B. and Ennis, D.M. (2001). A Thurstonian model for the dual pair (4IAX) discrimination method. *Perception & Psychophysics*, 63, 1083-1090.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, 243-258.

Toyoda, T. and Ohtani, K. (1986). Testing equality between sets of coefficients after a preliminary test for equality of disturbance variances in two linear regressions. *Journal of Econometrics*, 31, 67-80.

Witkovský, V. (2001). Computing the distribution of a linear combination of inverted gamma variables. *Kybernetika*, 37, 79-90.

Yatchew, A.J. (1986). Multivariate distributions involving ratios of normal variables. *Communications in Statistics-Theory and Methods*, 15, 1905-1926.

SORT 29 (2) July-December 2005, 201-216

On the role played by the fixed bandwidth in the Bickel-Rosenblatt goodness-of-fit test

Carlos Tenreiro

Universidade de Coimbra

Abstract

For the Bickel-Rosenblatt goodness-of-fit test with fixed bandwidth studied by Fan (1998) we derive its Bahadur exact slopes in a neighbourhood of a simple hypothesis $f = f_0$ and we use them to get a better understanding on the role played by the smoothing parameter in the detection of departures from the null hypothesis. When f_0 is a univariate normal distribution and we take for kernel the standard normal density function, we compute these slopes for a set of Edgeworth alternatives which give us a description of the test properties in terms of the bandwidth h. A simulation study is presented which indicates that finite sample properties are in good accordance with the theoretical properties based on Bahadur local efficiency. Comparisons with the quadratic classical EDF tests lead us to recommend a test based on a combination of bandwidths in alternative to Anderson-Darling or Cramér-von Mises tests.

MSC: 62G10, 62G20

Keywords: goodness-of-fit test, kernel density estimator, Bahadur efficiency.

1 Introduction

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent and identically distributed d-dimensional random vectors with unknown density function f. As it has been shown by Bickel and Rosenblatt (1973), a test of the simple hypothesis $H_0 : f = f_0$ against the alternative $H_a : f \neq f_0$, where f_0 is a fixed density function on $\mathbb{R}^d$, can be based on the L_2 distance between the kernel density estimator of f introduced by Rosenblatt (1956)

Address for correspondence: Carlos Tenreiro. Departamento de Matemática, Universidade de Coimbra, Apartado 3008, 3001-454 Coimbra, Portugal. Phone: (351) 239 791 155. Fax: (351) 239 832 568. E-mail: tenreiro@mat.uc.pt

Received: July 2004

Accepted: April 2005

and Parzen (1962), and its mathematical expectation under the null hypothesis (see also Fan (1994) and Gouriéroux and Tenreiro (2001)):

$$I_n^2(h_n) = n \int \{ f_n(x) - \mathrm{E}_0 f_n(x) \}^2 dx, \tag{1}$$

where, for $x \in \mathbb{R}^d$,

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(x - X_i),$$

$K_{h_n} = K(\cdot/h_n)/h_n^d$ with K a kernel, that is, a bounded and integrable function on $\mathbb{R}^d$, and (h_n) is a sequence of strictly positive real numbers converging to zero, when n goes to infinity (bandwidth). The Bickel-Rosenblatt test is asymptotically consistent and has a normal asymptotic distribution under the null hypothesis.

Following an idea of Anderson, Hall and Titterington (1994) that have used kernel density estimators with fixed bandwidth for testing the equality of two multivariate probability density functions, Fan (1998) uses the statistic (1) with a constant bandwidth for testing the composite hypothesis that f is a member of a general parametric family of density functions. He provides an alternative asymptotic approximation for the finite-sample properties of the Bickel-Rosenblatt test by showing that, for a fixed h, the asymptotic distribution of $I_n^2(h)$ is an infinite sum of weighted χ^2 random variables. Moreover, Fan (1998) proves that $I_n^2(h)$ can be interpreted as a L_2 weighted distance between the empirical characteristic function and the parametric estimate of the characteristic function implied by the null model with weight function $t \to |\phi_K(th)|^2$. In the important case of testing univariate or multivariate normality, and taking for K the standard normal density function, the role played by h in the power performance of the test is assessed in simulation studies by Epps and Pulley (1983), Henze and Zirkler (1990) and Henze and Wagner (1997).

Restricting our attention to the test of a simple hypothesis, the main purpose of this paper is to derive the Bahadur local exact slopes of goodness-of-fit tests based on $I_n^2(h)$, for a fixed $h > 0$, and use them to get a better understanding of the role played by the smoothing parameter in the detection of departures from the null hypothesis. For completeness reasons we give in Section 2 the asymptotic null distribution and the consistency of the test based on kernel density estimators with a fixed bandwidth. Using the integral and quadratic form of $I_n^2(h)$, we derive in Section 3 its Bahadur local exact slopes. They naturally depend on the smoothing parameter, on the kernel, on the null density f_0 and, finally, on the considered departure direction from the null hypothesis. In Section 4, in the particular case of a test for a simple univariate hypothesis of normality and taking for K the standard normal density function, the Bahadur local slopes are numerically evaluated for different values of h for a set of Edgeworth alternatives. These alternatives express departures from the null hypothesis in terms of each one of the first

four moments. The tests based on $I_n^2(h)$ for different values of h are compared with the corresponding ones of the quadratic EDF tests of Anderson-Darling (A^2) and Cramér-von Mises (W^2). The results we obtain suggest that a large bandwidth is adequate for detection of location alternatives whereas a small bandwidth is adequate for detection of alternatives for scale, skewness and kurtosis. A simulation study indicating that finite sample properties of tests I^2 are in good accordance with the theoretical properties based on the Bahadur local slopes is also presented. Moreover, if one does not know much about the unknown density function it suggests that a test based on a combination of bandwidths, that establish a compromise between the two opposite effects that the choice of h has in the detection of location and nonlocation alternatives, is a good practical recommendation in alternative to traditional A^2 or W^2 tests.

For convenience of presentation the proofs of some results in this article are given in Section 5. We denote by $\xrightarrow[n\to+\infty]{a.s.}$ the convergence with probability 1 and by $\xrightarrow[n\to+\infty]{d}$ the convergence in distribution.

2 Asymptotic null distribution and consistency

Consider the following assumptions on K which ensure that $d(f,g) = (\int\{K_h \star f(x) - K_h \star g(x)\}^2 dx)^{1/2}$, where $\star$ denotes the convolution product, is a distance on the set of integrable functions (see Anderson *et al.* (1994)).

Assumptions on K (K)

K is a bounded and integrable function on $\mathbb{R}^d$ with Fourier transform ϕ_K such that $\{t \in \mathbb{R}^d : \phi_K(t) = 0\}$ has Lebesgue measure zero.

In order to derive the asymptotic distribution of $I_n^2(h)$ under H_0 for a fixed $h > 0$, we first note that $I_n^2(h)$ is a V-statistic, that is,

$$I_n^2(h) = \frac{1}{n}\sum_{i,j=1}^{n} Q_h(X_i, X_j), \tag{2}$$

with kernel

$$Q_h(u,v) = \int k(x,u;h)k(x,v;h)dx,$$

where

$$k(x,u;h) = K_h(x-u) - K_h \star f_0(x), \tag{3}$$

for $u, v, x \in \mathbb{R}^d$. From the hypothesis on K, the kernel Q_h is bounded. Therefore the

functions $u \to Q_h(u,u)$ and Q_h are P_0 and $P_0 \otimes P_0$ integrable, respectively, where $P_0 = f_0\lambda$ and λ is the Lebesgue measure in $\mathcal{B}(\mathbb{R}^d)$. Moreover, Q_h is symmetric and degenerate, i.e., $\int Q_h(\cdot,v)dP_0(v) = 0,\ a.e.\ (P_0)$. From Gregory (1977), we know that the asymptotic distribution of $I_n^2(h)$ under H_0 can be characterized in terms of the eigenvalues of the symmetric Hilbert-Schmidt operator A_h defined, for $q \in L_2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_0) =: L_2(P_0)$, by

$$(A_h q)(u) = \int Q_h(u,v)q(v)dP_0(v). \tag{4}$$

In view of the degeneracy property of Q_h, $q_{0,h} = 1$ is an eigenfunction of A_h corresponding to the eigenvalue $\lambda_{0,h} = 0$. Denoting by $\langle\, 1\,\rangle$ the subspace generated by $q_{0,h}$ and $H(P_0) = \left\{g \in L_2(P_0) : \int g dP_0 = 0\right\}$ the tangent space of P_0, we have $L_2(P_0) = \langle\, 1\,\rangle \oplus H(P_0)$. The operator A_h is positive definite on $H(P_0)$ as follows from the integral form (3) of Q_h and assumption (K). In fact, if $\langle A_h q, q\rangle = 0$, for some $q \in H(P_0)$, where $\langle\cdot,\cdot\rangle$ denotes the usual inner product in $L_2(P_0)$, we have

$$\begin{aligned} 0 &= \int q(u)k(\cdot,u;h)dP_0(u) \\ &= K \star (q f_0)(\cdot),\ a.e.\ (\lambda), \end{aligned}$$

yielding $\phi_K(t)\phi_{qf_0}(t) = 0$, for all $t \in \mathbb{R}^d$.

From assumption (K) and the continuity of the Fourier transform, we deduce that $\phi_{qf_0}(t) = 0$, $t \in \mathbb{R}^d$, i.e., $q = 0,\ a.e.\ (P_0)$.

Finally, using the the infinite-dimensionality of $H(P_0)$ and the positivity of A_h on $H(P_0)$ we can conclude that A_h has a countable infinite collection $\{\lambda_{k,h}, k \in \mathbb{N}\}$ of strictly positive eigenvalues (see Dunford and Schwartz (1963), Corollary X.4.5).

The following result follows from the limit distribution of degenerate V-statistics (cf. Theorem 4.3.2 of Koroljuk and Borovskich (1989)). Remark that the asymptotic distribution presented by Fan (1998) in Theorem 4.2, is not correct. In general the P_0-integrability of $u \to Q_h(u,u)$ is not a sufficient condition for $\sum \lambda_{k,h} < \infty$.

Theorem 1 *If assumption* (K) *is fulfilled then, under* H_0 *we have*

$$I_n^2(h) \xrightarrow[n\to+\infty]{d} I_\infty,$$

with

$$I_\infty = \int Q_h(u,u)dP_0(u) + \sum_{k=1}^{\infty} \lambda_{k,h}(Z_k^2 - 1),$$

where the sequence $(\lambda_{k,h})$, *with* $\lambda_{1,h} \geq \lambda_{2,h} \geq \ldots$ *and* $\lambda_{k,h} \to 0, k \to +\infty$, *is described above and* (Z_k) *are i.i.d. standard normal variables. Moreover, the test* $I^2(h) = (I_n^2(h))$

defined by the critical regions $\{I_n^2(h) > c_\alpha\}$, *where* $\mathrm{P}(I_\infty > c_\alpha) = \alpha$, *is asymptotically of level* α *and consistent to test* H_0 *against* H_a.

Remark 1 If the density f_0 has a compact support S and Q_h is continuous in $S \times S$, from the Mercer's expansion for Q_h (see Dunford and Schwartz (1963), p. 1088) it follows that $\int Q_h(u,u)dP_0(u) = \sum_{k=1}^{\infty} \lambda_{k,h}$ and therefore I_∞ takes the form $I_\infty = \sum_{k=1}^{\infty} \lambda_{k,h} Z_k^2$.

3 Bahadur local efficiency

In order to compare the test $I^2(h)$ with other test procedures, or to compare $I^2(h)$ tests obtained for different values of h, we derive in the following its Bahadur exact slopes $C_{I^2(h)}(f)$, for f in a neighbourhood of f_0. They coincide with the Bahadur approximate slopes (and then with the Bahadur local approximate slopes) derived by Gregory (1980). For the description of Bahadur's concept of efficiency, see Bahadur (1967, 1971) or Nikitin (1995).

Throughout, $|\cdot|_p$ denotes the norm of the Lebesgue space $L_p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda) =: L_p(\lambda)$. The proof of the following result is given in Section 5.

Theorem 2 *We have*

$$C_{I^2(h)}(f) = \frac{b_{I^2(h)}(f)}{\lambda_{1,h}}(1 + o(1)), \ as \ |f - f_0|_1 \to 0,$$

where

$$b_{I^2(h)}(f) = \int \{K_h \star f(x) - K_h \star f_0(x)\}^2 dx,$$

and $\lambda_{1,h}$ *is the largest eigenvalue of the operator* A_h *defined by (4).*

If f_0 belongs to a family of probability density functions of the form $\{f(\cdot;\theta) : \theta \in \Theta\}$, where Θ is a nontrivial closed real interval and $f_0 = f(\cdot;\theta_0)$, for some $\theta_0 \in \Theta$, it is natural to compare a set of competitor tests through its Bahadur local exact slopes when $\theta \to \theta_0$.

Consider the following assumptions on the previous parametric family:

Assumptions on $\{f(\cdot;\theta) : \theta \in \Theta\}$ (P)

For all $x \in \mathbb{R}^d$ the function $\theta \to f(x;\theta)$ is continuously differentiable on Θ, and there exists a neighbourhood $V \subset \Theta$ of θ_0 such that the function $x \to \sup_{\theta \in V} \left|\frac{\partial f}{\partial \theta}(x;\theta)\right|$ is integrable on $\mathbb{R}^d$.

The following result comes easily from Theorem 2, assumption (P) and the dominated convergence theorem.

Corollary 1 *Under assumption* (P), *we have*

$$|f(\cdot;\theta) - f(\cdot;\theta_0)|_1 \to 0, \text{ when } \theta \to \theta_0,$$

and

$$C_{I^2(h)}(f(\cdot;\theta)) = \frac{b^o_{I^2(h)}(f(\cdot;\theta))}{\lambda_{1,h}} (\theta - \theta_0)^2 (1 + o(1)), \text{ when } \theta \to \theta_0$$

where

$$b^o_{I^2(h)}(f(\cdot;\theta)) = \int \left(K_h \star \frac{\partial f}{\partial \theta}(\cdot;\theta_0)(x)\right)^2 dx.$$

Let us denote by $\{q_{k,h}, k \in \mathbb{N}_0\}$ the orthonormal basis for $L_2(P_0)$ corresponding to the infinite collection of eigenvalues of A_h, i.e., for all k and j, $\int Q_h(\cdot, v) q_{k,h}(v) dP_0(v) = \lambda_{k,h} q_{k,h}$, $a.e.$ (P_0) and $\langle q_{k,h}, q_{j,h} \rangle = \delta_{kj}$, where δ_{kj} is the Kronecker symbol. In the following result, we establish a representation for the local slope $C_{I^2(h)}(f(\cdot;\theta))$ when $\theta \to \theta_0$, in terms of the weights $(\lambda_{k,h})$ and the principal components $(q_{k,h})$. It is proven in Section 5.

Corollary 2 *Under assumption* (P), *if* $\frac{\partial \ln f}{\partial \theta}(\cdot;\theta_0) \in L_2(P_0)$, *then*

$$C_{I^2(h)}(f(\cdot;\theta)) = \sum_{k=1}^{\infty} \frac{\lambda_{k,h}}{\lambda_{1,h}} a^2_{k,h} (\theta - \theta_0)^2 (1 + o(1)), \text{ when } \theta \to \theta_0,$$

where, for $k = 1, 2, \ldots,$

$$a_{k,h} = \left\langle q_{k,h}, \frac{\partial \ln f}{\partial \theta}(\cdot;\theta_0) \right\rangle.$$

From the previous representation, in particular from the fact that the weights $(\lambda_{k,h})$ converge to zero, it is clear that only a finite directions of alternatives effectively contribute to $C_{I^2(h)}(f(\cdot;\theta))$. The natural question, that we discuss in the next section for the test of a simple hypothesis of normality, is how rapidly the principal directions loose influence.

4 Testing a simple hypothesis of normality

In this section we consider the test of the simple hypothesis of normality. Without loss of generality we restrict our attention to the test of the hypothesis $H_0 : f = f_{N(0,1)}$ against the alternative hypothesis $H_a : f \neq f_{N(0,1)}$.

4.1 Local alternatives

In order to get a better understanding of the role played by the smoothing parameter in the detection of departures from the null hypothesis, we consider a set of alternatives that satisfy (P) with $f = f_j$ and $\theta_0 = 0$, such that

$$(\mathcal{A}.\, j) \quad \frac{\partial \ln f_j}{\partial \theta}(\cdot; 0) = H_j(\cdot)/\, j!, \tag{5}$$

for $j = 1, \ldots, 4$, where H_j is the jth Hermite polynomial defined by:

$$\begin{aligned} H_1(x) &= x; \\ H_2(x) &= x^2 - 1; \\ H_3(x) &= x^3 - 3x; \\ H_4(x) &= x^4 - 6x^2 + 3. \end{aligned}$$

These alternatives are based on the Edgeworth series for the density and the corresponding value of θ indicate departures from the null hypothesis in the jth moment (about Edgeworth expansion see Hall (1997) and the references therein). Remark that the location alternative $f(\cdot; \theta) = f_{N(\theta,1)}(\cdot)$ and the scale alternative $f(\cdot; \theta) = f_{N(0,1+\theta)}(\cdot)$, when $\theta \to 0$, satisfy $(\mathcal{A}.1)$ and $(\mathcal{A}.2)$, respectively. The alternative $f(\cdot; \theta) = 2\, f_{N(0,1)}(\cdot) F_{N(0,1)}(\theta \cdot)$, when $\theta \downarrow 0$, considered by Durio and Nikitin (2003), satisfies $(\mathcal{A}.1)$ up to the multiplication by a constant. Finally, the skew and kurtosis alternatives considered by Durbin *et al.* (1975) satisfy $(\mathcal{A}.3)$ and $(\mathcal{A}.4)$, respectively.

4.2 The test statistic

From now on we take for K the standard normal density $K = f_{N(0,1)}$. This choice for the kernel was mainly motivated by the fact that the function $b^o_{I^2(h)}(f(\cdot; \theta))$ given in Corollary 1 can be explicitly evaluated for the set of alternatives described above. Also remark that in this case the calculation of $I^2_n(h)$ does not involve any integration. In fact, the kernel Q_h given by (3) takes the form

$$Q_h(u, v) = f_{N(0,2h^2)}(u - v) - f_{N(0,2h^2+1)}(u) - f_{N(0,2h^2+1)}(v) + f_{N(0,2h^2+2)}(0), \tag{6}$$

for $u, v \in \mathbb{R}$ (see Bowman (1992), Bowman and Foster (1993) and Henze and Wagner (1997)).

4.3 Most significant weights

As described in Section 3, the Bahadur local slope of $I_n^2(h)$ depends on the weights $(\lambda_{k,h})$ and on the principal components $(q_{k,h})$. Numerical evaluations of the most significant weights are shown in Table 1 for four values of h. These approximations have been obtained through the projection method. We have considered the restriction, $A_{h|L}$, of the operator A_h defined by (4) with kernel given by (3) to the finite dimension subspace L of $H(P_0)$ given by $L = \{g \in H(P_0) : g = \sum_{i=1}^{n} g(\bar{x}_i)\mathbb{1}_{]x_i,x_{i+1}]}\}$, where $n = 1400$, $x_i = -7 + 0.01(i-1)$ and $\bar{x}_i = (x_i + x_{i+1})/2$, for $i = 1, \ldots, n$. The numerical calculation of the eigenvalues of $A_{h|L}$ have been performed using Lapack routines (cf. Anderson *et al.* (1999)).

From these values and the representation for the Bahadur local slopes given in Corollary 2, we expect that test $I_n^2(h)$ for small values of h could use information contained in others components different from the first ones. However, for moderate or large values of h, it appears that $I_n^2(h)$ might exclusively use information contained in the first components.

Table 1: *Weights for* $I^2(h)$ *with* $K = f_{N(0,1)}$ *and* $f_0 = f_{N(0,1)}$

	$h = 0.05$	$h = 0.2$	$h = 1.0$	$h = 2.0$
$\lambda_{1,h}$	3.59×10^{-1}	2.61×10^{-1}	5.53×10^{-2}	1.28×10^{-2}
$\lambda_{2,h}$	3.54×10^{-1}	2.36×10^{-1}	2.16×10^{-2}	1.93×10^{-3}
$\lambda_{3,h}$	3.11×10^{-1}	1.49×10^{-1}	3.97×10^{-3}	1.31×10^{-4}
$\lambda_{4,h}$	3.06×10^{-1}	1.29×10^{-1}	1.32×10^{-3}	1.65×10^{-5}
$\lambda_{5,h}$	2.70×10^{-1}	8.46×10^{-2}	2.85×10^{-4}	1.33×10^{-6}
$\lambda_{6,h}$	2.64×10^{-1}	7.15×10^{-2}	8.90×10^{-5}	1.57×10^{-7}
$\lambda_{7,h}$	2.35×10^{-1}	4.82×10^{-2}	2.05×10^{-5}	1.36×10^{-8}
$\lambda_{8,h}$	2.29×10^{-1}	4.00×10^{-2}	6.17×10^{-6}	1.55×10^{-9}
$\lambda_{9,h}$	2.04×10^{-1}	2.74×10^{-2}	1.47×10^{-6}	1.39×10^{-10}
$\lambda_{10,h}$	1.98×10^{-1}	2.24×10^{-2}	4.33×10^{-7}	1.54×10^{-11}
$\lambda_{11,h}$	1.77×10^{-1}	1.56×10^{-2}	1.06×10^{-7}	1.42×10^{-12}
$\lambda_{12,h}$	1.71×10^{-1}	1.26×10^{-2}	3.06×10^{-8}	1.54×10^{-13}

4.4 Bahadur local exact slopes

Similarly to the quadratic EDF tests of Anderson-Darling (A^2) and Cramér-von Mises (W^2) (see Nikitin (1995), p. 73–81), for each one of the alternatives (5) the Bahadur local exact slopes of the tests based on $I^2(h)$ take the form $\theta^2(1+o(1))$, up to the multiplication by a constant, when $\theta \to 0$. Therefore, for the comparison of such tests it is sufficient to compare the coefficients of θ^2. They are usually called local indices and are plotted in Figure 1 for $h \in [0.01, 3]$ and Q_h given by (6). We also plot the local indices for A^2 and W^2 tests.

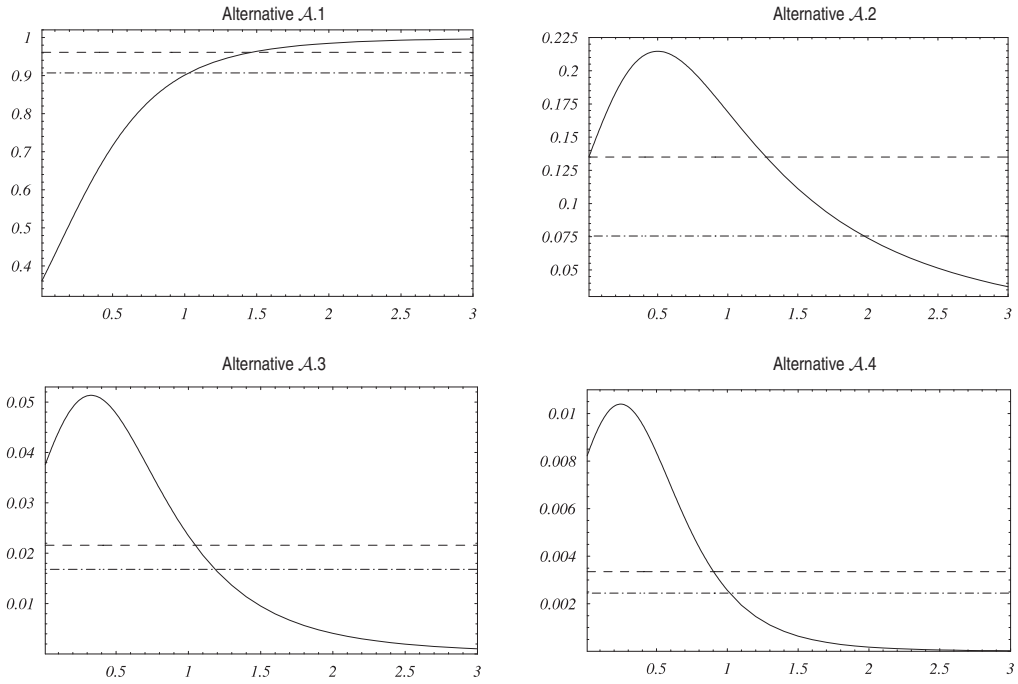

Figure 1: *Local indices for: $I^2(h)$ – solid line; A^2– broken line; W^2– broken and dotted line*

It is clear from Figure 1 that a large bandwidth leads to a strong predominance of the first principal component whereas a small bandwidth leads to a test that uses the information contained in the other components. For the location alternative, we note that the local indices obtained numerically for large values of h are close to one which is, from Bahadur-Raghavachari inequality (see Nikitin (1995), Theorem 1.2.3), the optimal Bahadur local efficiency for this alternative. However, the gain of efficiency in the location alternative by taking a large value of h implies a severe loss of efficiency in the other moment alternatives.

4.5 Combining bandwidths effects

A compromise between the two opposite effects that the choice of h has in the detection of location and nonlocation alternatives can be achieved by considering a test based on a combination of bandwidths, i.e., a test based on the statistic (2) with $K_h = (1-\alpha)K_{h_1} + \alpha K_{h_2}$, where h_1 (small bandwidth) and h_2 (large bandwidth) are two fixed bandwidths, and $\alpha \in [0,1]$.

Denoting by $I^2(\alpha; h_1, h_2)$ such test and assuming that $\{f(\cdot;\theta) : \theta \in \Theta\}$ satisfies (P), we have

$$
\begin{aligned}
& b^o_{I^2(\alpha;h_1,h_2)}(f(\cdot;\theta)) \\
&= (1-\alpha)^2\, b^o_{I^2(h_1)}(f(\cdot;\theta)) + \alpha^2\, b^o_{I^2(h_2)}(f(\cdot;\theta)) \\
&\quad + 2\alpha(1-\alpha)\int K_{h_1} \star \frac{\partial f}{\partial \theta}(\cdot;\theta_0)(x) K_{h_2} \star \frac{\partial f}{\partial \theta}(\cdot;\theta_0)(x)\, dx.
\end{aligned}
$$

For alternatives (5) we plot in Figure 2 the local indices for the combined test $I^2(\alpha; 0.3, 2.0)$ for $\alpha \in [0.7, 1]$. Notice that $h_1 = 0.3$ and $h_2 = 2.0$ are appropriated bandwidths for the detection of nonlocation and location alternatives, respectively (see Figure 1). It follows that the test $I^2(0.8; 0.3, 2.0)$ is superior to W^2 for all the considered alternatives ($\mathcal{A}$.1-4), and is superior to A^2 for alternatives ($\mathcal{A}$.2-4). Remark that this behaviour cannot be achieved by a test $I^2(h)$ for a fixed h (see Figure 1). The test $I^2(0.9; 0.3, 2.0)$ is superior to A^2 for alternative ($\mathcal{A}$.1) but is inferior to A^2 for alternatives ($\mathcal{A}$.2-4). However, the loss of efficiency for these last alternatives is not as significant as if we take a test $I^2(h)$ with a relative local Bahadur efficiency close to one with respect to $I^2(0.9; 0.3, 2.0)$ for alternative ($\mathcal{A}$.1).

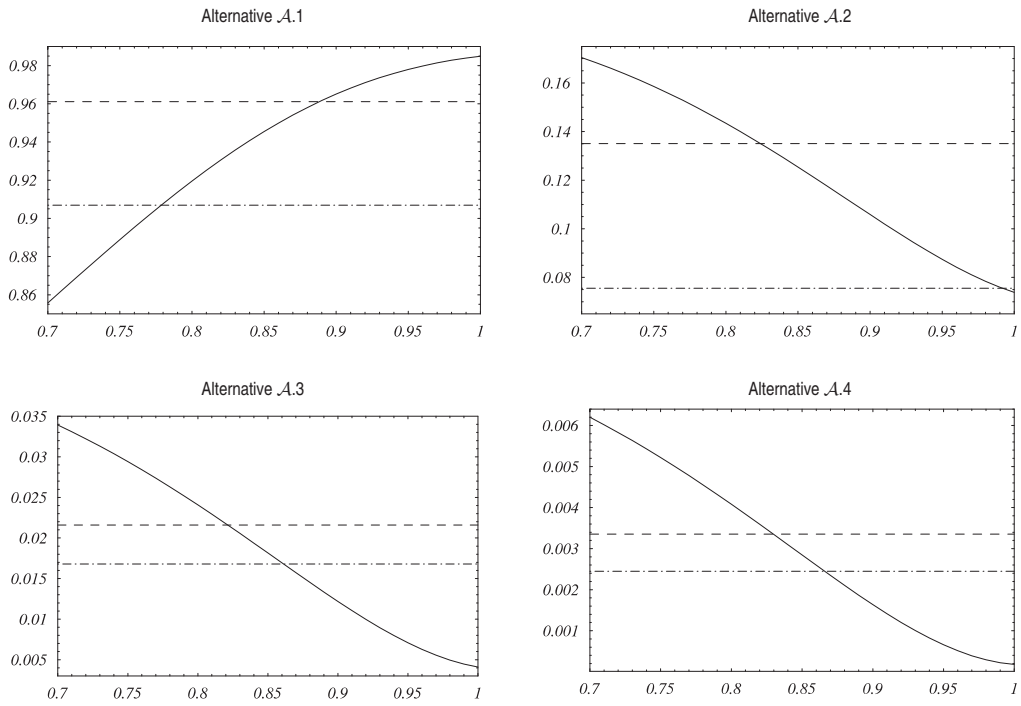

Figure 2: *Local indices for:* $I^2(\alpha; 0.3, 2.0)$ *– solid line;* A^2*– broken line;* W^2*– broken and dotted line*

4.6 Some simulation results

The main purpose of this section is to know if the finite sample properties of the I^2 tests for fixed alternatives are in accordance with the theoretical properties based on Bahadur local efficiency. For that reason we present a simulation study including the tests $I^2(0.3)$ (small bandwidth), $I^2(0.8)$ (medium bandwidth) and $I^2(2.0)$ (large bandwidth) based on fixed bandwidths, and the test $I^2(c) := I^2(0.8; 0.3, 2.0)$ based on a combination of bandwidths. Moreover, as before, the EDF tests A^2 and W^2 will be use for comparison.

To examine the performance of these tests when the null hypothesis is false, we consider three normal alternatives and four nonnormal alternative distribution shapes shown in Figure 3. The nonnormal distributions are members of the generalized lambda family discussed in Ramberg and Schmeiser (1974). The distributions of this family are easily generated because they are defined in terms of the inverses of the cumulative distribution functions: $F^{-1}(u) = \lambda_1 + (u^{\lambda_3} - (1-u)^{\lambda_4})/\lambda_2$, for $0 < u < 1$. The parameters defining the distributions used in the study and the associated mean (μ), variance (σ^2), skewness (α_3) and kurtosis (α_4) values, are given in Table 2. Some of these distributions are used in Fan (1994) to examine the performance of the Bickel-Rosenblatt test with a bandwidth converging to zero as n tends to infinity.

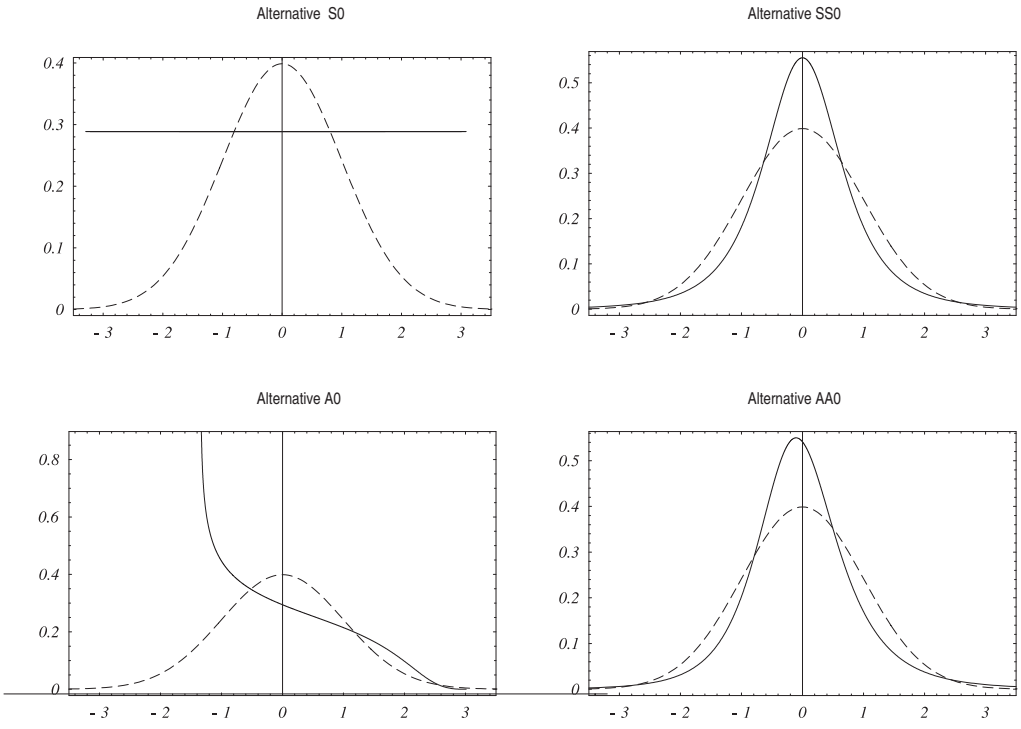

Figure 3: *Distribution shapes considered in the simulation study: Alternative density – solid line; Standard Normal density – broken line*

Table 2: *Distributions used in the simulation study*

Case	μ	σ^2	α_3	α_4	λ_1	λ_2	λ_3	λ_4
Normal distributions								
N1	0.4	1	”	”	—	—	—	—
N2	0	0.36	”	”	—	—	—	—
N3	0.4	0.36	”	”	—	—	—	—
Nonnormal distributions								
Symmetric distributions								
S0	0	1	0	1.8000	0	0.577350	1	1
S1	0.4	1	”	”	0.4	”	”	”
S2	0	0.36	”	”	0	0.962250	”	”
S3	0.4	0.36	”	”	0.4	”	”	”
SS0	0	1	0	11.6136	0	−0.397012	−0.16	−0.16
SS1	0.4	1	”	”	0.4	”	”	”
SS2	0	0.36	”	”	0	−0.663187	”	”
SS3	0.4	0.36	”	”	0.4	”	”	”
Asymmetric distributions								
A0	0	1	0.5129	2.2212	0.835034	0.459063	1.4	0.25
A1	0.4	1	”	”	1.235034	”	”	”
A2	0	0.36	”	”	0.501020	0.765105	”	”
A3	0.4	0.36	”	”	0.901020	”	”	”
AA0	0	1	0.7588	11.4308	−0.116734	−0.351663	−0.13	−0.16
AA1	0.4	1	”	”	−0.283266	”	”	”
AA2	0	0.36	”	”	−0.070040	−0.586106	”	”
AA3	0.4	0.36	”	”	0.329960	”	”	”

In Table 3 we present the Monte-Carlo empirical power results for the previous tests drawn from the considered alternatives. These results are based on 10^4 Monte-Carlo samples of different sizes for a significance level of 0.05. For the evaluation of the critical values of the I^2 tests we have used 10^4 replications. In applying the tests A^2 and W^2 we have followed Stephens (1986).

From Table 3, and Figures 1 and 2, we conclude that the theoretical results based on Bahadur local efficiency are in good accordance with empirical ones. The theoretical properties of I^2 tests are well transferred to finite sample situations.

In practice, the choice among the considered tests depends on the available information about the alternative to the null hypothesis. For alternatives f whose mean and variance satisfy $\mu_f \neq 0$ and $\sigma_f^2 = 1$ (Type I alternatives), A^2 is in general the best test, and each one of the tests $I^2(0.8)$, $I^2(2.0)$ or $I^2(c)$ is better than $I^2(0.3)$ test. For alternatives f satisfying $\mu_f = 0$ or $\sigma_f^2 \neq 1$ (Type II alternatives), $I^2(0.3)$ is globally the best test. Moreover, for these alternatives each one of the tests $I^2(0.3)$, $I^2(0.8)$ or $I^2(c)$ is better or significantly better than A^2 or W^2 tests.

Table 3: *Empirical power at level* 0.05 *for different values of n*

Case	n	$I^2(0.3)$	$I^2(0.8)$	$I^2(2.0)$	$I^2(c)$	A^2	W^2
N1	20	.210	.347	.423	.376	.404	.384
	50	.499	.722	.807	.759	.785	.761
N2	20	.432	.238	.005	.171	.080	.080
	50	.928	.930	.026	.811	.719	.536
N3	10	.362	.357	.183	.319	.245	.335
	20	.745	.805	.551	.750	.702	.743
S0	50	.452	.152	.060	.208	.148	.119
	100	.780	.277	.068	.450	.292	.211
	200	.982	.539	.067	.826	.666	.457
S1	20	.331	.359	.421	.413	.434	.363
	50	.712	.712	.784	.804	.816	.720
S2	20	.272	.138	.007	.087	.048	.052
	50	.986	.917	.030	.760	.636	.244
S3	10	.259	.299	.183	.260	.221	.276
	20	.671	.742	.531	.669	.633	.651
SS0	50	.316	.130	.032	.128	.094	.071
	100	.621	.326	.035	.335	.234	.170
	200	.915	.704	.041	.715	.613	.476
SS1	20	.391	.458	.455	.484	.494	.500
	50	.820	.885	.856	.896	.894	.898
SS2	20	.789	.463	.003	.461	.252	.246
	50	.998	.987	.027	.984	.955	.925
SS3	10	.596	.480	.167	.477	.339	.460
	20	.934	.903	.581	.899	.844	.881
A0	50	.581	.225	.069	.289	.207	.176
	100	.895	.442	.069	.613	.445	.323
	200	.998	.780	.082	.957	.905	.686
A1	20	.217	.223	.368	.315	.358	.251
	50	.588	.516	.740	.763	.790	.592
A2	20	.427	.187	.006	.156	.085	.086
	50	.995	.927	.030	.904	.790	.445
A3	10	.266	.241	.162	.210	.171	.229
	20	.870	.824	.519	.861	.739	.772
AA0	50	.320	.149	.038	.140	.107	.087
	100	.620	.350	.044	.346	.251	.184
	200	.909	.711	.054	.705	.619	.477
AA1	20	.315	.365	.394	.401	.425	.419
	50	.739	.828	.822	.857	.864	.857
AA2	20	.781	.455	.003	.449	.248	.241
	50	.998	.986	.028	.983	.954	.924
AA3	10	.575	.433	.150	.440	.301	.422
	20	.930	.903	.565	.899	.847	.881

If one does not know much about the unknown density function, the undertaken simulation study suggests that the test $I^2(c)$ is a good alternative to both A^2 and W^2 tests. In fact, for Type I alternatives the $I^2(c)$ performance is close to that one of A^2 or W^2, and for Type II alternatives $I^2(c)$ is better or significantly better than A^2 or W^2 tests.

The practical performance shown by the Bickel-Rosenblatt test with fixed bandwidth impels the generalization of the results presented in this paper to the test of a composite

null hypothesis. In case of location-scale null families of density functions, this demands the use of a kernel density estimator with data-dependent fixed bandwidth matrix which is out of the scope of this paper. In a future paper we intend to address this subject.

5 Proofs of Theorem 2 and Corollary 2

Proof of Theorem 2: In order to use Theorem 1.2.2 of Nikitin (1995) due to Bahadur (1967, 1971), we first note that from the strong law of large number for U-statistics (cf. Theorem 3.1.1 of Koroljuk and Borovskich (1989)) we have

$$n^{-1}I_n^2(h) \underset{n\to+\infty}{\overset{as}{\longrightarrow}} b_{I^2(h)}(f),$$

for all f, where $b_{I^2(h)}(\cdot)$ is given in Theorem 2. Secondly, it is necessary to solve the problem of determining large deviation asymptotics of the sequence $(I_n^2(h))$ under the null hypothesis. This problem can be solved by using the integral and quadratic form of $I_n^2(h)$ and a generalization of Chernoff large-deviation result due to Sethuraman (1964) (see also Nikitin (1995), p. 23). In fact, we have

$$I_n^2(h) = \left(n^{-1}|Z_{1,h} + \cdots + Z_{n,h}|_2\right)^2,$$

where $(Z_{i,h})$ are i.i.d. random variables taking values on $L_2(\lambda)$ given by $Z_{i,h}(x) = K_h(x - X_i) - K_h \star f_0(x)$, for $x \in \mathbb{R}^d$. Moreover, for all $g \in L_2(\lambda)$ we have $\int\int g(x)Z_{1,h}(x)dxdP = \int g(x)\int Z_{1,h}(x)dPdx = 0$, and, for all $z \in \mathbb{R}$, $\int \exp(z|Z_{1,h}|_2)dP \leq \exp(z\int |Z_{1,h}|_2 dP) < +\infty$, since $|Z_{1,h}|_2$ is a bounded random variable. The conditions of Sethuraman's theorem are thus fulfilled. Then, for all $a > 0$,

$$\lim_{n\to+\infty} n^{-1}\ln P(n^{-1}I_n^2(h) \geq a) = G(a),$$

where G is a continuous function in a neighbourhood V_0 of zero such that

$$G(a) = -\frac{a}{2\sigma_h^2}(1 + o(1)), \quad \text{as } a \to 0,$$

and

$$\begin{aligned}
\sigma_h^2 &= \sup\left\{\int\left(\int g(x)Z_{1,h}(x)dx\right)^2 dP : |g|_2 = 1\right\} \\
&= \sup\left\{\int\int g(x)g(y)\int Z_{1,h}(x)Z_{1,h}(y)dPdxdy : |g|_2 = 1\right\} \\
&= \sup\left\{\int\int g(x)g(y)\bar{Q}_h(x,y)dxdy : |g|_2 = 1\right\},
\end{aligned}$$

with $\bar{Q}_h(x,y) = \int k(x,u;h)k(y,u;h)dP_0(u)$ and k is given by (3).

By the Rayleigh equation (see Dunford and Schwartz (1963)), σ_h^2 is the largest eigenvalue of the integral operator $\bar{A}_h$, with kernel $\bar{Q}_h$, defined on $L_2(\lambda)$. Since the set of eigenvalues of $\bar{A}_h$ coincide with the corresponding one of the operator A_h defined by (4), we get

$$G(a) = -\frac{a}{2\lambda_{1,h}}(1 + o(1)), \quad \text{as } a \to 0,$$

where $\lambda_{1,h}$ is the largest eigenvalue of the operator A_h.

Finally, from the continuity in f_0 of the function $b_{I^2(h)}(\cdot)$ from $L_1(\lambda)$ to $[0, +\infty[$, there exists a neigbourhood V_{f_0} of f_0 such that $\{b_{I^2(h)}(f) : f \in V_{f_0}\} \subset V_0$, and therefore, from Theorem 1.2.2 of Nikitin (1995), we conclude that

$$C_{I^2(h)}(f) = -2\, G(b_{I^2(h)}(f)),$$

for all $f \in V_{f_0}$. □

Proof of Corollary 2: For $b^o_{I^2(h)}(f(\cdot;\theta))$ given in Corollary 1, we have

$$\begin{aligned} b^o_{I^2(h)}(f(\cdot;\theta)) &= \int\int Q_h(u,v)\frac{\partial \ln f}{\partial\theta}(u;\theta_0)\frac{\partial \ln f}{\partial\theta}(v;\theta_0)\, dP_0(u)dP_0(v) \\ &= \left\langle A_h\frac{\partial \ln f}{\partial\theta}(\cdot;\theta_0)\,,\ \frac{\partial \ln f}{\partial\theta}(\cdot;\theta_0)\right\rangle. \end{aligned}$$

The result follows now from Corollary 1 and the representation $A_h q = \sum_{k=1}^{\infty} \lambda_{k,h}\langle q, q_{k,h}\rangle\, q_{k,h}$, for all $q \in L_2(P_0)$. □

Acknowledgements

This work was partially supported by CMUC/FCT. The author would like to thank Paulo Oliveira for his helpful comments.

References

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. and Sorensen, D. (1999). *LAPACK users' guide*, Third edition, SIAM, Philadelphia.

Anderson, N.H., Hall, P., Titterington, D.M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates, *J. Multivariate Anal.*, 50, 41-54.

Bahadur, R.R. (1967). Rates of convergence of estimates and test statistics, *Ann. Math. Statist.*, 38, 303-324.

Bahadur, R.R. (1971). *Some Limit Theorems in Statistics*, SIAM, Philadelphia.

Bickel, P.J., Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *Ann. Statist.*, 1, 1071-1095.

Bowman, A.W. (1992). Density based tests for goodness-of-fit normality, *J. Statist. Comput. Simul.*, 40, 1-13.

Bowman, A.W., Foster, P.J. (1993). Adaptive smoothing and density-based tests of multivariate normality, *J. Amer. Statist. Assoc.*, 88, 529-537.

Dunford, N., Schwartz, J.T. (1963). *Linear Operators, Part II*, Interscience Publishers, New York.

Durbin, J., Knott, M., Taylor, C.C. (1975). Components of Cramér-von Mises Statistics II, *J. Roy. Statist. Soc. Ser. B*, 37, 216-237.

Durio, A., Nikitin, Ya.Yu. (2003). Local Bahadur efficiency of some goodness-of-fit tests under skew alternatives, *J. Statist. Plann. Inference*, 115, 171-179.

Epps, T.W., Pulley, L.B. (1983). A test for normality based on the empirical characteristic function, *Biometrika*, 70, 723-726.

Eubank, R.L., LaRiccia, V.N. (1992). Asymptotic comparison of Crámer-von Mises and nonparametric function estimation techniques for testing goodness-of-fit, *Ann. Statist.*, 20, 2071-2086.

Fan, Y. (1994). Testing the goodness of fit of a parametric density function by kernel method, *Econometric Theory*, 10, 316-356.

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters, *Econometric Theory*, 14, 604-621.

Gouriéroux, C., Tenreiro, C. (2001). Local power properties of kernel based goodness of fit tests, *J. Multivariate Anal.*, 78, 161-190.

Gregory, G.G. (1977). Large sample theory for *U*-statistics and tests of fit, *Ann. Statist.*, 5, 110-123.

Gregory, G.G. (1980). On efficiency and optimality of quadratic tests, *Ann. Statist.*, 8, 116-131.

Hall, P. (1997). *The Bootstrap and Edgeworth Expansion*, Springer, New York.

Henze, N., Zirkler, B. (1990). A class of invariante consistent tests for multivariate normality, *Commun. Stat. Theory Methods*, 19, 3595-3617.

Henze, N., Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality, *J. Multivariate Anal.*, 62, 1-23.

Koroljuk, V.S., Borovskich, Yu.V. (1989). *Theory of U-Statistics*, Kluwer, Dordrecht.

Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press.

Parzen, E. (1962). On estimation of a probability density function and mode, *Ann. Math. Statist.*, 33, 1065-1076.

Ramberg, J.S., Schmeiser, B.W. (1974). An approximate method for generating asymmetric random variables, *Commun. ACM*, 17, 78-82.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function, *Ann. Math. Statist.*, 27, 832-837.

Sethuraman, J. (1964). On the probability of large deviations of families of sample means, *Ann. Math. Statist.*, 35, 1304-1316.

Stephens, M.A. (1986). Tests based on EDF statistics, In: *Goodness-of-Fit Techniques*, ed. D'Agostino, R.B. and Stephens, M.A., Marcel Dekker, New York, 97-193.

SORT 29 (2) July-December 2005, 217-234

On sequential and fixed designs for estimation with comparisons and applications

Mekki Terbeche[1], Broderick O. Oluyede[2], Ahmed Barbour[3]

[1]*University of Oran,* [2]*Georgia Southern University,* [3]*U.A.E. University*

Abstract

A fully sequential approach to the estimation of the difference of two population means for distributions belonging to the exponential family of distributions is adopted and compared with the best fixed design. Results on the lower bound for the Bayes risk due to estimation and expected cost are presented and shown to be of first order efficiency. Applications involving the Poisson and exponential distributions with gamma priors as well as the Bernoulli distribution with beta priors are given. Finally, some numerical results are presented.

MSC: 62L12

Keywords: Estimation, Loss function, Prior distribution, Sequential design.

1 Introduction

The family of exponential type distributions play an important role in a wide variety of areas in probability and statistics. For example, the gamma distribution which belong to the family of exponential distributions is used to model lifetimes of various practical situations including but not limited to lengths of time between catastrophic events (floods, earthquakes and so on), lengths of time between emergency arrivals at a hospital and distance traveled by a wildlife ecologist between sighting of an endangered species. The exponential distribution which is a special case of the gamma distribution have

Address for correspondence: Mekki Terbeche. Department of Mathematics, University of Oran, Es-Senia, Algeria. Broderick O. Oluyede. Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460. *Boluyede@GeorgiaSouthern.edu. Ahmed Barbour. College of Information Technology, U.A.E. University, Al Ain, United Arab Emirates.

Received: June 2001
Accepted: June 2005

been used to describe the amount of time between occurrences of random events such as those described above. Further examples of the exponential type distributions include the Poisson and binomial distributions. The Poisson distribution provides a realistic model for many random phenomena such as number of fatal traffic accident per week at a busy intersection, the number of radioactive particle emissions per unit time, the number of telephone calls per hour arriving at a switchboard to mention a few. In this paper, we consider the problem of designing an experiment to estimate the difference between two population means for distributions belonging to the exponential family plus expected cost of drawing samples from either groups using a Bayesian approach. We explore and compare the Bayes risk due to estimation plus the expected cost of sampling. Numerical results on the relative efficiency are also presented.

This paper is organized as follows. Section 2 contains some preliminaries and basic results for the class of exponential type distributions. In Section 3, the problem is presented and the mathematical results on the fully sequential and best fixed designs derived. Some bounds are presented. In Section 4, we present applications and numerical results on the comparisons of the Bayes risk for the procedures described in Section 3. Applications are presented for the comparisons of two Poisson means, comparisons of two exponential means and the comparison of two Bernoulli means. Applications on the comparisons of the normal means with known variances as well as the comparisons of two normal variances with known means will be treated in the future. This paper concludes with a summary and discussion.

2 Preliminaries and basic results

In this section, we consider the family of exponential-type probability distributions on the real line, given by the family of densities $\mathcal{G}$ with respect to the Lebesgue measure. A natural form of an exponential family is as follows:

$$f(x,\theta) = \exp\{\theta T(x) + S(x) - \psi(\theta)\}, \tag{1}$$

where $f \in \mathcal{G}$. In this setting $E(T(X)) = \psi'(\theta)$ and $\mathrm{var}(T(X)) = \psi''(\theta)$. See Lehmann [3]. Consider two independent random variables X and Y with densities given by $f(x,\theta) = \exp\{\theta T(x) + S(x) - \psi(\theta)\}$, and $g(y,\omega) = \exp\{\omega T(y) + U(y) - \phi(\omega)\}$ respectively. Our objective is to estimate $\lambda = E_\theta[T(X)] - E_\omega[T(X)] = \psi'(\theta) - \phi'(\omega)$ with square error loss.

Definition 1 *The Bayes risk of an estimate $\hat{\lambda}$ with respect to the prior distribution $\pi(\theta)$ is*

$$r(\theta,\hat{\lambda}) = E[R(\theta,\hat{\lambda})], \tag{2}$$

where $R(\theta,\hat{\lambda}) = E[L(\theta,\hat{\lambda})]$ and $L(\theta,\hat{\lambda})$ is the loss function.

The adopted approach in this paper is Bayesian and it is assumed that the prior distributions of θ and ω are the conjugate priors given by: $\pi_1(\theta) \propto \exp[t(\theta\mu_1 - \phi(\theta))]$ and $\pi_2(\omega) \propto \exp[s(\omega\mu_2 - \phi(\omega))]$, where $\mu_1 = E_{\pi_1}[\phi^{'}(\theta)]$ and $\mu_2 = E_{\pi_2}[\psi^{'}(\omega)]$ are prior estimations of $E_\theta[T(X)]$ and $E_\omega[T(X)]$ respectively, if these densities and their derivatives decay to zero in the tails, (See West, 1985, 1986), and $t > 0$ and $s > 0$ are positive real numbers that can be interpreted as prior sample sizes.

If $X_1, X_2, \ldots, X_m$ is a random sample of X and $Y_1, Y_2, \ldots, Y_n$ is a random sample of Y, the Bayes estimator of λ is given by

$$\begin{aligned}\hat{\lambda}(x_1, \ldots, x_m, y_1, \ldots, y_n) &= E[\lambda | x_1, \ldots, x_m, y_1, \ldots, y_n] \\ &= E[\psi^{'}(\theta) | x_1, \ldots, x_m] - E[\psi^{'}(\omega) | y_1, \ldots, y_n], \end{aligned} \tag{3}$$

where

$$E[\psi^{'}(\theta) | x_1, \ldots, x_m] = \frac{m\overline{T}_m^X + t\mu_1}{m + t}$$

with $\overline{T}_m^X = \frac{T(x_1)+\cdots+T(x_m)}{m}$ and

$$E[\psi^{'})(\omega) | y_1, \ldots, y_n] = \frac{m\overline{T}_n^Y + s\mu_2}{n + s}$$

with $\overline{T}_n^Y = \frac{T(y_1)+\cdots+T(y_n)}{n}$.

If $\mathbf{X}=(X_1, \ldots, X_m)$ and $\mathbf{Y}=(Y_1, \ldots, Y_n)$, $\mathbf{x}=(x_1, \ldots, x_m)$ and $\mathbf{y}=(y_1, \ldots, y_n)$, the Bayes risk is given by

$$\begin{aligned} r(\pi_1, \pi_2) &= r(\hat{\lambda}(\mathbf{x}, \mathbf{y})) \\ &= E_{(\mathbf{X},\mathbf{Y})}\left[E_{\lambda|(\mathbf{X},\mathbf{Y})}\left[(\lambda - \hat{\lambda}(\mathbf{x}, \mathbf{y}))^2\right]\right] \\ &= E_{(\mathbf{X},\mathbf{Y})}[\mathrm{var}(\lambda|(\mathbf{X}, \mathbf{Y}))] \\ &= E_{(\mathbf{X},\mathbf{Y})}[\mathrm{var}(\psi^{'}(\theta)|\mathbf{X}) + \mathrm{var}(\phi^{'}(\omega)|\mathbf{Y})] \\ &= E_{\mathbf{X}}\left[E_{\theta|\mathbf{X}}\left[\frac{\psi^{''}(\theta)}{m + t}\right]\right] + E_{\mathbf{Y}}\left[E_{\omega|\mathbf{Y}}\left[\frac{\phi^{''}(\omega)}{n + s}\right]\right]. \end{aligned} \tag{4}$$

3 Sequential and best fixed designs

3.1 The problem

In order to set up the problem we adopt the notation given in Berger (1985, Chapter 7). The loss function is given by

$$L(\lambda, a, m, n) = (\lambda - a)^2 + c_1 m + c_2 n, \tag{5}$$

and the decision rule are sequential decision procedures $\Delta_S = (\tau, \delta)$ where τ is called the stopping rule of the procedure and consist of functions $\tau_{m,n}(x_1, \ldots, x_m, y_1, \ldots, y_n)$ that specify the probability of stopping sampling and making a decision after observing $(x_1, \ldots, x_m, y_1, \ldots, y_n)$; δ is the decision rule of the design Δ_S and consists of a series of decision functions $\delta_{m,n}(x_1, \ldots, x_m, y_1, \ldots, y_n)$ that specify the estimated value of λ when the sampling has stopped after observing $(x_1, \ldots, x_m, y_1, \ldots, y_n)$.

For the stopping rule τ, the Bayes risk is given by:

$$\begin{aligned} r(\tau, \pi_1, \pi_2) &= E_{(\mathbf{X},\mathbf{Y},\tau)}\left[\frac{U_m}{m+t} + \frac{V_n}{n+s} + c_1 m + c_2 n\right] \\ &= E_{(\mathbf{X},\tau_\mathbf{m})}\left[\frac{U_m}{m+t} + c_1 m\right] + E_{(\mathbf{Y},\tau_\mathbf{n})}\left[\frac{V_n}{n+s} + c_2 n\right], \end{aligned} \tag{6}$$

where τ_m and τ_n are the marginal stopping rules of τ, and $U_m = E_{(\mathbf{X},\tau_\mathbf{m})}[\psi''(\theta)]$, $V_n = E_{(\mathbf{Y},\tau_\mathbf{n})}[\phi''(\omega)]$, t and s are fixed and depend on the posteriors, m and n are unknown.

The fixed designs are particular cases of Δ_S where the stopping rules τ_m and τ_n are equal to one if $m = m^F$ and $n = n^F$ and zero otherwise and their optimal values $m_{opt}(\pi)$ and $n_{opt}(\pi)$ are given in this section.

3.2 Mathematical results

In this subsection, the mathematical results are presented. We compare the best fixed design with the sequential optimal random design. Let c_1 and c_2 be the cost of sampling per observation from populations 1 and 2 respectively. The Bayes risk due to estimation plus expected sampling cost is given by equation (6). The objective or goal is to minimize $r(\tau, \pi_1, \pi_2)$.

In the sequential allocation, for a fixed total number of observations the problem is to allocate the number of observations to be taken from each population to achieve or nearly achieve some optimality condition such as minimizing the Bayes risk when the allocation is done sequentially. That is, at each stage the decision to observe X or Y may depend on available information from all previous stages.

Note that at stage t:

a) If $U_m^{1/2} \geq c_1^{1/2}(m+t)$ take another observation of X; otherwise stop observing X.
b) If $V_n^{1/2} \geq c_2^{1/2}(n+s)$ take another observation of Y; otherwise stop observing X.

The rule takes an additional observation of X (respectively Y) if $m_{opt}(\theta|x_1,\ldots,x_m) \geq 1$ (respectively $n_{opt}(\theta|y_1,\ldots,y_n) \geq 1$, where $m_{opt}(\pi) = \left(E_\pi[\psi''(\theta)]/c_1\right)^{1/2} - t$ (respectively $n_{opt}(\pi) = \left(E_\pi[\phi''(\omega)]/c_2\right)^{1/2} - s$) are the sample sizes of the fixed design when the distribution of θ (respectively ω) is π. The sequential design achieves the lower bound. That is,

$$\liminf_{c_1,c_2\to 0}\left(\frac{r(\Delta)}{(c_1+c_2)^{1/2}}\right) = 2E[(\gamma_1\psi''(\theta))^{1/2} + (\gamma_2\phi''(\omega))^{1/2}]. \tag{7}$$

To see this, and for simplicity of the computations, we take the exponential family with probability distribution of the form $f_\theta(x) = \exp[\theta x - \psi(\theta)]$, $x \in \mathbb{R}$, $\theta \in \Omega$. Clearly, $E_\theta(X) = \psi'(\theta)$ and $\mathrm{var}_\theta(X) = \psi''(\theta)$, after differentiating the identity $\int e^{\theta x-\psi(\theta)}dx = 1$, once and twice with respect to θ and simplifying each expression respectively. Similarly, $E_\omega(Y) = \phi'(\omega)$ and $\mathrm{var}_\omega(Y) = \phi''(\omega)$. Following Diaconis and Ylvisaker [2], the form of the conjugate for exponential families, for $t > 0$ and $s > 0$ are

$$\pi(\theta) = \frac{e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}, \tag{8}$$

and

$$\gamma(\omega) = \frac{e^{s[\mu\omega-\phi(\omega)]}}{\int e^{s[\mu\omega-\phi(\omega)]}d\omega}, \tag{9}$$

respectively. We assume that θ and ω are independent random variables with conjugate prior distributions given above. If $(X_1, X_2, \ldots, X_m)$ is a random sample of X and $(Y_1, Y_2, \ldots, Y_n)$ is a random sample of Y, then

$$f_\theta(X_1,\ldots,X_m) = \exp[m(\theta\overline{X} - \psi(\theta))], \tag{10}$$

where $\overline{X} = (X_1,\ldots,X_m)/m$ and

$$g_\omega(Y_1,\ldots,Y_n) = \exp[n(\omega\overline{Y} - \phi(\omega))], \tag{11}$$

where $\overline{Y} = (Y_1,\ldots,Y_n)/n$.

The posterior distribution of θ when m observations $(X_1, X_2, \ldots, X_m)$ are sampled from population 1 is

$$\begin{aligned}\pi(\theta|X_1, X_2, \ldots, X_m) &= f_\theta(X_1, \ldots, X_m) / \int f_\theta(X_1, \ldots, X_m)\pi(\theta)d\theta \\ &= \frac{e^{m(\theta\overline{X}-\psi(\theta))}\frac{e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}}{\int e^{m(\theta\overline{X}-\psi(\theta))}\frac{e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}d\theta} \\ &= \frac{e^{\theta[m\overline{X}+t\mu]-(m+t)\psi(\theta)}}{\int e^{\theta[m\overline{X}+t\mu]-(m+t)\psi(\theta)}d\theta}. \qquad (12)\end{aligned}$$

Set $t_1 = m + t$ and $\mu_1 = \frac{t}{t_1}\mu + \frac{m}{t_1}\overline{X}$. Then

$$\pi(\theta|X_1, \ldots, X_m) = \frac{e^{t_1[\mu_1\theta-\psi(\theta)]}}{\int e^{t_1[\mu_1\theta-\psi(\theta)]}d\theta}. \qquad (13)$$

Similarly,

$$\gamma(\omega|Y_1, \ldots, Y_n) = \frac{e^{s_1[\mu_2\omega-\phi(\omega)]}}{\int e^{s_1[\nu_1\omega-\phi(\omega)]}d\omega}, \qquad (14)$$

where $s_1 = n + s$ and $\mu_2 = \frac{s}{s_1}\nu + \frac{n}{s_1}\overline{Y}$.

Next we show that the posterior mean and variance of $\psi'(\theta)$ given $x_1, \ldots, x_m$ are $E[\psi'(\theta)|x_1, \ldots, x_m] = \mu_1$ and $\text{var}[\psi'(\theta)|x_1, \ldots, x_m] = E_{\theta|\mathbf{X}}[\frac{\psi''(\theta)}{m+t}]$ respectively. First we state a useful lemma. For a proof of the lemma see Hajek and Sidak (1967).

Lemma 1 *If f is an absolutely continuous integrable and real valued function for which $\int |f(\omega)|d\omega < \infty$, then $\lim_{\omega\to-\infty} f(\omega) = 0$ and $\lim_{\omega\to+\infty} f(\omega) = 0$.*

The posterior mean of $\psi'(\theta)$ given $x_1, \ldots, x_m$ is

$$\begin{aligned}E[\psi'(\theta)|x_1, \ldots, x_m] &= \int \psi'(\theta)\pi(\theta|X_1, \ldots, X_m)d\theta \\ &= \int \psi'(\theta)\frac{e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}d\theta \\ &= \frac{1}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}\int \psi'(\theta)e^{t[\mu\theta-\psi(\theta)]}d\theta \\ &= -\frac{1}{t_1\int e^{t[\mu\theta-\psi(\theta)]}d\theta}\int [t_1(\mu_1 - \psi'(\theta))e^{t[\mu\theta-\psi(\theta)]} - t_1\mu_1 e^{t[\mu\theta-\psi(\theta)]}]d\theta\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{t_1 \int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int \frac{d}{d\theta}[e^{t[\mu\theta-\psi(\theta)]}]d\theta + \mu_1 \int \frac{e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}d\theta \\
&= -\frac{1}{t_1 \int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int \frac{d}{d\theta}[e^{t[\mu\theta-\psi(\theta)]}]d\theta + \mu_1 \\
&= \mu_1 - \frac{1}{t_1 \int e^{t[\mu\theta-\psi(\theta)]}d\theta}[\lim_{\theta\to+\infty} e^{t[\mu\theta-\psi(\theta)]} - \lim_{\theta\to+\infty} e^{t[\mu\theta-\psi(\theta)]}] \\
&= \mu_1, \qquad (15)
\end{aligned}
$$

since the two limits vanish by virtue of the lemma given above.

The posterior variance of $\psi'(\theta)$ given $x_1, \ldots, x_m$ is

$$
\begin{aligned}
\text{var}[\psi'(\theta)|x_1, \ldots, x_m] &= \int [\psi'(\theta) - E(\psi'(\theta))]^2 \pi(\theta|X_1, \ldots, X_m)d\theta \\
&= \frac{1}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int [\psi'(\theta) - E(\psi'(\theta))]^2 e^{t[\mu\theta-\psi(\theta)]}d\theta. \qquad (16)
\end{aligned}
$$

Let $\alpha(\theta) = t_1[\mu\theta - \psi(\theta)]$, then

$$
[E(\psi'(\theta)) - \psi'(\theta)]^2 = [\frac{1}{t_1}\frac{d\alpha(\theta)}{d\theta}]^2, \qquad (17)
$$

so that

$$
\begin{aligned}
\text{var}[\psi'(\theta)|x_1, \ldots, x_m] &= \frac{1}{t_1^2 \int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int [\frac{d\alpha(\theta)}{d\theta}]^2 e^{\alpha(\theta)}d\theta \\
&= \frac{1}{t_1^2 \int e^{t[\mu\theta-\psi(\theta)]}d\theta}[[\alpha'(\theta)e^{\alpha(\theta)}]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \alpha''(\theta)e^{\alpha(\theta)}d\theta] \\
&= \frac{1}{t_1^2 \int e^{t[\mu\theta-\psi(\theta)]}d\theta}[\lim_{\theta\to+\infty} \alpha'(\theta)e^{\alpha(\theta)}] - [\lim_{\theta\to-\infty} \alpha'(\theta)e^{\alpha(\theta)}] \\
&\quad - \int_{-\infty}^{\infty} \alpha''(\theta)e^{\alpha(\theta)}d\theta \\
&= -\frac{1}{t_1^2 \int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int_{-\infty}^{\infty} \alpha''(\theta)e^{\alpha(\theta)}d\theta \\
&= \frac{t_1}{t_1^2 \int e^{t[\mu\theta-\psi(\theta)]}d\theta} \int \psi''(\theta)e^{t[\mu\theta-\psi(\theta)]}d\theta
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{t_1}\int \frac{\psi''(\theta)e^{t[\mu\theta-\psi(\theta)]}}{\int e^{t[\mu\theta-\psi(\theta)]}d\theta}d\theta \\
&= \frac{1}{t_1}E[\psi''(\theta)|x_1,\ldots,x_m] \\
&= \frac{E[\psi''(\theta)|x_1,\ldots,x_m]}{m+t} \\
&= E[\frac{\psi''(\theta)}{m+t}|x_1,\ldots,x_m] \\
&= E_{\theta|\mathbf{x}}[\frac{\psi''(\theta)}{m+t}],
\end{aligned}
\tag{18}
$$

which also gives the proof of equation 4. □

In the best fixed design or policy the risk function $r(\Delta)$ is minimized as a function of fixed sample sizes m and n. This policy is asymptotically the best among the non-sequential or non-random policies. The best fixed design is determined by $m_{opt}(\pi) = (E[\psi''(\theta)]/c_1)^{1/2} - t$ and $n_{opt}(\pi) = (E[\phi''(\omega)]/c_2)^{1/2} - s$, and achieves the lower bound under suitable conditions.

Theorem 1 *Let c_1 and c_2 be such that $\frac{c_1}{c_1+c_2} \to \gamma_1$, as $c_1, c_2 \to 0$, $0 < \gamma_1 < 1$ and $\gamma_2 = 1 - \gamma_1$. Then for any random design Δ,*

$$
\liminf_{c_1,c_2\to 0}\left(\frac{r(\Delta)}{(c_1+c_2)^{1/2}}\right) \geq 2E[(\gamma_1\psi''(\theta))^{1/2} + (\gamma_2\phi''(\omega))^{1/2}].
$$

Proof. Observe that

$$
r(\Delta) \geq 2E\left[(c_1U_m)^{1/2} + (c_2V_n)^{1/2}\right] - tc_1 - sc_2. \tag{19}
$$

for any procedure Δ.

Now,

$$
\begin{aligned}
\left(\frac{r(\Delta)}{(c_1+c_2)^{1/2}}\right) &\geq 2E\left[\frac{(c_1U_m)^{1/2}}{(c_1+c_2)^{1/2}} + \frac{(c_2V_n)^{1/2}}{(c_1+c_2)^{1/2}}\right] \\
&- tc_1^{1/2}\left(\frac{c_1}{c_1+c_2}\right)^{1/2} - sc_2^{1/2}\left(\frac{c_2}{c_1+c_2}\right)^{1/2}.
\end{aligned}
\tag{20}
$$

Consequently,

$$\liminf_{c_1,c_2\to 0}\left(\frac{r(\Delta)}{(c_1+c_2)^{1/2}}\right) \geq 2E[(\gamma_1\psi^{''}(\theta))^{1/2} + (\gamma_2\phi^{''}(\omega))^{1/2}].$$

The last inequality follows from the application of Fatou's lemma.

Note that for any fixed design $\Delta_{\mathcal{F}}$,

$$\begin{aligned} r(\Delta_{\mathcal{F}}) &= 2[(c_1E_\theta[\psi^{''}(\theta)])^{1/2} + (c_2E_\omega[\psi^{''}(\omega)])^{1/2}] \\ &+ (m+t)^{-1}[E_\theta(\psi^{''}(\theta))^{1/2} - (m+t)c_1^{1/2}]^2 \\ &+ (n+s)^{-1}[E_\omega(\phi^{''}(\omega))^{1/2} - (n+s)c_2^{1/2}]^2 - (tc_1+sc_2). \end{aligned} \tag{21}$$

If $m = \left(E_\pi[\psi^{''}(\theta)]/c_1\right)^{1/2} - t$ and $n = \left(E_\pi[\phi^{''}(\omega)]/c_2\right)^{1/2} - s$, then

$$r(\Delta_{\mathcal{F}}) = 2E[(c_1E_\theta[\psi^{''}(\theta)])^{1/2} + (c_2E_\omega[\phi^{''}(\omega)])^{1/2}] - (tc_1+sc_2). \tag{22}$$

Moreover, if c_1 and c_2 are such that $\frac{c_1}{c_1+c_2} \to \gamma_1$, as $c_1, c_2 \to 0$, $0 < \gamma_1 < 1$ and $\gamma_2 = 1-\gamma_1$, then

$$\liminf_{c_1,c_2\to 0}\left(\frac{r(\Delta_{\mathcal{F}})}{(c_1+c_2)^{1/2}}\right) = 2[(\gamma_1E\theta[\psi^{''}(\theta)])^{1/2} + (\gamma_2E_\omega[\phi^{''}(\omega)])^{1/2}].$$

□

Theorem 2 *Let $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{F}}$ denote the first order sequential and fixed designs respectively. Then*

$$0 \leq \liminf_{c_1,c_2\to 0}\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \leq 1. \tag{23}$$

Proof. Note that

$$\liminf_{c_1,c_2\to 0}\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = \frac{(\gamma_1)^{1/2}E(\psi^{''}(\theta))^{1/2} + (\gamma_2)^{1/2}E(\varphi^{''}(\omega))^{1/2}}{(\gamma_1E\psi^{''}(\theta))^{1/2} + (\gamma_2E\varphi^{''}(\omega))^{1/2}}. \tag{24}$$

Applying Jensen's inequality, we have

$$0 \leq \liminf_{c_1,c_2\to 0}\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \leq 1. \tag{25}$$

□

Theorem 3 *If $c_1 = c_2$ then*

$$\liminf_{c_1 \to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{E(\psi''(\theta))^{1/2} + E(\varphi''(\omega))^{1/2}}{(E\psi''(\theta))^{1/2} + (E\varphi''(\omega))^{1/2}}. \tag{26}$$

Proof. We have

$$\liminf_{c_1, c_2 \to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{(\gamma_1)^{1/2} E(\psi''(\theta))^{1/2} + (\gamma_2)^{1/2} E(\varphi''(\omega))^{1/2}}{(\gamma_1 E\psi''(\theta))^{1/2} + (\gamma_2 E\varphi''(\omega))^{1/2}}. \tag{27}$$

If $c_1 = c_2$, then $\gamma_1 = \gamma_2$ and the result follows. □

Corollary 1 *If $c_1 = c_2$, $\psi''(\theta) = \varphi''(\omega)$, and $\pi_1 = \pi_2$, then*

$$\liminf_{c_1 \to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{E(\psi''(\theta))^{1/2}}{(E\phi''(\omega))^{1/2}}. \tag{28}$$

□

The asymptotic results in section 3 are derived in the following sense. Sampling sizes tending to infinity are achieved by taking the costs of sampling (c_1, c_2) tending to zero, since c_1 and c_2 may differ from population to population. Simultaneous control over c_1 and c_2 is maintained by assuming that $\frac{c_1}{c_1+c_2} \to \gamma_1$, $\frac{c_2}{c_1+c_2} \to \gamma_2$, so that c_1 and c_2 tend to zero at the same rate.

Theorem 2 states that the lower bound for the sequential design is smaller than the lower bound for the best fixed design. This makes sense due to the fact that we use all previous information about the population for the sequential design as well as the information on the priors for the best fixed design.

4 Application

In this section, we present applications of the results in Sections 2 and 3. Specifically, applications involving the Poisson and exponential distributions with gamma priors as well as the Bernoulli distribution with beta priors are given. Some numerical results on the relative efficiency of the estimation problem concerning the Poisson and exponential distributions with gamma priors are also presented.

4.1 Comparison of two Poisson means

Let the distribution of the random variables X and Y be given by $f(x, \theta)$ and $g(y, \omega)$ respectively, where

$$f(x,\theta) = \frac{\theta^x e^{-\theta}}{x!}, \tag{29}$$

$x = 0, 1, 2, \ldots, \theta > 0$ and

$$g(y,\omega) = \frac{\omega^y e^{-\omega}}{y!}, \tag{30}$$

$y = 0, 1, 2, \ldots, \omega > 0$. We assume that θ and ω are independent and distributed as $Gamma(a, p)$, $a > 0$, $p > 0$ and $Gamma(c, q)$, $c > 0$, $q > 0$. It follows therefore from Theorem 1 that

$$\liminf_{c_1,c_2\to 0} \frac{R(\Delta_{\mathcal{S}})}{R(\Delta_{\mathcal{F}})} = \frac{\left(\frac{\Gamma(a+1/2)}{p^{1/2}\Gamma(a)} + \frac{\Gamma(c+1/2)}{q^{1/2}\Gamma(c)}\right)}{((a/p)^{1/2} + (c/q)^{1/2})}, \tag{31}$$

$a > 0, c > 0, p > 0, q > 0$. □

Note that (*a*) If $a/p = c/q = d > 0$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = (1/2)\left(\frac{\Gamma(a+1/2)}{a^{1/2}\Gamma(a)} + \frac{\Gamma(c+1/2)}{c^{1/2}\Gamma(c)}\right). \tag{32}$$

(*b*) If $a \to 0$ and $c \to \infty$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \to \frac{1}{2}. \tag{33}$$

(*c*) If $a = c$ and $p = q$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = \frac{\Gamma(a+1/2)}{a^{1/2}\Gamma(a)}. \tag{34}$$

(*d*) If $a = c = p = q = 1/2$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = 0.7979. \tag{35}$$

□

4.2 Comparison of two Bernoulli means

Let the distribution of the random variables X and Y be given by $f(x,\theta)$ and $g(y,\omega)$ respectively, where

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}, \tag{36}$$

$x = 0, 1$, $0 < \theta < 1$ and

$$g(y,\omega) = \omega^y(1-\omega)^{1-y}, \tag{37}$$

$y = 0, 1$, $0 < \omega < 1$. We assume that θ and ω are independent and distributed as $Beta(a,b)$, $a > 0$, $b > 0$ and $Beta(c,d)$, $c > 0$, $d > 0$. It follows therefore from Theorem 3 that

$$\liminf_{c_1,c_2\to 0} \frac{R(\Delta_{\mathcal{S}})}{R(\Delta_{\mathcal{F}})} = \frac{E(\theta(1-\theta))^{1/2} + E(\omega(1-\omega))^{1/2}}{(E[\theta(1-\theta)])^{1/2} + (E[\omega(1-\omega)])^{1/2}}, \tag{38}$$

where

$$E(\theta(1-\theta))^{1/2} = \frac{\Gamma(a+1/2)\Gamma(b+1/2)}{(a+b)\Gamma(a)\Gamma(b)}, \tag{39}$$

for $a > 0$, $b > 0$, and

$$E(\omega(1-\omega))^{1/2} = \frac{\Gamma(c+1/2)\Gamma(d+1/2)}{(c+d)\Gamma(c)\Gamma(d)}, \tag{40}$$

for $c > 0$, $d > 0$. Similarly,

$$E[\theta(1-\theta)] = ab/(a+b+1)(a+b), \tag{41}$$

for $a > 0$, $b > 0$, and

$$E[\omega(1-\omega)] = cd/(c+d+1)(c+d), \tag{42}$$

for $c > 0$, $d > 0$. For the beta distribution, that is, $\theta \sim Beta(a,b)$, it is well known that $E(\theta) = \frac{a}{a+b}$, and $\mathrm{var}(\theta) = \frac{ab}{(a+b+1)(a+b)^2}$. Similarly, if $\omega \sim Beta(c,d)$, then $E(\omega) = \frac{c}{c+d}$, and $\mathrm{var}(\omega) = \frac{cd}{(c+d+1)(c+d)^2}$.

The ratio of the sequential to the best fixed design is

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = \frac{\dfrac{\Gamma(a+1/2)\Gamma(b+1/2)}{(a+b)\Gamma(a)\Gamma(b)} + \dfrac{\Gamma(c+1/2)\Gamma(d+1/2)}{(c+d)\Gamma(c)\Gamma(d)}}{(ab/(a+b+1)(a+b))^{1/2} + (cd/(c+d+1)(c+d))^{1/2}}, \tag{43}$$

$a > 0$, $b > 0$, $c > 0$, $d > 0$.

Note that (*a*) If $a = c$ and $b = d$, then

$$\liminf_{c_1,c_2\to 0} \frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = \frac{\Gamma(a+1/2)\Gamma(b+1/2)}{a^{1/2}\Gamma(a)b^{1/2}\Gamma(b)}. \tag{44}$$

(*b*) For any fixed b

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \to \frac{\Gamma(b+1/2)}{b^{1/2}\Gamma(b)}, \tag{45}$$

as $a \to \infty$, and as $a, b \to \infty$

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \to 1, \tag{46}$$

(*c*) If $a = b = c = d$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = \left(\frac{\Gamma(a+1/2)}{a^{1/2}\Gamma(a)}\right)^2 \left(\frac{2a+1}{2a}\right)^{1/2}. \tag{47}$$

(*d*) If $a = b = c = d = 1$, then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} = [\Gamma(3/2)]^2(3/2)^{1/2} = 0.9619. \tag{48}$$

(*e*) If $a, b \to 0$ then

$$\frac{r(\Delta_{\mathcal{S}})}{r(\Delta_{\mathcal{F}})} \to 0. \tag{49}$$

□

4.3 Comparison of two exponential means

We next consider the estimation of the difference of the means of two exponential populations with gamma priors. Let the distribution of X and Y be given by

$$f(x,\theta) = \theta e^{-\theta x}, \tag{50}$$

for $x > 0$, $\theta > 0$ and

$$g(y,\omega) = \omega e^{-\omega y}, \tag{51}$$

for $y > 0$, $\omega > 0$ respectively. We assume the prior distributions are $Gamma(a, p)$, $a > 2$,

$p > 0$ and $Gamma(c, q)$, $c > 2$, $q > 0$ respectively. The ratio of the sequential to the best fixed design is given by

$$\liminf_{c_1,c_2\to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{E(\theta^{-1}) + E(\omega^{-1})}{(E(\theta^{-2}))^{1/2} + (E(\omega^{-2}))^{1/2}}. \tag{52}$$

Therefore the ratio of the sequential to the best fixed design becomes

$$\liminf_{c_1,c_2\to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{\dfrac{p}{a-1} + \dfrac{q}{c-1}}{\dfrac{p}{((a-1)(a-2))^{1/2}} + \dfrac{q}{((c-1)(c-2))^{1/2}}}, \tag{53}$$

$a > 2$, $p > 0$, $c > 2$, $q > 0$.

If $a = p$ and $q = c$ then the ratio becomes

$$\liminf_{c_1,c_2\to 0} \frac{r(\Delta_S)}{r(\Delta_F)} = \frac{\dfrac{a}{a-1} + \dfrac{c}{c-1}}{\dfrac{a}{((a-1)(a-2))^{1/2}} + \dfrac{c}{((c-1)(c-2))^{1/2}}}, \tag{54}$$

$a > 2$, and $c > 2$.

4.4 Numerical comparisons

In this section we examine the ratio of the sequential to the best fixed designs for the estimation problem. We consider the case of balanced and unbalanced designs. This numerical study is conducted for the case of exponential distribution means with gamma priors and Poisson distribution means with gamma priors. For the balanced designs, $E(\theta) = E(\omega)$ and $\text{var}(\theta) = \text{var}(\omega)$, that is $a = c$ and $p = q$. Note that for the Poisson means with gamma priors with $a/p = c/q = k$, where $k > 0$ is fixed, the ratio $\frac{r(\Delta_S)}{r(\Delta_F)}$ is given by

$$\frac{r(\Delta_S)}{r(\Delta_F)} = \frac{\Gamma(a + 0.5)}{2a^{1/2}\Gamma(a)} + \frac{\Gamma(c + 0.5)}{2c^{1/2}\Gamma(c)}. \tag{55}$$

Table 1 gives the $\frac{r(\Delta_S)}{r(\Delta_F)}$ for the exponential distribution with gamma priors when $a = p$ and $c = q$.

In the tables below, we present the results of numerical comparisons of the best fixed and fully sequential procedures for several values of the parameters. The tables depict the efficiency for the balanced and unbalanced designs. The results are presented for the comparisons of Poisson means with gamma priors.

Table 1: *Relative efficiency when $a = p$ and $c = q$.*

a	2.0001	2.0010	2.0100	2.1000	10	50	100	200
2.0001	0.0100	0.0152	0.0181	0.0189	0.0155	0.0150	0.0150	0.0149
2.0010	0.0152	0.0316	0.0479	0.0563	0.0483	0.0470	0.0468	0.0468
2.0100	0.0181	0.0479	0.0995	0.1481	0.1461	0.1431	0.1428	0.1426
2.1000	0.0189	0.0562	0.1481	0.3015	0.4021	0.3979	0.3973	0.3971
10	0.0155	0.0483	0.1461	0.4021	0.9428	0.9647	0.9670	0.9680
50	0.0150	0.0470	0.1431	0.3979	0.9647	0.9900	0.9923	0.9936
100	0.0150	0.0468	0.1428	0.3973	0.9670	0.9923	0.9949	0.9962
200	0.0149	0.0468	0.1426	0.3971	0.9680	0.9936	0.9962	0.9975

Table 2: *Relative efficiency when $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$, $a = 2p$, $c = 2q$, and $a > c$.*

(a, c)	*Ratio*
$(2.0001, 2.000001)$	0.0048
$(2.001, 2.0005)$	0.0262
$(2.01, 2.005)$	0.0825
$(2.1, 2.05)$	0.2527
$(10, 5)$	0.9005
$(50, 10)$	0.9647
$(100, 50)$	0.9923
$(200, 100)$	0.9962

Table 3 gives the efficiency $\frac{r(\Delta_S)}{r(\Delta_{\mathcal{F}})}$ for the Poisson distribution with gamma priors when $a = p$ and $c = q$.

Table 3: *Relative efficiency when $a = p$ and $c = q$.*

a	10^{-10}	.001	.010	.100	1	10	50	100
10^{-10}	$2 * 10^{-5}$	.0280	.0874	.2475	.4431	.4938	.4988	.4994
0.001	.0280	.0560	.1154	.2755	.4711	.5218	.5267	.5274
0.010	.0874	.1154	.1748	.3349	.5305	.5812	.5868	.5868
0.100	.2475	.2755	.3349	.4950	.6906	.7413	.7463	.7469
1	.4431	.4711	.5305	.6906	.8862	.9369	.9419	.9425
10	.4938	.5218	.5812	.7413	.9369	.9876	.9925	.9932
50	.4988	.5267	.5868	.7463	.9419	.9925	.9975	.9981
100	.4994	.5274	.5868	.7469	.9425	.9933	.9981	.9988

The comparisons in Table 3 are for the balanced design.
Table 4 is given for $p = 10^{-10}$ with $a/p = c/q$ and $p > q$.
Table 6 gives the numerical values of the efficiency for $c = 4a$ and $q = 2p$.

Table 4: *Relative Efficiency when* $\mu_1 = \mu_2$ *and* $\sigma_1 < \sigma_2$.

(a, c)	*Ratio*
$(10^{-10}, 10^{-10})$	0.0000
$(0.0010, 10^{-5})$	0.0308
$(0.0100, 10^{-6})$	0.1333
$(0.1000, 10^{-7})$	0.4510
$(1.0000, 10^{-8})$	0.8591
$(10.0000, 10^{-9})$	0.9778
$(50.0000, 10^{-10})$	0.9931
$(100.0000, 10^{-10})$	0.9956

Table 5: *Relative Efficiency when* $\mu_1 < \mu_2$ *and* $\sigma_1 = \sigma_2$, $a = 2p$, $c = 2q$, *and* $a > c$.

(a, c)	*Ratio*
$(10^{-10}, 4 * 10^{-10})$	0.0000
$(0.0010, 0.0040)$	0.0885
$(0.0100, 0.0400)$	0.2694
$(0.1000, 0.0400)$	0.6513
$(1.0000, 4.0000)$	0.9349
$(10.0000, 40.0000)$	0.9930
$(50.0000, 200.0000)$	0.9903
$(100.0000, 400.0000)$	0.9993

Table 6: *Relative Efficiency when* $\mu_1 < \mu_2$ *and* $\sigma_1 < \sigma_2$, $a < c$, *and* $p < q$.

(a, c)	*Ratio*
$(10^{-10}, 4 * 10^{-10})$	0.0000
$(0.0010, 0.0040)$	0.1039
$(0.0100, 0.0400)$	0.3142
$(0.1000, 0.0400)$	0.6793
$(1.0000, 4.0000)$	0.9436
$(10.0000, 40.0000)$	0.9940
$(50.0000, 200.0000)$	0.9989
$(100.0000, 400.0000)$	0.9991

5 Concluding remarks

We have shown that the sequential procedure for the problem of estimating the difference of the means of two independent populations from the exponential family with conjugate priors when compared with the best fixed design reveal the superiority of the random design. The lower bound for the Bayes risk plus the expected costs

determined. Application of the results to the Poisson and exponential distributions using gamma priors as well as the Bernoulli distribution with beta priors are given. Numerical comparisons of the best fixed and fully sequential procedures for several values of the parameters conducted. There are other random designs that are of interest including the two stage design, and the myopic design (see Terbeche, 2000). These designs seem to perform better than the best fixed design.

Acknowledgement

The authors wish to thank the referees for constructive criticisms which lead to substantial improvement of the paper.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, New York: Springer-Verlag.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. *Ann. Statist.*, 6, 269-281.

Hajek, J. & Sidak. Z. (1967). *Theory of Rank Tests*, New York-London: Academic Press.

Lehmann, E.L. (1959). *Testing Statistical Hypothesis*, New York: John Wiley and Sons Inc.

Terbeche, M. (2000). *Sequential Design for Estimation*, Thesis, Florida Institute of Technology.

West, M. (1985). *Generalized Linear Models: Outlier Accommodation, Scale Parameters and Prior Distributions*. Bayesian Statistics 2, J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith (Eds), North-Holland, Amsterdam and Valencia University Press.

West, M. (1986). Bayesian Model Monitoring. *J. Roy. Statist. Soc.*, (Ser. B) 28, 70-78.

SORT 29 (2) July-December 2005, 235-248

On the probability of reaching a barrier in an Erlang(2) risk process

M.M. Claramunt[1], M. Mármol[1] and R. Lacayo[2]

[1]*Universitat de Barcelona,* [2]*Universitat Autònoma de Barcelona*

Abstract

In this paper the process of aggregated claims in a non-life insurance portfolio as defined in the classical model of risk theory is modified. The Compound Poisson process is replaced with a more general renewal risk process with interoccurrence times of Erlangian type. We focus our analysis on the probability that the process of surplus reaches a certain level before ruin occurs, $\chi(u, b)$. Our main contribution is the generalization obtained in the computation of $\chi(u, b)$ for the case of interoccurrence time between claims distributed as Erlang(2, β) and the individual claim amount as Erlang (n, γ).

MSC: 91B30, 62P05

Keywords: risk theory, Erlang distribution, upper barrier, ordinary differential equation, boundary conditions.

1 Introduction

Ruin theory is concerned basically with the study of the insurer's solvency through the analysis of the level of reserves as a function of time and other important aspects such as the probability of ruin, the time of ruin and the severity of ruin.

*Address for correspondence:*M.M. Claramunt, M. Mármol. Departament de Matemàtica Econòmica, Financera i Actuarial. Universitat de Barcelona. Avda. Diagonal, 690. 08034 Barcelona. Tlf: 93.4035744. E-mail: mmclaramunt@ub.edu, mmarmol@ub.edu.

R. Lacayo. Departament de Matemàtiques. Universitat Autònoma de Barcelona. Edifici C. 08193 Bellaterra. Barcelona. Tlf. 93.5812534. E-mail: rlacayo@mat.uab.es.

Received: October 2004
Accepted: July 2005

One of the most important probabilities related with ruin is the probability that the process of surplus reaches a certain level before ruin occurs (Dickson and Gray (1984), Dickson (1992) and Dickson and Egidio dos Reis (1994) analyzed this probability in the classical risk model). The aim of this paper is the study of this probability, $\chi(u,b)$.

In this paper the Poisson number process of the classical risk model is replaced with a more general renewal risk process with interoccurrence times of Erlangian type (see, e.g., Dickson and Hipp (1998, 2001), Dickson (1998), Cheng and Tang (2003), Sun and Yang (2004), Albrecher *et al.* (2005)). Dickson (1998) analyzed $\chi(u,b)$ for the particular case in which the interoccurrence times between claims are distributed as Erlang(2,2) and the individual claim amount has also an Erlang(2,2) distribution. Our main contribution in this paper is the generalization obtained in the computation of $\chi(u,b)$ for the case of interoccurrence time distributed as Erlang$(2,\beta)$ and the individual claim amount as Erlang (n,γ). Note that the Erlang distribution is a special case of the Gamma distribution where the shape parameter n is a positive integer.

The organization of the paper is as follows. In Section 2 we summarize the main results related to $\chi(u,b)$ in the classical risk model. In Section 3, we obtain an integro-differential equation for $\chi(u,b)$ in an ordinary Erlangian$(2,\beta)$ model, i.e. with interoccurrence time Erlang $(2,\beta)$. In Section 3.1 we obtain and solve the corresponding differential equation for $\chi(u,b)$ assuming a general Erlang(n,γ) distribution for the individual claim amount. In Section 3.2 we provide numerical results for the particular case when the individual claim amount is distributed as an Erlang$(2,\gamma)$, and in Section 3.3 for the case of an Erlang $(1,\gamma)$, i.e. exponential(γ) distribution. In Section 3.4 we analyze the influence of the individual claim amount distribution on $\chi(u,b)$ by comparing the numerical results.

2 Classical model

In the classical model of risk theory, the surplus, $R(t)$, at a given time $t \in [0,\infty)$ is defined as $R(t) = u + ct - \sum_{i=1}^{N(t)} X_i$, with $u = R(0)$ being the insurer's initial surplus. $N(t)$, the number of claims occurred until time t, follows a Poisson process with parameter λ, and X_i is the amount of the *i-th* claim and has density function $f(x)$ with mean μ. The instantaneous premium rate, c, is $c = \lambda\mu(1+\rho)$, where ρ, called the security loading, is a positive constant.

In this model, and in the more general ordinary renewal model, the interoccurrence time between claims, T_i, $i = 1,2,...$ is modeled as a sequence of independent and identically distributed random variables. T_1 denotes the time until the first claim and, in general, T_i denotes the time between the $i-1$-*th* and *i-th* claims. Note that in a Poisson process with parameter λ, $T_i, i \geq 1$ has an exponential distribution with mean $\frac{1}{\lambda}$.

Given that the time of the first claim, T_1, follows an exponential distribution with density function $f_{T_1}(t) = \lambda e^{-\lambda t}$, the probability $\chi(u,b)$ that the surplus process reaches

the level $b > u$ before the time until ruin, defined as $\tau = \inf\{t : R(t) < 0\}$, can be obtained as

$$\chi(u,b) = \int_0^{t_0} \lambda e^{-\lambda t} \int_0^{u+ct} \chi(u+ct-x,b)\, f(x)\, dx dt + \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt, \tag{1}$$

where $u + ct_0 = b$, so that the surplus process will reach b at time t_0 if no claims occur by time t_0 (Dickson and Gray, 1984).

The function $\chi(u,b)$ has also been related with ruin probabilities. The probability of ultimate ruin is defined as

$$\psi(u) = P[R(t) < 0 \text{ for some } t > 0],$$

and $\delta(u) = 1 - \psi(u)$ denotes the survival probability. It can be proved that (Dickson and Gray, 1984),

$$\delta(u) = \chi(u,b)\,\delta(b). \tag{2}$$

It is clear then that $\chi(u,b)$ can be computed as this ratio of survival probabilities as an alternative to using expression (1).

In a model with upper absorbing barrier b, such that when the reserve level reaches this barrier the process is finished, the quantity $1 - \chi(u,b)$ is also, by definition, the probability of ruin given that the initial reserve is u.

$\chi(u,b)$ plays also an important role in the model with a constant dividend barrier. In this model whenever the surplus reaches the level b, dividends are paid out in such amount that surplus stays at the barrier until the next claim. Obviously, the present value of the dividends paid out, $D(u,b)$, is a random variable that has a non-null probability at zero. This is the probability that dividends paid out are zero (Mármol *et al.*, 2003), i.e.,

$$P[D(u,b) = 0] = 1 - \chi(u,b).$$

3 Ordinary renewal model with $T_i \sim$ Erlang $(2,\beta)$

The classical Poisson risk model is an ordinary renewal process, with $T_i \sim Erlang(1,\lambda)$. In this section we assume that the number of claims is an ordinary renewal process in which the T_i are i.i.d. Erlang$(2,\beta)$ with density function,

$$k(t) = \beta^2 t e^{-\beta t}, \quad t > 0, \tag{3}$$

and distribution function

$$K(t) = 1 - e^{-\beta t}(\beta t + 1) \qquad \text{for } t \geq 0.$$

Then, as in expression (1), in Dickson (1998) it is obtained

$$\chi(u,b) = \int_0^{t_0} k(t) \int_0^{u+ct} \chi(u + ct - x, b)\, f(x)\, dx dt + \int_{t_0}^{\infty} k(t)\, dt. \tag{4}$$

Substituting $s = u + ct$ in (4) and differentiating twice with respect to u,

$$c^2\chi''(u,b) - 2\beta c\chi'(u,b) + \beta^2\chi(u,b) = \beta^2 \int_0^u \chi(u - x, b) f(x)\, dx. \tag{5}$$

Notice that equation (2), which in the classical model relates the survival probability with $\chi(u,b)$, is not true in the Ordinary Erlangian $(2,\beta)$ model because the lack of memory property is exclusive of the Exponencial distribution and does not hold for the general Erlang distribution (Dickson, 1998). As a result, $\chi(u,b)$ cannot be obtained as a ratio of survival probabilities, and for its calculation expression (5) must be used.

From (5) we obtain and solve the differential equation assuming that the individual claim amount is Erlang(n,γ), following the procedure presented by Dickson (1998).

3.1 Individual claim amount Erlang (n, γ)

In this section we assume that the individual claim amount follows an Erlang(n,γ) distribution with pdf

$$f(x) = \frac{\gamma^n x^{n-1} e^{-\gamma x}}{(n-1)!}. \tag{6}$$

To solve (5) let us define

$$h(u) = \beta^2 \int_0^u \chi(x,b) f(u - x)\, dx. \tag{7}$$

Substituting (6) in (7) yields

$$h(u) = \frac{\beta^2 \gamma^n e^{-\gamma u}}{(n-1)!} \int_0^u \chi(x,b)\,(u - x)^{n-1} e^{\gamma x} dx.$$

Later on we will need an expression for the n-th derivative of the function above in terms of the lower order derivatives. This result is the essence of the following lemma.

Lemma 1 *The n-th derivative of the function* $h(u)$ *is given by*

$$h^{(n)}(u) = -\sum_{i=0}^{n-1} \binom{n}{i} h^{(i)}(u)\,\gamma^{n-i} + \beta^2\gamma^n \chi(u,b) \tag{8}$$

(see the proof in Appendix A)

After rewriting equation (5) in the form

$$c^2\chi''(u,b) - 2\beta c\chi'(u,b) + \beta^2\chi(u,b) = h(u),$$

it is clear that after differentiating i and n times, respectively, we obtain

$$c^2\chi^{(i+2)}(u,b) - 2\beta c\chi^{(i+1)}(u,b) + \beta^2\chi^{(i)}(u,b) = h^{(i)}(u), \tag{9}$$

and

$$c^2\chi^{(n+2)}(u,b) - 2\beta c\chi^{(n+1)}(u,b) + \beta^2\chi^{(n)}(u,b) = h^{(n)}(u). \tag{10}$$

Substitution in (10) of the value of $h^{(n)}(u)$ found in (8) and the value of $h^{(i)}(u)$ from (9) yields the following ordinary differential equation of order $(n+2)$ for $\chi(u,b)$:

$$a_{n+2}\chi^{(n+2)}(u,b) + a_{n+1}\chi^{(n+1)}(u,b) + a_n\chi^{(n)}(u,b) - \sum_{j=1}^{n-1} a_j\chi^{(j)}(u,b) = 0. \tag{11}$$

The value of the constant coefficients is given by

$$a_{n+2} = c^2$$

$$a_{n+1} = c^2\gamma n - 2\beta c$$

$$a_n = \beta^2 - 2\beta c\gamma n + \binom{n}{n-2}c^2\gamma^2$$

$$a_j = -c^2\binom{n}{j-2}\gamma^{n+2-j} + \binom{n}{j-1}2\beta c\gamma^{n+1-j} - \beta^2\binom{n}{j}\gamma^{n-j},\ \ j = 1,\ldots,n-1.$$

If all the roots of the characteristic equation of (11), $\{r_i\}_{i=0}^{n+1}$, are different, it is a trivial matter to write down the solution for the ordinary differential equation above, namely:

$$\chi(u,b) = \sum_{i=0}^{n+1} \alpha_i e^{r_i u}, \tag{12}$$

where $\{r_i\}_{i=0}^{n+1}$ are functions of γ, β and c, while the $\{\alpha_i\}_{i=0}^{n+1}$ depend additionally on b. To obtain the values of $\{\alpha_i\}_{i=0}^{n+1}$ we need $(n+2)$ equations.

The first of them is obtained from the boundary condition $\chi(b,b) = 1$. Then,

$$\sum_{i=0}^{n+1} \alpha_i e^{r_i b} = 1. \tag{13}$$

From (12) we know $\chi'(u,b)$ and $\chi''(u,b)$. Substituting in (5), after rearranging terms, one easily obtains n equations, namely,

$$\sum_{i=0}^{n+1} \frac{\alpha_i}{(r_i+\gamma)^s} = 0 \quad , \; s = 1, \ldots, n. \tag{14}$$

From (4) , differentiating with respect to u, and considering (12) and its first and second derivatives, we obtain the last equation

$$1 = \alpha_0 + \frac{1}{\beta} \sum_{i=1}^{n+1} \alpha_i \left(\beta - c r_i\right) e^{r_i b}. \tag{15}$$

Consequently, after the combination of (13) , (14) and (15), we obtain the required set of $(n+2)$ equations from which to calculate the coefficients $\{\alpha_i\}_{i=0}^{n+1}$. They are

$$\begin{cases} \sum\limits_{i=0}^{n+1} \alpha_i e^{r_i b} = 1 \\ \sum\limits_{i=0}^{n+1} \dfrac{\alpha_i}{(r_i+\gamma)^s} = 0 \quad , s = 1, 2, \ldots, n \\ \frac{1}{\beta} \sum\limits_{i=0}^{n+1} \alpha_i \left(\beta - c r_i\right) e^{r_i b} = 1. \end{cases} \tag{16}$$

3.2 $T_i \sim$ Erlang $(2,\beta)$ and $X \sim$ Erlang $(2,\gamma)$

In this section we study the case $n = 2$. The ODE can be obtained directly from (11) as

$$\begin{gathered} c^2 \chi''''(u,b) + \left(2\gamma c^2 - 2\beta c\right) \chi'''(u,b) + \\ \left(\beta^2 - 4\gamma\beta c + \gamma^2 c^2\right) \chi''(u,b) + \left(2\gamma\beta^2 - 2\gamma^2 \beta c\right) \chi'(u,b) = 0. \end{gathered} \tag{17}$$

This equation generalizes the one obtained by Dickson (1998) for the particular case $c = 1.1, \beta = 2$ and $\gamma = 2$.

The solution of (17) gives

$$\chi(u,b) = \sum_{i=0}^{3} \alpha_i e^{r_i u},$$

where $\{r_i\}_{i=0}^{3}$ are the roots of the characteristic equation of (17). From (16), the system of equations required to find $\{\alpha_i\}_{i=0}^{3}$ is,

$$\begin{cases} \sum_{i=0}^{3} \alpha_i e^{r_i b} = 1 \\ \sum_{i=0}^{3} \dfrac{\alpha_i}{(r_i+\gamma)} = 0 \\ \sum_{i=0}^{3} \dfrac{\alpha_i}{(r_i+\gamma)^2} = 0 \\ \dfrac{1}{\beta} \sum_{i=0}^{3} \alpha_i (\beta - c r_i) e^{r_i b} = 1. \end{cases} \tag{18}$$

For $\gamma = 2, \beta = 2$, and $c = 1.1$, solving (18) and finding the roots of the characteristic equation (17), we obtain in Table 1 the results for $\chi(u,b)$ for different values of u and b,

Table 1.

u/b	0	1	2	3	4	5
0	1	0.5802	0.3694	0.2805	0.2335	0.2049
1		1	0.7600	0.5828	0.4854	0.4258
2			1	0.8472	0.7096	0.6228
3				1	0.8939	0.7875
4					1	0.9224
5						1

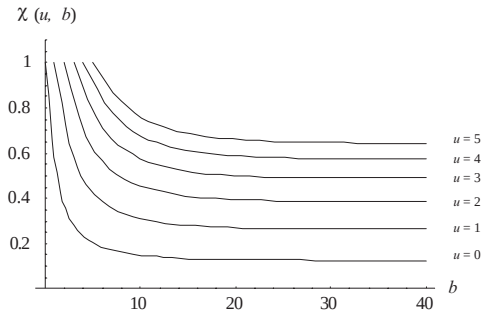

Figure 1: $\chi(u,b)$ *for* $u = 0, 1, 2, 3, 4, 5$.

Graphically we can represent the evolution of $\chi(u,b)$ with respect to b for different values of the initial surplus u. In Figure 1 $\chi(u,b)$ is plotted for $u = 0,1,2,3,4,5$. For a given value of b, the probability of the reserves reaching that value before ruin, $\chi(u,b)$, is increasing in u. On the other hand, for a given value of u, the probability $\chi(u,b)$ is decreasing in b, and for each value of u tends toward a limiting value, as shown in Table 2 below.

Table 2.

u	0	1	2	3	4	5
$\lim_{b\to\infty}\chi(u,b)$	0.1268	0.2636	0.3855	0.4876	0.5727	0.6438

Obviously, as b tends to infinity, the probability $\chi(u,b)$ includes only those trajectories of the reserve process which do not lead to ruin. In other words,

$$\lim_{b\to\infty}\chi(u,b) = \delta(u). \tag{19}$$

Consequently, the limiting values for $\chi(u,b)$ just obtained are the values of the survival probability for the corresponding initial reserves u (they can be found in the discussion section written by De Vylder and Goovaerts in Dickson (1998)).

3.3 $T_i \sim$ Erlang $(2,\beta)$ and $X \sim \exp(\gamma)$

Here we study the case $n = 1$. The corresponding ODE, from (11) is

$$c^2\chi'''(u,b) + \left(\gamma c^2 - 2\beta c\right)\chi''(u,b) + \left(\beta^2 - 2\beta\gamma c\right)\chi'(u,b) = 0,$$

with solution

$$\chi(u,b) = \alpha_0 + \sum_{i=1}^{2}\alpha_i e^{r_i u},$$

where r_1 and r_2 are the roots of

$$c^2 r^2 + \left(\gamma c^2 - 2\beta c\right)r + \left(\beta^2 - 2\beta\gamma c\right) = 0.$$

In order to obtain $\{\alpha_i\}_{i=0}^{2}$, we put $n = 1$ in (16),

$$\begin{cases} \sum_{i=0}^{2} \alpha_i e^{r_i b} = 1 \\ \sum_{i=0}^{2} \dfrac{\alpha_i}{(r_i + \gamma)} = 0 \\ \frac{1}{\beta} \sum_{i=0}^{2} (\beta - c r_i) \cdot \alpha_i \cdot e^{r_i \cdot b} = 1. \end{cases}$$

For $\gamma = 1, \beta = 2$, and $c = 1.1$, we obtain in Table 3 the following results of $\chi(u, b)$

Table 3.

u/b	0	1	2	3	4	5	⋯	∞
0	1	0.6363	0.4318	0.3339	0.2779	0.2419	⋯	0.1199
1		1	0.7838	0.6106	0.5083	0.4425	⋯	0.2194
2			1	0.8518	0.7125	0.6204	⋯	0.3076
3				1	0.8906	0.7781	⋯	0.3858
4					1	0.9155	⋯	0.4552
5						1	⋯	0.5168

The behaviour of this probability in this case turns out to be the same as in Section 2.1 where the Erlang$(2, \gamma)$ distribution was assumed.

3.4 Numerical comparison

In order to study the influence of the distribution of the claim amount on the probability $\chi(u, b)$ we find the behaviour of the latter when the individual claim amount follows an Erlang$(n, \gamma), n = 1, 2, 3, 4, 5$ distribution. To ensure that the results can be compared to one another we set $n = \gamma$ and call the resulting distribution simply an Erlang(n). Note that in this case the mean of the claim n/γ is 1.

For $n = 1$ and $n = 2$ the probability $\chi(u, b)$ behaves as indicated in Sections 2.3 and 2.2, respectively, i.e., takes the value 1 for $u = b$, and for a fixed u, it is decreasing in b and tends to a limiting value which is the same as the survival probability in the model without a barrier.

The effect of the claim amount distribution on $\chi(u, b)$ depends, as expected, on the initial reserve and barrier levels u and b, and on the difference $u-b$ as well. The following three figures show the behavior of $\chi(u, b)$ as a function of b for initial reserve levels $u = 0, u = 1$ and $u = 2$, respectively. For the given u the graphs in each figure show the

dependence of $\chi(u,b)$ on the Erlang parameter n for $n = 1,2,3,4,5$. These values have been chosen for illustration only, and carry no special significance.

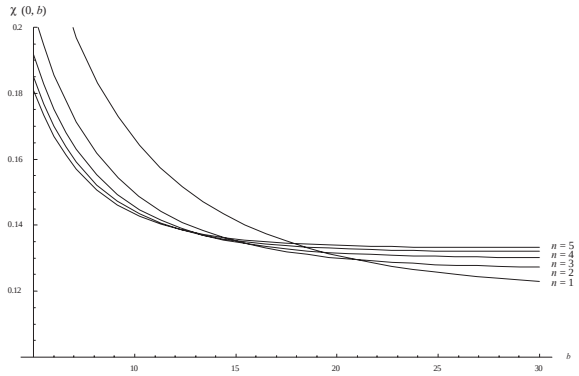

Figure 2: $\chi(u,b)$ *for* $u = 0$, *assuming* $n = 1,2,3,4,5$.

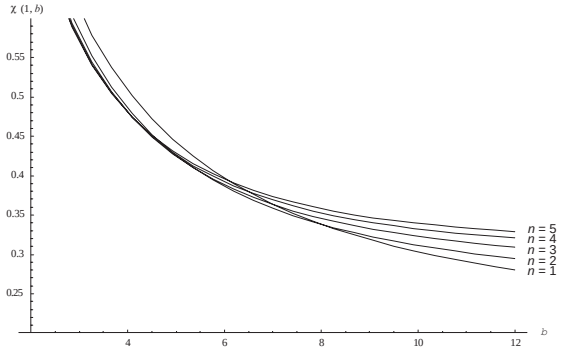

Figure 3: $\chi(u,b)$ *for* $u = 1$, *assuming* $n = 1,2,3,4,5$.

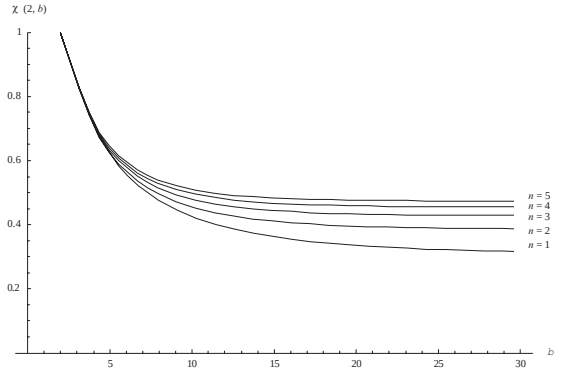

Figure 4: $\chi(u,b)$ *for* $u = 2$, *assuming* $n = 1,2,3,4,5$.

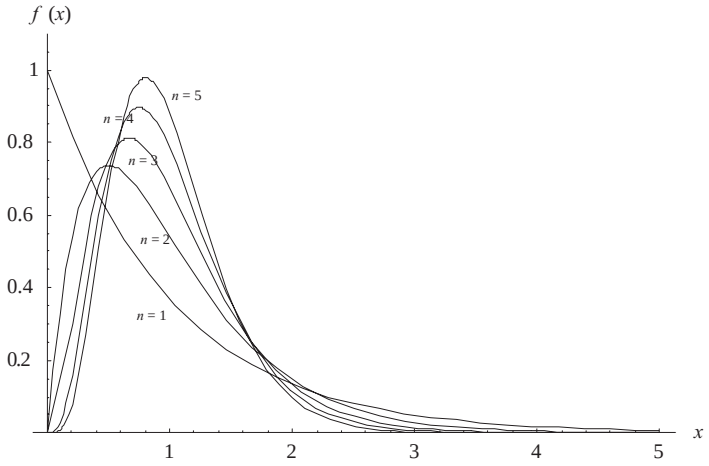

Figure 5: *pdf Erlang*(*n*) *for* $u = 1$, *assuming* $n = 1, 2, 3, 4, 5$.

Before analyzing our results further, in Figure 5 we provide the graphs of an Erlang(n) pdf with mean $E[\cdot] = 1$ and variance $Var[\cdot] = \frac{1}{n}$, also for $n = 1, 2, 3, 4, 5$. It is clear from the figure that with increasing n both the variance and the asymmetry decrease, and the pdf concentrates more and more around its mean 1.

Moreover, from Figures 2, 3 and 4, it follows that for values of u near zero and small b, $\chi(u, b)$ decreases as n increases. This behaviour is reversed as b grows larger (the graphs intersect at different points, and eventually those corresponding to larger n appear on top). A plausible explanation for this behaviour can be found in Figure 5, from which we see that for small n (recall that $n = 1$ coincides with the exponential case) the probability of occurrence of small and large claims is greater than that corresponding to large n. As a consequence, for values of u near zero and small b, the probability of reaching b before ruin occurs is greater for small n. For $b \gg u$, large claims take preponderance in reaching the ruin state and they are more likely for small n, thus $\chi(u, b)$ is smaller for small n.

Table 4.

$\delta(u)$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
$\delta(0)$	0.1199	0.1268	0.1300	0.1319	0.1332
$\delta(1)$	0.2194	0.2636	0.2882	0.3041	0.3153
$\delta(2)$	0.3076	0.3855	0.4282	0.4552	0.4738
$\delta(3)$	0.3858	0.4876	0.5409	0.5736	0.5956
$\delta(4)$	0.4552	0.5727	0.6314	0.6663	0.6892
$\delta(5)$	0.5168	0.6438	0.7041	0.7388	0.7612

As u increases, the inversion process with increasing b disappears rapidly. In fact the graphs intersect very close to the initial abscissa b. This fact may be taken to mean that for initial reserves of substantial magnitude, the greater probability of small claims for small n loses relevance.

Below, in Table 4, we provide an additional table with the survival probabilities for all cases $n = 1, 2, 3, 4, 5$ (recall expression (19)). Note that they represent the survival probability in the absence of a barrier. In our case, as the table clearly shows, in the limit $\chi(u, b)$ decreases with increasing n in accordance with the results above.

Appendix A

Proof of Lemma 1.

Since the function $h(u)$ depends explicitly on n, for notational convenience we rewrite it as

$$\begin{aligned} h_n(u) &= \frac{\beta^2 \gamma^n e^{-\gamma u}}{(n-1)!} \int_0^u \chi(x,b)\,(u-x)^{n-1}\, e^{\gamma x} dx \\ &= A_n e^{-\gamma u} \int_0^u B\,(u-x)^{n-1}\, dx, \end{aligned} \tag{20}$$

where, $A_n = \dfrac{\beta^2 \gamma^n}{(n-1)!}$ and $B = \chi(x,b)e^{\gamma x}$.
For $n = 1$ we have

$$h_1(u) = \beta^2 \gamma e^{-\gamma \upsilon} \int_0^u \chi(x,b) e^{\gamma x} dx,$$

and it readily follows through differentiation and substitution that

$$h_1'(u) = -\gamma h_1(u) + \beta^2 \gamma \chi(u,b) \tag{21}$$

For $n > 1$, differentiating (20) once yields

$$h_n'(u) = A_n e^{-\gamma u}\,(n-1) \int_0^u B\,(u-x)^{n-2}\, dx - \gamma A_n e^{-\gamma u} \int_0^u B\,(u-x)^{n-1}\, dx,$$

which, after dropping the argument u, can be written as the recurrence relation

$$h_n' = -\gamma h_n + \gamma h_{n-1} \quad , n > 1 \tag{22}$$

Obviously we can obtain all the required derivatives of $h_n(u)$ by successive differentiation of (22).

From (22) we have

$$\begin{aligned} h_n'' &= -\gamma h_n' + \gamma h_{n-1}' = -\gamma h_n' + \gamma\left(-\gamma h_{n-1} + \gamma h_{n-2}\right) \\ &= -\gamma h_n' - \gamma\left(h_n' + \gamma h_n\right) + \gamma^2 h_{n-2} \\ &= -2\gamma h_n' - \gamma^2 h_n + \gamma^2 h_{n-2}. \end{aligned} \tag{23}$$

In the first line we used (22) to obtain h_{n-1}', and again in the second line to obtain γh_{n-1}.

In the same manner, from (23) we get

$$\begin{aligned} h_n''' &= -2\gamma h_n'' - \gamma^2 h_n' + \gamma^2 h_{n-2}' = -2\gamma h_n'' - \gamma^2 h_n' + \gamma^2\left(-\gamma h_{n-2} + \gamma h_{n-3}\right) \\ &= -2\gamma h_n'' - \gamma^2 h_n' - \gamma\left(h_n'' + 2\gamma h_n' + \gamma^2 h_n\right) + \gamma h_{n-3}\Big) \\ &= -3\gamma h_n'' - 3\gamma^2 h_n' - \gamma^3 h_n + \gamma^3 h_{n-3}. \end{aligned} \tag{24}$$

Here we used in the first line (22) to obtain h_{n-2}', and in the second (23) to obtain $\gamma^2 h_{n-2}$.

It is clear that in the fourth derivative we would have to make use of (22) and of (24) and, in general, each derivative requires the use of (22) and of the previous one. Moreover, it follows easily from (22), (23) and (24), after transposing, that the rightmost term in the derivative of order k can be formally written as the k-th derivative of the product γh_n (this follows from Leibniz formula) provided the derivatives of γ are interpreted as regular powers. In other words, recalling that $h_n^{(0)} = h_n$,

$$\gamma^k h_{n-k} = (\gamma h_n)^{(k)} = \sum_{j=0}^{k} \binom{k}{j} \gamma^j h_n^{(k-j)}, \ k < n. \tag{25}$$

In particular, for $k = n - 1$, (25) becomes

$$\gamma^{n-1} h_1 = (\gamma h_n)^{(n-1)} = \sum_{j=0}^{n-1} \binom{n-1}{j} \gamma^j h_n^{(n-1-j)}. \tag{26}$$

Note that in (21) h_1' is given in terms of h_1, that is, in terms of the summation appearing in (26). If we now differentiate (26) and make the appropriate substitutions, after transposing we obtain

$$\beta^2 \gamma^n \chi(u,b) = \sum_{j=0}^{n-1} \binom{n-1}{j} \gamma^j h_n^{(n-j)} + \sum_{j=0}^{n-1} \binom{n-1}{j} \gamma^{j+1} h_n^{(n-1-j)}.$$

In the first summation above we can safely change the upper limit to n because $\binom{n-1}{n} = 0$. In the second summation, by substitution of the dummy variable j with $j - 1$ the summation limits are changed from $j = 1$ to $j = n$. But since $\binom{n-1}{-1} = 0$, we put the lower limit at $j = 0$. Combining the resulting summations and given that $\binom{n-1}{j} + \binom{n-1}{j-1} = \binom{n}{j}$ we

finally obtain

$$\beta^2\gamma^n\chi(u,b) = \sum_{j=0}^{n}\binom{n}{j}\gamma^j h_n^{(n-j)}.$$

The equivalence of this expression to (8) is evident. The proof of the Lemma is complete.

References

Albrecher, H., Claramunt, M.M., Mármol, M. (2005). On the distribution of dividend payments in a Sparre Andersen model with generalized Erlang(n) interclaim time. *Insurance: Mathematics and Economics*. To appear.

Cheng, Y., Tang, Q. (2003). Moments of the surplus before ruin and the deficit at ruin in the Erlang(2) risk process. *North American Actuarial Journal*, 7, 1-12.

Dickson, D.C.M., Gray, J.R. (1984). Approximations to ruin probability in the presence of an upper absorbing barrier. *Scandinavian Actuarial Journal*, 105-115.

Dickson, D.C.M. (1992). On the distribution of the surplus prior to ruin. *Insurance: Mathematics and Economics*, 11, 197-207.

Dickson, D.C.M. (1998). On a class of renewal risk process. *North American Actuarial Journal*, 2 (3), 60-73.

Dickson, D.C.M., Egídio dos Reis, A.D. (1994). Ruin problems and dual events. *Insurance: Mathematics and Economics*, 14, 51-60.

Dickson, D.C.M, Hipp, C. (1998). Ruin probabilities for Erlang(2) risk process. *Insurance: Mathematics and Economics*, 22, 251-262.

Dickson, D.C.M., Hipp, C. (2001). On the time to ruin for Erlang(2) risk process. *Insurance: Mathematics and Economics*, 29, 333-344.

Mármol, M., Claramunt, M.M., Alegre, A. (2003). Reparto de dividendos en una cartera de seguros no vida. Obtención de la barrera constante óptima bajo criterios económico-actuariales. *Documents de treball de la Divisió de Ciències Jurídiques, Econòmiques i Socials*, E03/99.

Sun, L., Yang, H. (2004). On the joint distributions of surplus immediately before ruin and the deficit at ruin for Erlang(2) risk processes. *Insurance: Mathematics and Economics*, 34, 121-125.

SORT 29 (2) July-December 2005, 249-268

Factorial experimental designs and generalized linear models

S. Dossou-Gbété and W. Tinsson

Université de Pau et des Pays de l'Adour

Abstract

This paper deals with experimental designs adapted to a generalized linear model. We introduce a special link function for which the orthogonality of design matrix obtained under Gaussian assumption is preserved. We investigate by simulation some of its properties.

MSC: 62J12, 62K15

Keywords: generalized linear model, exponential family, Fisher-Scoring algorithm, factorial designs, regular fraction.

1 Introduction

Experimental designs are usually used in a linear context, *i.e.* assuming that the mean response can be correctly fitted by a linear model (polynomial of degree one or two in most cases). This assumption is often associated with the normality of the observed responses. Some classical and efficient experimental designs are then well known in this context (see the books of Box and Draper (1987) or Khuri and Cornell (1996)). However, it is clear that these linear assumptions are inadequate for some applications.

Many books and papers deal with the question of relaxing the linear model and the gaussian model framework (see, for example, chapter 10 of the book of Khuri and Cornell (1996) for a synthesis). But there are two main difficulties with this approach. First, the choice of a good nonlinear model is not always easy. Second, assuming the

Address for correspondence: S. Dossou-Gbété and W. Tinsson Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées, avenue de l'Université-64000 Pau-France.

Received: March 2004
Accepted: November 2005

nonlinear model is given, using a classical design (factorial, central composite, *etc.*) is not in general the best choice. This fact can be problematic when industrial results are first obtained with a classical design. If a linear model turns out to be inappropriate it is then impossible in general to make new experiments because they are too expensive.

Our goal in this paper is to propose another class of solutions. These solutions have to be, on the one hand, more general than the linear case and the gaussian framework and, on the other hand, easier to improve than nonlinear modeling.

This intermediate solution consists of the choice of a generalized linear model (see, for example, McCullagh and Nelder (1989) or Green and Silverman (1994)). In other words, we assume that the image of the mean response by a given "link function" can be modelled *via* a linear relationship. Such an assumption allows us to consider any responses with a distribution in the exponential family (Bernoulli, binomial, Poisson, Gamma, *etc.*) and then we do not have the restrictions of the classical linear case. These models have been studied in order to construct D-optimal designs (see the book of Pukelsheim (1993) for the general problem of optimality). The main problem of this approach is the fact that the information matrix depends on the unknown parameters of the model. Some authors have then developed Bayesian methods (see Chaloner and Larntz (1989)) or robust designs (see Chipman and Welch (1996) or Sebastiani and Settimi (1997)) but these are available only for logistic regression. Our goal in this paper is to propose a general method of analysis with a simple information matrix, independent of the parameters of the model. When there is no prior knowledge the canonical link function is classically used for the modelization of the mean. We prove in the following that if we use the alternative choice of an appropriate link function, called the *surrogate function*, then classical factorial designs can be advantageously used.

Our paper is organized as follows. Section 2 is devoted to notations and preliminary results concerning the generalized linear model, the Fisher scoring algorithm and factorial designs. Section 3 makes a link between these methods and the choice of an experimental design. At the end we present an example of application.

2 Experimental designs and GLM

2.1 The generalized linear model

We consider in the following a generalized linear model as it was introduced by Nelder and Wedderburn (1972). Suppose that we have n observed responses y_i $(i = 1, \ldots, n)$ associated with the independent random variables Y_i having the same distribution, a member of exponential family. Denoting $m_i = E(Y_i)$, we then have a generalized linear model if and only if:

$$\forall\ i = 1, \ldots, n\ ,\ g(m_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i \in \mathbb{R}^r$ is the vector of independant variables, $\boldsymbol{\beta} \in \mathbb{R}^r$ is the vector of unknown parameters of the model and g is the link function (assumed to be bijective and differentiable). Because Y_i $(i = 1, \ldots, n)$ is a member of exponential family we have the following classe of density functions:

$$f(y_i, \theta_i, \phi) = h(y_i, \phi) \exp\left(\frac{y_i\theta_i - v(\theta_i)}{\phi}\right) \text{ with } \phi \text{ known.} \tag{1}$$

We say that θ_i is the canonical parameter of the distribution (associated with Y_i) and that ϕ is a dispersion parameter. It is usual to use the canonical link function which means that:

$$\forall\ i = 1, \ldots, n\ ,\ g(m_i) = \theta_i.$$

Recall that for every element of an exponential family we have the following relations:

$$E(Y_i) = m_i = v'(\theta_i) \text{ and } \operatorname{Var}(Y_i) = \phi v''(\theta_i)\,. \tag{2}$$

Hence we can write $\operatorname{Var}(Y_i) = V(m_i)$ with $V(m_i) = \phi m_i'(\theta_i)$.

Example 1 Consider the common case of binary responses. Every observed response y_i is then a realization of a Bernoulli distribution of parameter p_i (unknown in most cases). Such a distribution belongs to the exponential family because its density satisfies relation (1) with :

$$\theta_i = \ln\frac{p_i}{1-p_i}\ ,\ v(\theta_i) = -\ln\left(1 + e^{\theta_i}\right)\ ,\ \phi_i = 1 \text{ and } h(y_i, \phi_i) = \mathbb{I}_{\{0,1\}}(y_i, \phi_i)\,.$$

For the function V and the canonical link function we have $m_i = p_i$ and $\operatorname{Var}(Y_i) = p_i(1-p_i)$ so:

$$V(t) = t(1-t) \text{ and } g(t) = \ln\frac{t}{1-t}.$$

2.2 Estimation of the parameters

For a given generalized linear model, our problem is then to estimate the unknown parameters for the specification of the mean. Using the maximum likelihood method, our goal is then to maximize the likelihood of the sample or (equivalently) its logarithm, that is:

$$L(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{y_i \theta_i - v(\theta_i)}{\phi_i} + \sum_{i=1}^{n} \ln(h(y_i, \phi_i)). \tag{3}$$

The likelihood maximization involves a nonlinear equation for which the solution is not in closed form. Nelder and Wedderburn (1972) proposed the Fisher-scoring algorithm in order to find a numerical approximation of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$. Fisher-scoring is one of the best known quasi-Newton method to solve the likelihood maximization problem (see Smyth (2002)). For the implementation of this algorithm we have to choose an initial value $\boldsymbol{\beta}^{(0)}$ for the parameters of the model and then to apply iteratively the relation:

$$\forall\, k \in \mathbb{N}^*, \quad \boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left(\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{q}^{(k)} \tag{4}$$

where $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^r$ is an approximation of the solution at iteration k, $\mathbf{X}$ is the model matrix (with n rows and r columns), $\mathbf{W}^{(k)}$ and $\mathbf{q}^{(k)}$ depend on the vector $\boldsymbol{\beta}$ at iteration k as follows:

$$\mathbf{W}^{(k)} = diag(\omega_i, i = 1, \ldots, n) \text{ with } \omega_i = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial m_i}{\partial \eta_i}\right)^2,$$

$$\eta_i = g(m_i) \text{ and } \mathbf{q}^{(k)} \text{ as } \frac{\partial L(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \beta_j} \text{ for } j\text{-th element}(j = 1, \ldots, r).$$

Note that the matrix $\mathbf{W}^{(k)}$ has to be computed at every iteration because it depends on m_i and $m_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ depends on the value of the approximation of the solution at iteration k (vector $\boldsymbol{\beta}^{(k)}$).

Remark 1 It is also possible to find a vector $\mathbf{z}^{(k)}$ such that relation (4) becomes:

$$\boldsymbol{\beta}^{(k+1)} = \left(\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{z}^{(k)}.$$

In other words, the Fisher scoring algorithm is also an iteratively reweighted least squares method.

2.3 Factorial designs

We assume now that every variable is coded in such a way that its values always belong to the interval $[-1, 1]$ (this can be done in a very simple way by using a linear transformation, see chapter 2 of Khuri and Cornell (1996)). A complete factorial design, for m factors, is then constituted by all the vertices of the cube $[-1, 1]^m$. Nevertheless,

using such designs is not possible when the number of factors m becomes high (because of the 2^m experimental units). So we also consider in the following some regular fractions of these factorial designs (see Box and Hunter, 1961*a*, *b*). In other words, we are now working with configurations given by:

1) 2^{m-q} vertices of the cube $[-1,1]^m$,
2) n_0 central replications of the experimental domain.

Example 2 For $m = 3$ factors we can consider first the complete factorial design associated to the design matrix $\mathbf{D}_C$ (*i.e.* the $n \times m$ matrix with row i made up from the m coordinates of the i-th design point). Another choice is given by a regular fraction associated with the matrix $\mathbf{D}_F$ (with $n_0 = 1$ central point in our case):

$$\mathbf{D}_C = \begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{D}_F = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

This fraction is obtained by keeping the experimental units such that $x_1 x_2 x_3 = +1$ where x_i denotes the i-th coordinate of each design point of the factorial part. We say in a classic way that this regular fraction is generated by the relation $123 = \mathbb{I}$ where 1, 2 and 3 are the three different columns of the design matrix and 123 is $1 \odot 2 \odot 3$ with $\odot$ the Hadamard product operator (also called elementwise product).

In the framework of linear models these factorial designs can be used in order to fit linear (L) or interaction (I) models such that:

$$(L)\ ,\ \forall\, i = 1, \ldots, n\ ,\ m_i = E(Y_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}\ ,$$

$$(I)\ ,\ \forall\, i = 1, \ldots, n\ ,\ m_i = E(Y_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum\sum_{j<l} \beta_{jl} x_{ij} x_{il}.$$

We denote in the following by $\mathbf{D}$ the design matrix, by $\mathbf{D}_j$ $(1 \leq j \leq m)$ the j-th column of this matrix and we put $\mathbf{Q}_{jl} = \mathbf{D}_j \odot \mathbf{D}_l$ $(1 \leq j < l \leq m)$. The model matrix is then given by:

$$\mathbf{X} = [\mathbb{I}_n \mid \mathbf{D}_1 \ldots \mathbf{D}_m] \text{ for the model } (L),$$
$$\mathbf{X} = [\mathbb{I}_n \mid \mathbf{D}_1 \ldots \mathbf{D}_m \mid \mathbf{Q}_{12} \ldots \mathbf{Q}_{(m-1)m}] \text{ for the model } (I).$$

It is well known that the matrix model $\mathbf{X}$ is of full rank (*i.e.* $\mathbf{X}^T\mathbf{X}$ is regular) for the two models when the factorial design is complete. In the case of a regular fraction then it will be of resolution at least III for the model (L) and at least V for the model (I) in order the obtain a full rank matrix $\mathbf{X}$ (see Box and Hunter, 1961*a*, *b*). When a factorial regular design is used we have also an orthogonal configuration such that (for the two models):

$$\mathbf{X}^T\mathbf{X} = \text{diag}\left(2^{m-q} + n_0, 2^{m-q}, \cdots, 2^{m-q}\right).$$

Example 3 (continuation) The complete factorial design associated with matrix $\mathbf{D}_C$ can be used to fit model (L) or (I). For the regular fraction associated with the matrix $\mathbf{D}_F$ it is a fraction of resolution III because it has only one generator (123) and this generator is a word of length 3. So such a fraction can be used to fit model (L) but is not able to fit model (I). In other words the following model matrix $\mathbf{X}_F^1$ for model (L) is of full rank but $\mathbf{X}_F^2$ for model (I) is not (because, for example, columns 2 and 7 are the same):

$$\mathbf{X}_F^1 = \left[\begin{array}{c|ccc} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{array}\right], \mathbf{X}_F^2 = \left[\begin{array}{c|ccc|ccc} 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

3 The surrogate link function

3.1 Modified Fisher-scoring method

Our goal is now to simplify the algorithm of Fisher scoring by dropping out the diagonal weighting matrix $\mathbf{W}$. This can be done by a judicious choice of the link function. In fact our objective is:

$$\mathbf{W} = I_d \Leftrightarrow \forall\, i = 1, \ldots, n\,, \frac{1}{\text{Var}\,(Y_i)}\left(\frac{\partial m_i}{\partial \eta_i}\right)^2 = 1. \tag{5}$$

But we know, from relation (2), that $\text{Var}\,(Y_i) = V(m_i)$. Then $m_i = g^{-1}(\eta_i)$ implies that:

$$(5) \Leftrightarrow \frac{\partial m_i}{\partial \eta_i} = \sqrt{V(m_i)} \Longleftrightarrow \frac{1}{g'(m_i)} = \sqrt{V(m_i)}.$$

Our proposal relies on the following lemma:

Lemma 1 *The matrix* $\mathbf{W}$ *is the identity matrix if and only if the link function g satisfies:*

$$\forall\ i = 1, \ldots, n\ ,\ g'(m_i) = V^{-1/2}(m_i)\,.$$

Such a function is then called the surrogate link function.

Table 1 gives, for some exponential families of distributions, the surrogate link functions (depending on t) verifying the differential equations of lemma 1 (with the additive constant chosen to be zero). We also recall in this table the associated canonical link functions.

Table 1: *Surrogate link function for different distributions.*

Distribution of Y_i	*Function* V	*Surrogate link fn.*	*Canonical link fn.*
Bernoulli (p)	$t(1-t)$	$\arcsin(2t-1)$	$\ln\left(\frac{t}{1-t}\right)$
Binomial $\mathcal{B}(n,p)$	$t\left(1-\frac{t}{n}\right)$	$\sqrt{n}\arcsin\left(\frac{2t}{n}-1\right)$	$\ln\left(\frac{t}{n-t}\right)$
Neg. Bin. (n,p)	$t\left(\frac{t}{n}+1\right)$	$\sqrt{n}\,\mathrm{arccosh}\left(\frac{2t}{n}+1\right)$	$\ln\left(\frac{t}{n+t}\right)$
Poisson $\mathcal{P}(\lambda)$	t	$2\sqrt{t}$	$\ln t$
Gamma $\mathcal{G}(a,p)$	$\frac{t^2}{p}$	$\sqrt{p}\ln t$	$\frac{p}{t}$

Remark 2 We have seen in Section 2.2 that the Fisher-scoring algorithm is in fact an iteratively reweighted least squares method. Then, the use of the surrogate link function allows us to have an iteratively unweighted least squares method.

The algorithm of Fisher scoring needs also the use of a vector $\mathbf{q}$ with j-th element $\partial L(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi})/\partial\beta_j$ for $j = 1, \ldots, r$ (see section 2.2). We have the following relation, using the chain rule:

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial m_i}\frac{\partial m_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Then we obtain immediately for the likelihood of every sample of the exponential family (see formula (3)):

$$\forall\ j = 1, \ldots, r\ ,\ \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n}\frac{(y_i - m_i)}{\mathrm{Var}\,Y_i}\frac{\partial m_i}{\partial \eta_i}[\mathbf{X}]_{ij}$$

where $[\mathbf{X}]_{ij}$ is the element of row i and column j of the matrix model $\mathbf{X}$. This general relation can be simplified in our case because we have:

$$\eta_i = g(m_i) \text{ with } g'(m_i) = V^{-1/2}(m_i) \text{ so } \frac{\partial m_i}{\partial \eta_i} = \sqrt{V(m_i)}.$$

Thus, we can state the following lemma:

Lemma 2 *If the link function is the surrogate link function the vector* $\mathbf{q}$ *is then defined by:*

$$\forall\ j = 1, \ldots, r\ ,\ \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} [\mathbf{X}]_{ij}\, y_i^* \text{ with } y_i^* = \frac{y_i - m_i}{\sqrt{V(m_i)}}.$$

We see from lemma 2 that the vector $\mathbf{q}$ has a very simple expression when the surrogate link function is used. It needs only the observations y_i^* in their standardized and centred form.

Example 4 (continuation) Consider a random binary phenomenon such that every observed response y_i is a realization of a Bernoulli distribution with parameter p_i (unknown). Here we make the assumption that this phenomenon depends on three factors and the true response is given by:

$$\forall\ i = 1, \ldots, n\ ,\ p_i = 0.2x_{i1} - 0.1x_{i2} - 0.1x_{i3} + 0.6$$

where x_{i1}, x_{i2} and x_{i3} are the coded levels for the three factors. In other words, we assume that the probabilities associated with each Bernoulli distribution can be correctly fitted by a Taylor series of order one in the experimental domain. We also assume that the effects of factors 2 and 3 are opposite (and lower) to the effect of the factor 1 on the response. In order to make a modelization of this phenomenon using the surrogate link function (such that $g(t) = \arcsin(2t - 1)$ in our case) we can consider the following model (with $m_i = \mathrm{E}(Y_i) = p_i$):

$$\forall\ i = 1, \ldots, n\ ,\ \arcsin(2m_i - 1) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

3.2 Application to factorial designs

Consider a random phenomenon of m factors that may be checked by the experimenter. We have seen that the choice of the surrogate link function allows us to put the matrix $\mathbf{X}^T\mathbf{X}$ in place of the initial matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ in the algorithm of Fisher scoring. The optimal situation is then reached when a complete factorial design or a well chosen regular

fraction is used, because we have seen in Section 2.3 that $\mathbf{X}^T\mathbf{X}$ is then a diagonal matrix (*i.e.* the design is orthogonal). Now we consider in the following the two non-linear models given below:

$$(L^*)\ ,\ \forall\ i = 1, \ldots, n\ ,\ g(m_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}\ ,$$

$$(I^*)\ ,\ \forall\ i = 1, \ldots, n\ ,\ g(m_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \sum\sum_{j<l} \beta_{jl} x_{ij} x_{il},$$

Models (L^*) and (I^*) are then two generalized linear models with a polynomial linear part of degree one for (L^*) and of degree two with interactions for (I^*). Using relation (4) and lemmas 1 and 2 we can state the following simplified iterative treatment when factorial designs are used:

Proposition 3 *Consider the model* (L^*) *or* (I^*) *used with the surrogate link function. For a complete factorial design or a regular fraction of resolution at least III, the Fisher scoring algorithm is given for the model* (L^*) *by:*

1) $\beta_0^{(k+1)} = \beta_0^{(k)} + \dfrac{1}{2^{m-q} + n_0} \displaystyle\sum_{i=1}^{n} y_i^*,$

2) $\forall\ j = 1, \ldots, m\ ,\ \beta_j^{(k+1)} = \beta_j^{(k)} + \dfrac{1}{2^{m-q}} \displaystyle\sum_{i=1}^{n} x_{ij} y_i^*.$

For a complete factorial design or a regular fraction of resolution at least V, the algorithm of Fisher scoring for the model (I^*) *verifies, in addition to the two previous relations:*

3) $\forall\ j, l = 1, \ldots, m \text{ with } j < l\ ,\ \beta_{jl}^{(k+1)} = \beta_{jl}^{(k)} + \dfrac{1}{2^{m-q}} \displaystyle\sum_{i=1}^{n} x_{ij} x_{il} y_i^*.$

The implementation of the Fisher scoring algorithm is then very simple in our case because we only have to apply iteratively results from this proposition and no use of matrix calculus is needed (in particular we do not have to invert any matrix). Note also that factorial designs have only two levels, so the values for the coded variables x_{ij} are only -1, 0 (if at least one central point is used) or 1. This algorithm has to be initialized by judicious values for $\beta^{(0)}$. This can be done, for example, by a classic linear regression on the transformed response (*i.e.* on the $g(y_i)$ with g surrogate link function in place of the y_i). It can also be stopped using different criteria: when the likelihood seems to be constant (*i.e.* when $|L_{\max}^{(k+1)} - L_{\max}^{(k)}| < \varepsilon$ with ε small positive) or when the estimated parameters seem to be constant (*i.e.* when $\|\beta^{(k)} - \beta^{(k+1)}\| < \varepsilon$ where $\|.\|$ is a chosen norm) for example.

Example 5 (continuation) For the binary responses we assume that the experimenter has conducted the experiment according to a complete factorial design with two centre

points (the low number of factors allow us to consider the complete design in this case). We have then a total of 10 trials given in Table 2 with the probabilities p_i associated for each experimental unit (column p_i) and simulated results for the different responses (column y_i).

Table 2: *Results for the complete factorial design.*

Trial	Fac. 1	Fac.2	Fac. 3	p_i	$\mathbf{y}_i$	$\hat{p}_i$	$\hat{\mathbf{y}}_i$
1	1	1	1	0.60	**1**	0.54 *(0.75)*	**1** *(1)*
2	−1	1	1	0.20	**0**	0.27 *(0.00)*	**0** *(0)*
3	1	−1	1	0.80	**1**	1.00 *(1.00)*	**1** *(1)*
4	1	1	−1	0.80	**1**	1.00 *(1.00)*	**1** *(1)*
5	−1	−1	1	0.30	**0**	0.03 *(0.00)*	**0** *(0)*
6	−1	1	−1	0.30	**0**	0.03 *(0.00)*	**0** *(0)*
7	1	−1	−1	1.00	**1**	0.59 *(1.00)*	**1** *(1)*
8	−1	−1	−1	0.60	**1**	0.60 *(0.75)*	**1** *(1)*
9	0	0	0	0.60	**1**	0.57 *(0.75)*	**1** *(1)*
10	0	0	0	0.60	**0**	0.57 *(0.75)*	**1** *(1)*

If we have no information concerning the choice for the initial values of the algorithm, we can take, for example:

$$\beta_0^{(0)} = 1\,,\, \beta_1^{(0)} = \beta_2^{(0)} = \beta_3^{(0)} = 0.$$

Then the iterative treatment of Proposition 3 leads us very quickly (in two iterations) to the maximum likelihood solution:

$$\hat{\beta}_0 = 0.143\,,\, \hat{\beta}_1 = 1.376\,,\, \hat{\beta}_2 = -0.719 \text{ and } \hat{\beta}_3 = -0.719.$$

In other words, the best fitted model satisfies ($\forall\ x_1,\ x_2 \in [-1, 1]$):

$$\hat{p}\,(x_1, x_2) = \frac{\sin\,(0.143 + 1.376x_1 - 0.719x_2 - 0.719x_3) + 1}{2}.$$

Predicted values of the probabilities p_i are given in Table 2 (column $\hat{p}_i$) with the predicted responses (column $\hat{y}_i$), that is the values of $\hat{p}_i$ rounded to the nearest integer. We also present, in brackets, results obtained by the classical analysis with the canonical link function (these results come from the SAS software). We observe the good global quality of the results since observed responses y_i and predicted responses $\hat{y}_i$ are always the same (except, of course, for the two last trials where it is impossible to predict at once 0 and 1). If we consider probabilities p_i we note, on the one hand, that predictions are very good for half of the experiments (*i.e.* trials 1, 2, 8, 9 and 10). On the other hand, these results are not so good for trials 3 and 4 and they are bad for trials 5, 6 and 7. These problems of prediction are principally due to the small number of trials, and also to the nature of the responses which give poor information because we have only two levels.

We can finally note that the adjusted model allows us to find again the correct effect of each factor (*i.e.* factor 1 has a preponderant effect on the response and factors 2 and 3 have equal effects, opposite to factor 1).

3.3 Dispersion of the estimations

We know (see Green and Silverman (1994)) that asymptotically the maximum likelihood estimator of $\boldsymbol{\beta}$ has a Gaussian distribution and a dispersion given by:

$$\mathrm{Var}\hat{\boldsymbol{\beta}} = \phi\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}.$$

If ϕ is unknown then it can be estimated by means of Pearson statistics. This result is very interesting in our case because we know that $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is a diagonal matrix and the diagonal elements are given in the last subsection. So we have the following proposition:

Proposition 4 *Consider the model* (L^*) *or* (I^*) *used with the surrogate link function and a complete factorial design or an appropriate regular fraction (of resolution at least III for* (L^*) *and at least V for* (I^*)*). The maximum likelihood estimator* $\hat{\beta}$ *satisfies asymptotically the following properties:*

1) its components are non-correlated,

2) its dispersion is given by:

$$\mathrm{Var}\,\hat{\beta}_0 = \frac{\phi}{2^{m-q}+n_0} \text{ and } \forall\ j,l = 1,\ldots,m\ ,\ j<l\ ,\ \mathrm{Var}\,\hat{\beta}_j = \mathrm{Var}\,\hat{\beta}_{jl} = \frac{\phi}{2^{m-q}}.$$

Remark 3 The dispersion parameter ϕ is needed in order to obtain these different dispersions. This is not a serious problem in practice because ϕ has often a simple form (for example, $\phi = 1$ for a binomial distribution, a Poisson distribution and a negative-binomial distribution).

3.4 Considering submodels

For m quantitative factors, models (L^*) and (I^*) are not often the best choice because some linear effects or interactions may be sometimes removed. Then it is preferable to use a submodel. Propositions 3 and 4 follow from the orthogonal properties of the model matrix $\mathbf{X}$ when the model is complete. These results are then *a fortiori* true in the case of a submodel. The main problem for an experimenter is the validation of such a submodel. In other words, is the chosen submodel really better than the complete model? To answer to this question it is usually advised to use the notion of deviance (see Green

and Silverman [6]). For a given model and a submodel, the deviance is defined as:

$$D = -2\left[L_{\max}\left(\text{submodel}\right) - L_{\max}\left(\text{model}\right)\right] \tag{6}$$

where $L_{\max}(.)$ is the maximal value of the likelihood for the (sub)model (*i.e.* $L(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator). The choice of the submodel is then a good alternative when the deviance D is close to zero. In order to quantify this notion we usually use the following rule: when the model has r unknown parameters and the submodel has $r' < r$ unknown parameters, the submodel is then a better one if and only if:

$$D < \chi^2_{p-p',0.05}$$

with $\chi^2_{p-p',0.05}$ the upper 5% point of a χ^2 distribution with $(r - r')$ degrees of freedom.

4 Application to the geometric distribution

4.1 Utilization of a full model

We consider in this part responses with a binomial negative distribution and, more precisely, the particular case of the geometric distribution. It is, once again, a very classical situation and such a distribution is in the exponential family because its density satisfies relation (1) with:

$$\theta_i = \ln\left(1 - p_i\right)\ ,\ v\left(\theta_i\right) = -\ln\left(1 - e^{\theta_i}\right)\ ,\ \phi = 1 \text{ and } h\left(y_i, \phi\right) = \mathbb{I}_{\mathbb{N}}\left(y_i, \phi\right).$$

We can illustrate such model by considering experiments made in order to test the tensile strength of ropes. The experimenter makes identical tractions and the response is then the number of tractions endured by the rope before breaking. We assume in the following that the tensile strength of the rope depends mainly on five concentrations of chemicals (called now factors 1, 2, 3, 4 and 5).

From section 2.3 we consider the surrogate link function $g(t) = \operatorname{arccosh}(2t + 1)$ and we can use the interaction model (with $m_i = E(Y_i) = (1 - p_i)/p_i$):

$$\forall\ i = 1, \ldots, n\ ,\ \ \operatorname{arccosh}\left(2m_i + 1\right) = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \sum\sum_{j<k} \beta_{jk} x_{ij} x_{ik}.$$

We consider in the following the experimental design obtained by the regular fraction of the factorial design such that $\mathbb{I}_{16} = 12345$. With the addition of three central replications, we have then a total of 19 experiments given in Table 3 (with 1 denoted by + and −1 by

−). Responses given in this table are obtained by simulation of geometric distributions with p_i parameters such that :

$$\forall\, i = 1, \ldots, 19\,, \quad \operatorname{arccosh}(2m_i + 1) = \\ x_{i1} + 0.5x_{i2} + 0.5x_{i3} + 0.5x_{i4} - 0.5x_{i5} + x_{i1}x_{i2} + 0.5x_{i1}x_{i4} + 2. \tag{7}$$

In other words, we make two important assumptions in this part. First, we assume that there are only two interaction effects in this phenomenon and they are associated with the pair of factors $\{1, 2\}$ and $\{1, 4\}$. Secondly, we assume here that we are in "optimal" conditions because the true model uses the surrogate link function (simulations with another link function will be used later).
Now we can implement the Fisher-scoring algorithm with a set of simulated responses (given in column y_i of Table 3). Concerning the initial values, we take:

$$\beta_0^{(0)} = \operatorname{arccosh}(2\overline{y} - 1) \text{ and all the others components of } \boldsymbol{\beta}^{(0)} \text{ are zero.}$$

So, the algorithm is initiated with the best choice for a constant model.

Table 3: *Results for the fractional factorial design, the full model and the submodel (in brackets).*

Exp	f 1	f 2	f 3	f 4	f 5	p_i	y_i	$\hat{p}_i$	$\hat{y}_i$
1	+	+	+	+	+	0.02	**40**	0.03 *(0.02)*	**39** *(64)*
2	−	−	+	+	+	0.60	**1**	0.51 *(0.50)*	**1** *(1)*
3	−	+	−	+	+	0.94	**0**	0.99 *(0.94)*	**0** *(0)*
4	−	+	+	−	+	0.94	**0**	0.99 *(0.99)*	**0** *(0)*
5	−	+	+	+	−	0.60	**1**	0.51 *(0.63)*	**1** *(1)*
6	+	−	−	+	+	0.60	**2**	0.34 *(0.52)*	**2** *(1)*
7	+	−	+	−	+	0.94	**0**	0.99 *(0.97)*	**0** *(0)*
8	+	−	+	+	−	0.11	**7**	0.13 *(0.10)*	**7** *(9)*
9	+	+	−	−	+	0.28	**4**	0.21 *(0.29)*	**4** *(2)*
10	+	+	−	+	−	0.02	**72**	0.01 *(0.01)*	**70** *(76)*
11	+	+	+	−	−	0.04	**20**	0.05 *(0.05)*	**19** *(19)*
12	−	−	−	−	+	0.94	**0**	0.99 *(0.89)*	**0** *(0)*
13	−	−	−	+	−	0.60	**1**	0.51 *(0.44)*	**1** *(1)*
14	−	−	+	−	−	0.28	**5**	0.17 *(0.26)*	**5** *(3)*
15	−	+	−	−	−	0.94	**0**	0.99 *(0.97)*	**0** *(0)*
16	+	−	−	−	−	0.94	**0**	0.99 *(0.64)*	**0** *(0)*
17	0	0	0	0	0	0.42	**0**	0.44 *(0.41)*	**1** *(1)*
18	0	0	0	0	0	0.42	**2**	0.44 *(0.41)*	**1** *(1)*
19	0	0	0	0	0	0.42	**1**	0.44 *(0.41)*	**1** *(1)*

The iterations continue until the likelihood increases by only a small amount (*i.e.* until $L_{\max}^{(k+1)} - L_{\max}^{(k)} < \varepsilon$ with $\varepsilon = 0.001$). Then, we obtain the following estimates after 10 iterations:

$$\begin{array}{lllll} \hat{\beta}_0 = 1.946 & \hat{\beta}_1 = 0.971 & \hat{\beta}_{12} = 1.097 & \hat{\beta}_{23} = -0.094 & \hat{\beta}_{34} = -0.145 \\ & \hat{\beta}_2 = 0.469 & \hat{\beta}_{13} = -0.186 & \hat{\beta}_{24} = -0.067 & \hat{\beta}_{35} = -0.233 \\ & \hat{\beta}_3 = 0.441 & \hat{\beta}_{14} = 0.435 & \hat{\beta}_{25} = 0.021 & \hat{\beta}_{45} = 0.072 \\ & \hat{\beta}_4 = 0.731 & \hat{\beta}_{15} = 0.113 & & \\ & \hat{\beta}_5 = -0.514 & & & \end{array}$$

The predicted probabilities $\hat{p}_i$ and the predicted mean responses (i.e. the rounded values to the nearest integer of $(1 - \hat{p}_i)/\hat{p}_i$) are then reported in Table 3 (results in brackets will be discussed in the next subsection). We note the global good fit of the model: observed responses and predicted responses are always very close. The maximum likelihood associated with this model is equal to −32.523.

4.2 Utilization of a submodel

Our goal in this part is to find a good submodel of the previous full polynomial model. Again we find that the submodel containing only the interactions x_1x_2 and x_1x_4 is interesting because it is associated with a maximum likelihood equal to −33.531. In other words, the deviance between the full model (with $r = 16$ parameters) and this submodel (with $r' = 8$ parameters) is:

$$D = -2\,(-33.531 + 32.523) = 2.016.$$

But we have $\chi^2_{8,0.05} = 15.51$ so results from Section 3.4 show us that this submodel is a good alternative to the full model. The Fisher-scoring algorithm leads us (after 8 iterations and until $L_{\max}^{(k+1)} - L_{\max}^{(k)} < \varepsilon$ with $\varepsilon = 0.001$) to the following estimates:

$$\begin{array}{llll} \hat{\beta}_0 = 2.039 & \hat{\beta}_1 = 0.987 & \hat{\beta}_4 = 0.612 & \hat{\beta}_{12} = 1.095 \\ & \hat{\beta}_2 = 0.396 & \hat{\beta}_5 = -0.518 & \hat{\beta}_{14} = 0.507 \\ & \hat{\beta}_3 = 0.429 & & \end{array}$$

In conclusion, the best fitted model is then given by the following formula (for every $x = (x_1, x_2, x_3, x_4, x_5) \in [-1, 1]^5$) :

$$\hat{p}\,(x) = \frac{2}{\cosh\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{14} x_1 x_4\right) + 1}$$

Results concerning this submodel are given in brackets in Table 3. Once again, we note the good quality of predicted probabilities and estimated responses.

4.3 Quality of the parameter estimations

Results from Sections 5.1 and 5.2 are obtained with only one simulation of the responses y_i ($i = 1, \ldots, 19$). So it is natural to perform now a large number of simulations in order to evaluate the global quality of this method. Table 4 presents, for each parameter of the model, the basic statistical results (mean and dispersion) for 1000 simulations of the geometric distribution.

Table 4: *Simulation results.*

Param.	*Mean*	*Variance*	*Param.*	*Mean*	*Variance*
$\hat{\beta}_0$	1.712	0.074	$\hat{\beta}_4$	0.532	0.100
$\hat{\beta}_1$	0.967	0.095	$\hat{\beta}_5$	−0.457	0.099
$\hat{\beta}_2$	0.556	0.085	$\hat{\beta}_{12}$	0.981	0.089
$\hat{\beta}_3$	0.470	0.095	$\hat{\beta}_{14}$	0.485	0.096

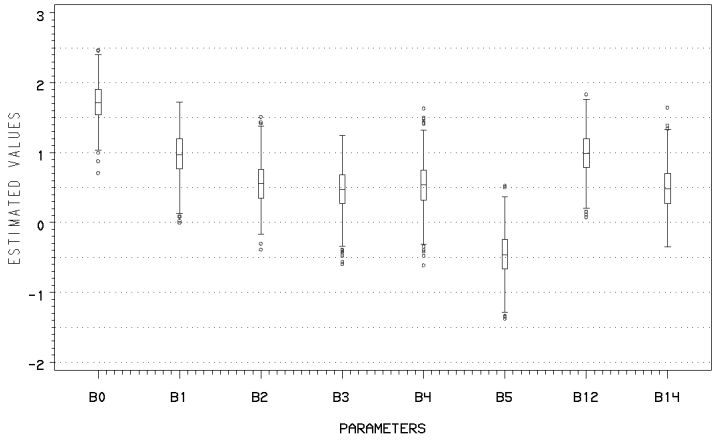

Figure 1: *Boxplots of values of the estimated parameters; the length of the whiskers is 1.5 times the interquartile range.*

A graphical representation of these results, using boxplots, is also given (Figure 1). We deduce from Table 4 and Figure 1 that this method of estimation is adequate concerning the stability of the estimated parameters (i.e. only a very few of these parameters are outside the whiskers). We have also a good convergence speed of this iterative method because, for the 1000 simulations, the algorithm needs an average of 8.7 iterations in order to converge (less than 10 iterations are needed in 94 % of the cases and the maximum does not exceed 20). Concerning the linear and interaction effects we note the good quality of the estimated values, with mean and median very close from the theoretical values of the model of Section 5.1. The only imprecision concerns the general mean effect β_0 which is underestimated.

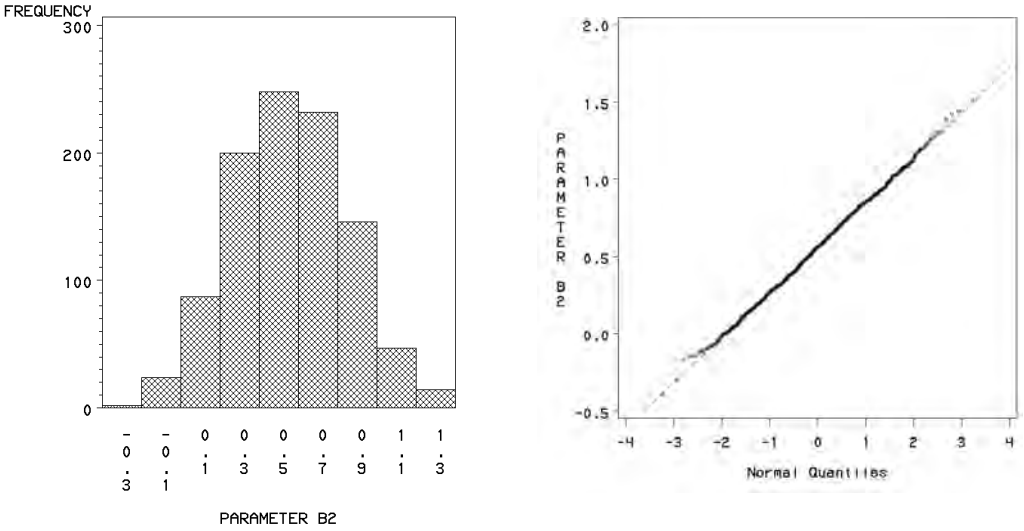

Figure 2: *Estimation of β_2 (histogram and QQ-plot).*

Another important problem concerns the validity of Proposition 4. The goal in statistical planning is to reduce the number of experiments so we must be very careful with asymptotic results. Nevertheless, some properties of this proposition are true in our case. First, we find again that dispersions of linear and interaction effects seem to be very close whereas the dispersion of the general mean effect is smaller (because of the three central replications). Secondly, Figure 2 (the histogram and QQ-plot for the estimated values of the parameter β_2) shows us that we can assume that this parameter follows a normal distribution, and we obtain similar results in the case of the other parameters).

We have the same problem for the choice of a submodel with the deviance criterion. We know that D follows asymptotically a $\chi^2_{r-r'}$ distribution but is this result true for our 19 experiments ? We have computed, for each simulation, the deviance between the full model and the chosen submodel with only interactions x_1x_2 and x_1x_4. Figure 3 gives a graphical representation (the histogram and QQ-plot) for these deviances. The line of the QQ-plot represents the best fitted χ^2 distribution and we find that it has 6 degrees of freedom (i.e. it is a gamma distribution with parameters $1/2$ and 3). In conclusion, we note that the deviance is close to a χ^2 distribution but we have to be careful because the observed degrees of freedom (6) are smaller than the theoretical ones ($r - r' = 8$). This fact implies that theoretical results lead us to reject the submodel when $D > \chi^2_{8,0.05} = 15.51$ but it seems more adequate to reject it as soon as $D > \chi^2_{6,0.05} = 12.59$. Note that it has a weak influence for the validation of the submodel because, for our 1000 simulations, we have only 16 values of D in the interval $[12.59, 15.51]$.

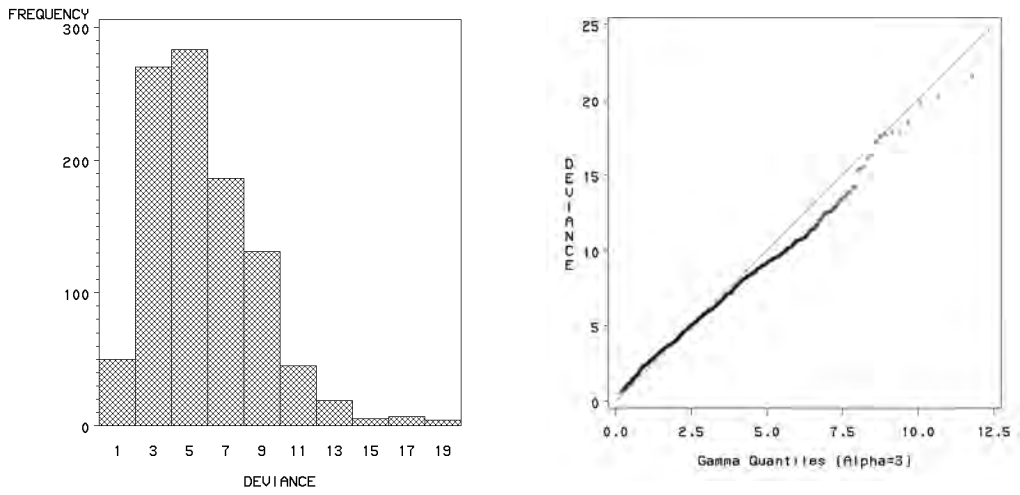

Figure 3: *Deviance (histogram and QQ-plot).*

4.4 Comparison with the classical method

The previous Sections 5.1, 5.2 and 5.3 use simulated responses given by the model (7) (i.e. responses obtained with the surrogate link function). Now we are going to use other simulations in order to compare our method to a classical one. For the classical method we use the GENMOD procedure of the SAS program. In the case of a geometric distribution this procedure uses by default the logarithm function for the link. So, responses are now obtained by simulation of a geometric distribution with probabilites given by (with $m_i = E(Y_i) = (1 - p_i)/p_i$):

$$\begin{gathered} \forall\, i = 1, \ldots, 19\,, \quad \ln(m_i) = \\ x_{i1} + 0.5x_{i2} + 0.5x_{i3} + 0.5x_{i4} - 0.5x_{i5} + x_{i1}x_{i2} + 0.5x_{i1}x_{i4} + 2. \end{gathered} \tag{8}$$

This model is then the optimal choice concerning the classical method because it uses the same link function. The values of the p_i are given for each point of the design in Table 5. For each of these parameters 1000 simulations have been made and the means of the estimated values for the p_i parameters are given in Table 5. We can note that these two methods gives very close estimations. Note also that the convergence is obtained for the surrogate link function with a mean of 8.7 iterations (and the number of iterations is always between 2 and 34) but for the classical method the SAS software stops the algorithm, in most of the cases, after 50 iterations because the convergence is not reached.

Figure 4 and 5 allow us to compare these two methods with a graphical representation using boxplots of the 17 estimated probabilities associated with every experimental unit. Once again the two results seem to be very close. Figure 6 is a graphical representation for the estimated values of the model parameters. We note that the stability of the estimated parameters is again satisfactory (i.e. all the observed distributions are very close to a normal distribution).

Table 5: *Simulation results (mean values of $\hat{p}_i$) SLF: surrogate link function. CLF: classical link function.*

Exp	p_i	**SLF** $\hat{p}_i$	CLF $\hat{p}_i$	Exp	p_i	**SLF** $\hat{p}_i$	CLF $\hat{p}_i$
1	0.06	**0.10**	0.11	10	0.06	**0.10**	0.11
2	0.32	**0.42**	0.41	11	0.10	**0.16**	0.16
3	0.56	**0.64**	0.61	12	0.44	**0.51**	0.50
4	0.44	**0.54**	0.52	13	0.32	**0.41**	0.40
5	0.32	**0.43**	0.42	14	0.22	**0.31**	0.31
6	0.32	**0.39**	0.40	15	0.44	**0.52**	0.51
7	0.44	**0.53**	0.50	16	0.44	**0.52**	0.50
8	0.15	**0.22**	0.23	17	0.27	**0.32**	0.35
9	0.22	**0.29**	0.30				

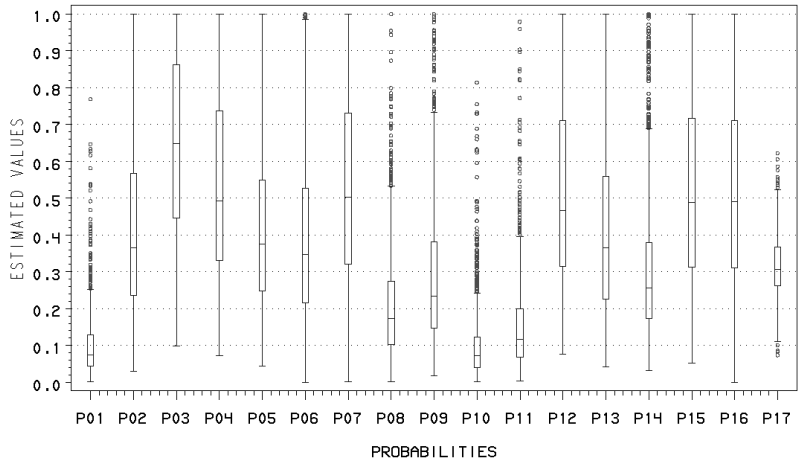

Figure 4: *Estimated probabilities for the surrogate link function.*

4.5 Conclusion

In this paper we have presented a new method in order to extend the classical one associated with the analysis of a linear model. The constraint of this new method

concerns the utilization of the surrogate link function. Neverthless, the two examples of this paper have suggested that this choice for the link is a good choice. This method has two principal advantages:

1) the Fisher-scoring algorithm is now very easy to improve (and then computations can be done faster),
2) classical designs like factorial designs, well known in the linear case, can be used.

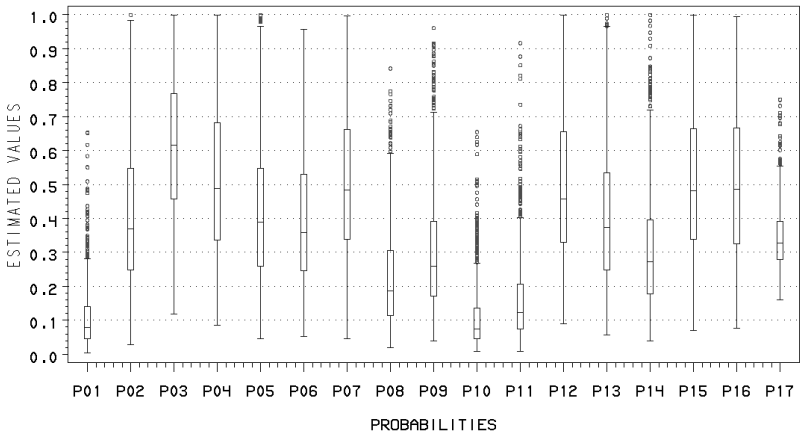

Figure 5: *Estimated probabilities for the canonical link function.*

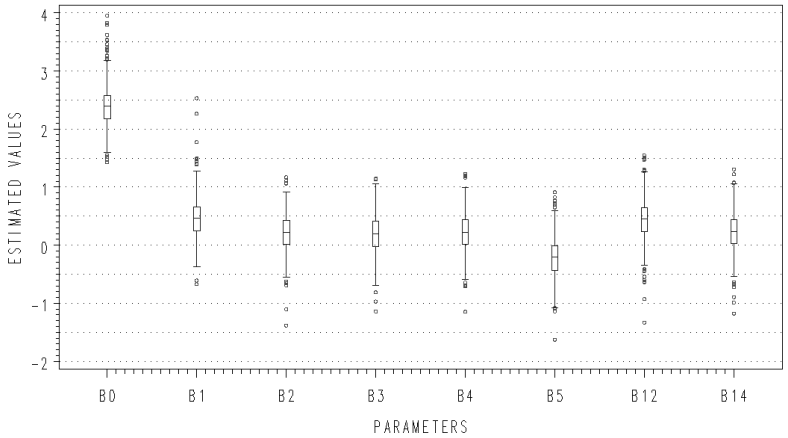

Figure 6: *Values of the estimated parameters (natural link function).*

References

Box G. and Draper N. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley, New-York.

Box G. E. P. and Hunter J. S. (1961*a*). The 2^{k-p} fractionnal factorial designs, Part *I*. *Technometrics*, 3, 311-351.

Box G. E. P. and Hunter J. S. (1961*b*). The 2^{k-p} fractionnal factorial designs, Part *II*. *Technometrics*, 3, 449-458.

Chipman H. A. and Welch W. J. (1996). D-Optimal Design for Generalized Linear Models, Unpublised. http://www.stats.uwaterloo.ca/~hachipma/publications.html.

Chaloner K. and Larntz K. (1989). Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21, 191-208.

Green P.J. and Silverman B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Monographs on Statistics and Applied Probability, 58. London: Chapman & Hall.

Khuri A. and Cornell J. (1996). *Response Surfaces: Designs and Analyses*. Dekker, Statistics: textbooks and monographs, 152, New-York.

McCullagh P. and Nelder J. A. (1989). *Generalized Linear Models (second edition)*. Monographs on Statistics and Applied Probability, 37. London: Chapman & Hall.

Nelder J. A. and Wedderburn R. W. M. (1972). Generalized linear models. *J. Roy. Stat. Soc.* A, 135, 370-384.

Pukelsheim F. (1993). *Optimal Design of Experiments*. New York: John Wiley.

Sebastiani P. and Settimi R. (1997), A note on D-optimal designs for a logistic regression model. *Journal of statistical planning and inference*, 59, 359-368.

Smyth G. K. (2002). Optimization. *Encyclopedia of Environmetrics*, 3, 1481-1487.

SORT 29 (2) July-December 2005, 269-288

A comparison of parametric models for mortality graduation. Application to mortality data for the Valencia Region (Spain)

A. Debón[1], F. Montes[2] and R. Sala[2]

[1]*Universidad Politécnica de Valencia,* [2]*Universitat de València*

Abstract

The parametric graduation of mortality data has as its objective the satisfactory estimation of the death rates based on mortality data but using an age-dependent function whose parameters are adjusted from the crude rates obtainable directly from the data. This paper proposes a revision of the most commonly used parametric methods and compares the results obtained with each of them when they are applied to the mortality data for the Valencia Region. As a result of the comparison, we conclude that the Gompertz-Makeham functions estimated by means of generalized linear models lead to the best results. Our working method is of additional interest for being applicable to mortality data for a wide range of ages from any geographical conditions, allowing us to select the most appropriate life table for the case in hand.

MSC: 62P05

Keywords: Gompertz-Makeham functions, Heligman and Pollard's laws, parametric graduation.

1 Introduction

Historically, Actuarial Science has worked with the mortality data of a population. The first step, and perhaps one of the fundamental ways in which statistics plays a part, is the graduation of mortality data. We define graduation (Haberman and Renshaw, 1996) as

*Address for correspondence:*A. Debón. Dpt. Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia. E-46022. Valencia. Spain. Tlf: +34 963877007 (Ext. 74961). Fax: +34 963877499. E-mail: andeau@eio.upv.es.

Received: February 2005

Accepted: October 2005

the set of principles and methods by which the observed (or crude) probabilities are fitted to provide a smooth basis for making practical inferences and calculations of premiums and reserves. Graduation is necessary (London, 1985) because the sequence of crude death probabilities generally presents brusque changes, which do not correspond the plausible hypothesis that the probabilities of death for two consecutive ages should be very close.

The graduation methods suggested in the literature, and used in practice, can be classified into two fundamental types: parametric and non-parametric, depending on whether they adjust the data to a function or simply achieve smoothness. Within the first type are the now classic Gompertz (1825) and Makeham (1860) models, used especially for advanced age groups: the former postulate that the force of mortality would grow exponentially with age, and the second adds a constant, an age independent component, to the exponential growth. These authors' proposals gave good results for data from the late 19th and early 20th centuries. Over time a mortality pattern evolved with an increase in mortality among the young and a relative hump among the middle-aged, such that it was difficult to obtain a good graduation with the Makeham formula, which in turn led to the introduction of new models known as the Heligman and Pollard laws (Heligman and Pollard, 1980). The Gompertz-Makeham function described in Forfar *et al.* (1988) generalizes the original models proposed by Gompertz and Makeham. In Renshaw (1991) and Renshaw *et al.* (1997), generalized linear and non-linear models are used for adjusting these functions. An example of non-parametric graduation by means of kernel smoothing can be found in Gavin *et al.* (1993, 1995).

The objective of this paper is to revise and compare different parametric graduation models by applying them to real mortality data for the Valencia Region, on the Spanish Mediterranean coast. The paper is organized as follows. In Section 2 we present the parametric graduation models: the methodology developed by the Continuous Mortality Investigation (CMI) Bureau and its extension to generalized linear models, and the so-called Heligman and Pollard model. Section 3 is devoted to obtaining crude estimations of the probability of death in the Valencia Region for the period 1999-2001. We apply the different graduation methods to these estimations, commenting on their advantages and disadvantages, as well as on their suitability for the mortality analysis in question. In Section 4 the different fittings are compared by means of the usual non-parametric tests, and in Section 5 the most relevant conclusions are presented.

2 A review of parametric models for mortality graduation

The representation of mortality data by means of parametric models attracted the attention of actuaries, demographers and statisticians throughout the past century. These methods are based on the hypothesis that the chosen measurement of mortality is a function of age, x, $f_\alpha(x)$ with $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ being parameters to be determined. In

short, obtaining the graduation consists of applying the regression techniques which are widely described and used in statistics literature to the particular case of mortality data.

The objective of applying these procedures is to obtain the best possible fitting with the minimum number of parameters. It is therefore necessary to obtain a balance between the number of parameters and the goodness of fit. Congdon (1993) warns how many demographic graduation studies have emphasized the goodness of fit without considering the statistical stability of the parameters involved in the regression, usually leading to an overparameterization of the model, which shows up when the following are observed: standard errors for the parameter estimates that are too big, high correlations between the parameter estimates and failures of convergence in the iterative routines of non-linear fitting. An overparameterization also has practical implications on the use of graduation. For example, in the comparison of the time series of the parameters obtained when fitting mortality data corresponding to different years, the prediction of values for future years can show irregular erratic fluctuations that can make prediction difficult. There is also a strong relation between overparameterization and the instability of the parameters over time. There are therefore reasons to prefer parsimonious functions, with few parameters, despite producing slight losses in the goodness of fit.

The form of the functions that fit the data are diverse and fundamentally based on the profile presented by the crude estimations of the mortality measure used. The different models proposed by various authors are collected together in Gerber (1997) and Benjamin and Pollard (1992).

2.1 CMI Bureau Methodology

The Continuous Mortality Investigation (CMI) Bureau of the Institute and Faculty of Actuaries of London was created in 1924, when the continuous collection of mortality data began. It is responsible for constructing standard life tables for use in Great Britain's insurance industry. Forfar *et al.* (1988) have given an easily-understood description of the methodology that is normally used by the CMI to produce such tables. This methodology is a generalization of the Gompertz (1825) and Makeham (1860) models. It was applied to Spanish data by Navarro (1991) and to data of the Valencia Region by Navarro *et al.* (1995).

In order to get the graduation, the CMI Bureau uses the *Gompertz-Makeham functions of the type (r,s)*. They are functions with $r + s$ parameters of the form

$$GM_{\alpha}^{r,s}(x) = \sum_{i=1}^{r} \alpha_i x^{i-1} + \exp\left(\sum_{j=r+1}^{r+s} \alpha_i x^{j-r-1} \right),$$

with the convention that if $r = 0$, they only present an exponential part, and if $s = 0$, they only possess a polynomial term. The *Logit Gompertz-Makeham of the type (r,s)* are

alternative models that can be derived from the GM functions, the general expression of which is

$$LGM_{\alpha}^{r,s}(x) = \frac{GM_{\alpha}^{r,s}(x)}{1 + GM_{\alpha}^{r,s}(x)}.$$

In order to estimate the value of the parameters included in these functions, two optimization criteria are considered, that of maximum likelihood or that of minimum χ^2, which in practice produce very similar graduations (a detailed discussion is presented by Forfar *et al.*, (1988)). The minimum χ^2 criterium is the usual χ^2 statistic, that is the sum of squared standardized residuals.

This methodology can be reformulated and extended by using the schemes of generalized linear and non-linear models. The experience in graduation using generalized linear models has been compiled in actuarial literature by Renshaw (1991), Renshaw and Hatzopoulos (1996), Renshaw *et al.* (1997) and Verrall (1996). The use of generalized linear models (GLM) for the graduation of both the probability of death at age x, q_x, and the force of mortality at age x, μ_x, is justified because both response variables are not normal. Details about modelling and probability distribution assumptions for both mortality measures follow.

2.1.1 GLM for μ_x

Let us suppose that E_x^c persons enter observation under hypothesis that the force of mortality (instantaneous mortality rate) is a constant, $\mu_{x+\frac{1}{2}}$, during the period of observation and that the death or survival of each one is independent. In this case E_x^c represents those central exposed to risk, which can get modified throughout the duration of the study, meaning that the number of individuals in the study is not determined. The number of deaths which occur in the period of observation, D_x, will have a Poisson distribution with average and variance equal to $E_x^c\mu_{x+\frac{1}{2}}$. We consider the graduation of μ_x, with $D_x \sim Po(E_x^c\mu_{x+\frac{1}{2}})$ independent, the link utilized being $log\left(\mu_{x+\frac{1}{2}}\right)$, which is the canonical link of the Poisson family, and the model which is used is $\mu_{x+\frac{1}{2}} = GM(r, s)$, which gives rise to a linear predictor when $r = 0$.

When the predictor is not linear, Renshaw (1991) suggests an iterative method which enables the application of a similar methodology that is based on Makeham's historical formula $\eta_x = A + Bc^x$. Given that is not possible to transform this non-linear form into a linear one unless $A = 0$, it is possible to introduce a trivial reparametrization in exponential form and write

$$\eta_x = \alpha + \beta exp(\phi x). \tag{1}$$

The non-linear term

$$g(x;\phi) = exp(\phi x), \tag{2}$$

can be approximated

$$g(x;\phi) \simeq g(x;\phi_0) + (\phi - \phi_0)\left(\frac{\partial g}{\partial \phi}\right)_{\phi=\phi_0},$$

so that $\beta exp(\phi x)$ can be replaced in (1) by $\beta u + \gamma v$ with

$$u = g(x;\phi_0), \qquad v = \left(\frac{\partial g}{\partial \phi}\right)_{\phi=\phi_0} \qquad \text{and} \qquad \gamma = \beta(\phi - \phi_0)$$

In this way, the non-linear term (2) has been converted into a linear expression which can be inserted in the predictor of a generalized linear model.

So, starting from an initial value ϕ_0, we calculate the covariables

$$u = \exp(\phi_0 x) \quad \text{and} \quad v = x\exp(\phi_0 x),$$

and the parameters β and γ estimated following the adjustment of the model as in any linear estimator.

We then update

$$\phi_1 = \phi_0 + \frac{\hat{\gamma}}{\hat{\beta}}$$

and this process is repeated until convergence, which is not guaranteed for very distant initial values. We found that an initial value of $\phi_0 = 0.0005$ produced convergence in many of the sets of typical data which we graduated in this way. This method enables the graduation of μ_x, with a Poisson distribution and identity link, through models $GM_x(r,2)$ with $r \neq 0$.

Another alternative consists of considering D_x as fixed and equal to the number of observed deaths, d_x, and assuming therefore that E_x^c follows a Gamma distribution with parameters $\alpha = d_x$ y $\beta = \mu_{x+\frac{1}{2}}$. Gerber (1997) considered this distribution and it was used by Renshaw *et al.* (1997) to graduate $1/\mu_x$, force of vitality according to Lambert (1772) terminology, through a generalized linear model. We can therefore use response E_x^c variables with averages $\lambda_x = d_x\dfrac{1}{\mu_{x+\frac{1}{2}}}$, variances $\sigma_x^2 = d_x\dfrac{1}{\mu^2_{x+\frac{1}{2}}}$ and weights $\omega_x = d_x$. Taking the log link, we get

$$\log\lambda_x = \log d_x - \log\mu_{x+\frac{1}{2}} = \log d_x + \eta_x,$$

where η_x is the linear predictor. Renshaw (1991) obtained results for his data which were very similar to both of the μ_x graduation proposals.

2.1.2 GLM for q_x

Let us suppose that E_x^i persons come under observation at age x and continue under observation until they survive to $x + 1$ or die before. In this case we denote initial exposed to risk as E_x^i, which determines the number of individuals in the study. Also, let us suppose that the probability of death during the year for each one of them is q_x, and that the death or survival of one is independent of the death or survival of the others. If we call D_x the random variable which represents the number of deaths that occur in the year, we will get $D_x \sim B(E_x^i, q_x)$.

We perform the graduation of q_x using the function

$$q_x = LGM(r, s) = \frac{GM(r, s)}{1 + GM(r, s)}, \tag{3}$$

using the transformation $\text{logit}(q_x)$ as the link, which is the canonical link of the binomial family. From (3) we easily obtain

$$\frac{q_x}{1 - q_x} = GM(r, s),$$

so that if $r = 0$, $\text{logit}(q_x)$ corresponds to a linear predictor.

Heligman and Pollard's laws. An alternative to the previous functions are the Heligman and Pollard laws (Heligman and Pollard, 1980). These laws have been used by various countries (England, Sweden, Germany, Spain, United States of America and Australia) since the UN promoted the fitting of mortality through Heligman and Pollard's first law. Heligman and Pollard, inspired by Thiele (1972), adjusted a new mortality law to post-war Australian data with the general expression

$$\frac{q_x}{1 - q_x} = \sum_{i=1}^{n} A_i \exp\left(-B_i \left(f_i(x) - C_i\right)^{D_i}\right),$$

where A_i, B_i, C_i, D_i, $i = 1, 2, ..., n$, are the parameters to be estimated, and where $f_i(x)$ is usually the identity function, $f_i(x) = x$, or $f_i(x) = \ln(x)$

The three expressions that really fitted Australian mortality were as follows:

Heligman and Pollard's first law

$$\frac{q_x}{1-q_x} = A^{(x+B)^C} + D\exp\left(-E(\ln x - \ln F)^2\right) + GH^{(x-x_0)}$$

an expression that they consider cannot be distinguished from

$$q_x = A^{(x+B)^C} + D\exp\left(-E(\ln x - \ln F)^2\right) + \frac{GH^x}{1+GH^x}.$$

Heligman and Pollard's second law

$$q_x = A^{(x+B)^C} + D\exp\left(-E(\ln x - \ln F)^2\right) + \frac{GH^x}{1+KGH^x} \tag{4}$$

Heligman and Pollard's third law

$$q_x = A^{(x+B)^C} + D\exp\left(-E(\ln x - \ln F)^2\right) + \frac{GH^{x^k}}{1+GH^{x^k}}$$

The first term models childhood mortality, the second one the accident hump and the third term natural mortality caused by senescence (Heligman and Pollard, 1980). The graph in Figure 1 shows this decomposition. The interpretation of the parameters is as follows: A represents the infant mortality rate; B represents death probability for children who are 1 year old; C is closely related with the rate at which an individual adapts to his environment, three parameters taking values in the interval (0,1). D, E and F are referred as the accident hump, D indicates the severity of the accident hump with values in (0,1), E with large values, in (0,∞), indicate a concentrated accident hump and F from 15 to advanced age indicates the location of the hump maximum. Finally, G indicates the base level of later adult mortality, and H is the rate of increase in mortality at the later adult ages and its domains are (0,1) and (0,∞) respectively.

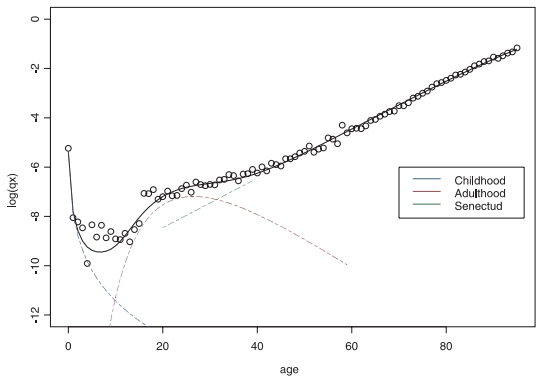

Figure 1: *Decomposition of Heligman and Pollard's Law.*

In order to estimate the coefficients, given the heterocedasticity of the data, the error structure should be accommodated by differential weighting of the rates for different ages (Congdon, 1993). Using criteria of weighted least squares $WLS(\alpha) = \sum_x \omega_x(\dot{q}_x - f_\alpha(x))^2$, with $f_\alpha(x)$ as in Heligman and Pollard's laws, and weights inversely proportional to binomial sampling variances

$$var(\dot{q}_x) = \frac{\dot{q}_x(1-\dot{q}_x)}{e_x},$$

and taking into consideration that $(1 - q_x) \approx 1$, we obtain the following expressions:

$$\begin{aligned} &a) \quad \sum_x \frac{e_x}{\dot{q}_x}(\dot{q}_x - f_\alpha(x))^2 \\ &b) \quad \sum_x (\dot{q}_x - f_\alpha(x))^2 \\ &c) \quad \sum_x \frac{e_x}{\dot{q}_x^2}(\dot{q}_x - f_\alpha(x))^2 \\ &d) \quad \sum_x \frac{1}{\dot{q}_x^2}(\dot{q}_x - f_\alpha(x))^2 \end{aligned} \tag{5}$$

including unweighted least squares in item *b)*. In these expressions, $\dot{q}_x$ is the crude estimate of q_x and e_x is the estimate of initial exposure to risk E_x^i.

An example of the application of these laws to mortality data of our geographic and social surroundings can be found in Felipe and Guillén (1999). They apply the second law to Spanish data for the period 1979-82. A Bayesian approach for Heligman and Pollard's laws has been proposed by Dellaportas *et al.* (2001), using Markov chain Monte Carlo simulation for avoiding the numerical problems that arise in classical methods.

3 Application to mortality data of the Valencia Region

The comparative study of the different parametric models of graduation is done by applying them to the mortality data of the Valencia Region, using aggregate population and death figures corresponding to the three-year period 1999-2001. These two data sets were published by the Spanish National Institute of Statistics (INE) and are classified by age (ranging from 0 to 100 or older) and sex. They both refer to the Valencia Region as the place of residence, which means the two sets of figures correspond to each other coherently.

As the population census takes place every 10 years and during the first year of the ten year period, only the data for 2001 are real counts, the data for 1999 and 2000 being

Table 1: *Age, x, initial number exposed to risk, e_x, and number of deaths, d_x, observed in the period 1999-2001.*

	MEN		WOMEN			MEN		WOMEN	
x	e_x	d_x	e_x	d_x	Age	e_x	d_x	e_x	d_x
0	39199.70	180.00	36955.70	182.00					
1	38315.50	21.00	36211.50	17.00	49	48797.50	212.00	50280.50	137.00
2	38139.50	5.00	35822.50	11.00	50	48223.00	234.00	49786.00	132.00
3	38096.00	7.00	35797.00	10.00	51	47697.50	254.00	49363.50	151.00
4	38345.00	6.00	36114.50	6.00	52	47301.50	298.00	49073.50	160.00
5	39066.00	8.00	36755.50	7.00	53	46565.50	265.00	48400.50	180.00
6	39971.50	9.00	37689.00	1.00	54	45335.00	318.00	47270.00	195.00
7	40772.50	10.00	38576.00	11.00	55	43837.00	345.00	45854.00	189.00
8	41449.00	8.00	39300.50	7.00	56	42854.00	349.00	44903.00	205.00
9	42046.50	2.00	39956.00	8.00	57	42080.50	376.00	44140.00	199.00
10	42696.50	8.00	40522.50	9.00	58	40376.50	359.00	42484.00	227.00
11	43574.50	5.00	41207.00	3.00	59	38943.50	457.00	41196.00	268.00
12	44781.00	5.00	42271.00	4.00	60	38785.00	440.00	41292.50	248.00
13	46223.00	10.00	43662.50	9.00	61	38629.50	476.00	41424.50	303.00
14	47927.00	11.00	45348.50	8.00	62	38341.50	487.00	41368.00	347.00
15	50014.50	20.00	47414.00	19.00	63	39109.00	591.00	42455.00	398.00
16	52505.50	34.00	49799.50	19.00	64	39885.50	656.00	43643.50	416.00
17	55265.50	50.00	52413.50	26.00	65	39758.50	706.00	43911.00	488.00
18	58202.00	66.00	55256.00	24.00	66	39325.00	744.00	43958.00	507.00
19	61265.00	49.00	58248.50	29.00	67	38834.50	868.00	43998.50	614.00
20	64133.00	66.00	61040.00	30.00	68	37958.00	939.00	43610.00	665.00
21	66470.00	60.00	63360.50	22.00	69	36676.50	922.00	42844.00	748.00
22	68160.50	58.00	65063.50	36.00	70	35290.50	994.00	42032.50	821.00
23	68999.50	59.00	66045.50	31.00	71	33869.00	1043.00	41144.50	881.00
24	69101.50	70.00	66303.00	31.00	72	32269.50	1146.00	40023.00	977.00
25	68737.50	56.00	66035.00	33.00	73	30646.50	1245.00	38763.00	1136.00
26	68178.00	71.00	65598.00	29.00	74	29109.00	1276.00	37516.50	1281.00
27	67544.00	73.00	65109.00	33.00	75	27468.00	1273.00	36204.50	1383.00
28	66988.50	69.00	64620.50	40.00	76	25565.00	1327.00	34564.00	1542.00
29	66568.50	81.00	64415.50	40.00	77	23332.00	1484.00	32491.50	1579.00
30	66311.00	71.00	64476.50	37.00	78	20987.50	1488.00	30287.50	1715.00
31	66111.00	99.00	64526.00	41.00	79	18440.50	1290.00	27841.50	1704.00
32	65994.00	77.00	64630.00	52.00	80	15972.00	1199.00	25308.50	1796.00
33	65963.00	93.00	64775.00	65.00	81	13771.50	1148.00	22870.00	1856.00
34	65564.00	111.00	64497.50	41.00	82	12148.50	1169.00	20833.00	2095.00
35	64618.50	109.00	63797.00	75.00	83	10820.00	1040.00	18958.50	2096.00
36	63513.00	123.00	63059.50	72.00	84	9785.50	1094.00	17317.00	2256.00
37	62503.50	102.00	62347.00	68.00	85	8764.50	1074.00	15721.50	2325.00
38	61467.00	117.00	61572.50	69.00	86	7716.50	1021.00	14113.00	2332.00
39	60471.00	128.00	60830.50	83.00	87	6670.50	973.00	12400.50	2306.00
40	59447.50	136.00	59983.00	89.00	88	5589.50	845.00	10596.00	2141.00
41	58063.00	128.00	58719.00	85.00	89	4631.00	733.00	8933.00	2051.00
42	56272.50	164.00	57007.00	90.00	90	3737.00	593.00	7320.00	1809.00
43	54271.00	162.00	55168.50	104.00	91	2995.50	550.00	5883.00	1654.00
44	52509.00	165.00	53575.00	102.00	92	2301.00	462.00	4558.50	1396.00
45	51236.50	165.00	52366.50	122.00	93	1702.50	334.00	3416.50	1097.00
46	50214.50	173.00	51369.50	118.00	94	1269.50	234.00	2535.00	936.00
47	49358.50	202.00	50587.50	116.00	95	852.50	179.00	1740.00	667.00
48	48993.50	188.00	50376.00	133.00	96	576.50	130.00	1148.50	530.00

MEN												
Poisson (GLM)		**GM(0,2)**	**GM(0,3)**	**GM(0,4)**	**GM(0,5)**	**GM(0,6)**	**GM(0,7)**	**GM(0,8)**	**GM(0,9)**	**GM(0,10)**	**GM(0,11)**	**GM(0,12)**
deviance	102212	2144.98	960.44	793.61	782.45	636.61	519.51	379.36	227.25	174.14	165.85	165.35
d.f.	96	95	94	93	92	91	90	89	88	87	86	85
log-likelihood	218846.9	217774.4	218366.7	218450.1	218455.7	218528.6	218587.1	218657.2	218733.3	218759.8	218764	218764.2
χ^2		13552.89	2348.77	1202.75	1058.72	685.13	530.36	380.66	224.14	176.54	170.07	169.22
Poisson (GNLM)			**GM(1,2)**		**GM(3,2)**	**GM(4,2)**	**GM(5,2)**	**GM(6,2)**	**GM(7,2)**	**GM(8,2)**	**GM(9,2)**	**GM(10,2)**
deviance	102212		862.7652		736.7106	605.2137	342.1944	311.1831	339.0331	330.6377	318.0727	274.3146
d.f.	96		95		94	93	92	91	90	89	88	87
log-likelihood	218846.9		218415.5		21478.5	218544.3	218675.8	218691.3	218677.4	218681.6	218687.9	218709.7
χ^2			1378.717		1066.584	803.3898	347.6592	280.7295	338.2825	324.1219	309.911	268.9084
Gamma (GLM)		**GM(0,2)**	**GM(0,3)**	**GM(0,4)**	**GM(0,5)**	**GM(0,6)**	**GM(0,7)**	**GM(0,8)**	**GM(0,9)**	**GM(0,10)**	**GM(0,11)**	**GM(0,12)**
deviance	102212	2311.18	1426.80	1423.07	1547.64	1392.25	1156.79	731.03	323.09	271.37	279.22	278.12
d.f.	96	95	94	93	92	91	90	89	88	87	86	85
log-likelihood	218846.9	217691.3	218133.5	218135.4	218073.1	218150.8	218268.5	218481.4	218685.4	218711.2	218707.3	218707.8
χ^2		10391.08	7198.96	7846.74	6854.49	3499.79	1885.87	933.81	381.43	318.12	331.78	331.58

WOMEN												
Poisson (GLM)		**GM(0,2)**	**GM(0,3)**	**GM(0,4)**	**GM(0,5)**	**GM(0,6)**	**GM(0,7)**	**GM(0,8)**	**GM(0,9)**	**GM(0,10)**	**GM(0,11)**	**GM(0,12)**
deviance	172821.4	4481.88	1133.22	863.537	862.63	409.18	324.47	283.97	185.99	155.38	113.46	113.46
d.f.	96	95	94	93	92	91	90	89	88	87	86	85
log-likelihood	312737.5	310496.6	312170.9	312305.8	312306.2	312532.9	312575.3	312595.6	312644.5	312659.8	312680.8	312680.8
χ^2		100664.9	2573.70	1151.41	1198.49	414.03	317.73	276.38	178.98	152.13	111.54	111.53
Poisson (GNLM)			**GM(1,2)**		**GM(3,2)**	**GM(4,2)**	**GM(5,2)**	**GM(6,2)**	**GM(7,2)**	**GM(8,2)**	**GM(9,2)**	**GM(10,2)**
deviance	172821.4		1129.71		736.71	611.27	603.07	352.42	337.97	270.23	262.90	215.51
d.f.	96		95		94	93	92	91	90	89	88	87
log-likelihood	312737.5		312172.7		310226.6	312431.9	312436	312561.3	312568.6	312602.4	312606.1	312629.8
χ^2			1782.39		1066.58	787.88	771.41	390.49	362.17	274.40	261.50	206.54
Gamma (GLM)		**GM(0,2)**	**GM(0,3)**	**GM(0,4)**	**GM(0,5)**	**GM(0,6)**	**GM(0,7)**	**GM(0,8)**	**GM(0,9)**	**GM(0,10)**	**GM(0,11)**	**GM(0,12)**
deviance	172821.4	11628.81	1694.593	1691.821	2014.139	1080.164	1018.142	897.1084	408.2796	203.1785	173.3137	159.9273
d.f.	96	95	94	93	92	91	90	89	88	87	86	85
log-likelihood	312737.5	306923.1	311890.2	311891.6	311730.5	312197.5	312228.5	312289	312533.4	312635.9	312650.9	312657.6
χ^2		29977.59	7965.66	9101.77	8058.54	2837.94	1600.56	1048.8	438.68	220.34	188.99	180.21

***Table 2**: Goodness of fit measurements for the different models of μ_x.*

inter-census estimations obtained from various INE publications, (INE, 1997) and (INE, 2001).

The first step is to calculate the crude estimates of q_x from these data. From among the different existing proposals for carrying out such estimates, we have used that of Navarro *et al.* (1995):

$$\dot{q}_x = \frac{D_{x(t-1)} + D_{xt}}{1/2P_{x(t-1)} + P_{xt} + 1/2P_{x(t+1)} + 1/2(D_{x(t-1)} + D_{xt})}, \tag{6}$$

where P_{xt} is the population of people whose ages are between x and $x + 1$ years old on 1st January of the year t, and D_{xt} is the number of individuals deaths whose ages were between x and $x + 1$ during the year t. This choice is made because as we do not have the deaths classified according to the year of birth, but according to age and sex, the expression (6) allows us to avoid this difficulty because it supposes uniform death distribution throughout the year. The denominator of the expression is e_x, an estimation of E_x^i.

The same expression, adequately corrected in denominator, can be used for the crude estimation of μ_x,

$$\dot{\mu}_x = \frac{D_{x(t-1)} + D_{xt}}{1/2P_{x(t-1)} + P_{xt} + 1/2P_{x(t+1)}}, \tag{7}$$

where the denominator is now $e_x - d_x/2$, an estimation of E_x^c.

The graphic representation of the logarithms of the crude estimations led us to take a range of between 0 to 96 years old for age, which seems to us compatible with the use of the maximum possible and with the demand for relatively stable behavior. Beyond this age the logarithms decrease, showing behaviour which is difficult to explain.

In the period under study, there were nearly 3.96 million men and 4.11 million women exposed to risk. 77% of these were over 20 years of age. In the same period, nearly 39.3 thousand men and 51.4 women died, with the great majority, approximately 99%, doing so after the age of 20 (see Table 1).

3.1 Modelling μ_x

The modelling μ_x has been done by means of $GM(r, s)$ functions, using GLM and generalized non-linear models (GNLM) of the Poisson and Gamma families. The goodness of fit of the models involved has been measured by means of the log-likelihood and the χ^2. Since the fitting must improve as the number of parameters increase, we must to see if that improvement is significant and to do so we use the deviance and Mallow's C_p statistic, both testing the improvement of the fitting in relation to the increase in complexity of the model.

Table 2 summarizes the results obtained. It is divided into two parts, according to sex, and then each part into three groups of results.

1. Those corresponding to the functions $GM(0,s)$, $s = 2,\ldots,12$, fitted through the generalized linear models of the Poisson family using the log as a link.
2. Those corresponding to the functions $GM(r,2)$, $r = 1,\ldots,10$, fitted through generalized non linear models of the Poisson family using the identity as a link;
3. Those corresponding to the functions $GM(0,s)$, $s = 2,\ldots,12$ fitted through generalized linear models of the Gamma family using the log as a link, even though what we adjust in this case is the vitality force, $1/\mu_x$.

The first column of the table contains the initial reference values corresponding to the null model for the deviance and to the saturated model for the log-likelihood. We conclude that the best model is $GM(0,11)$, obtained through generalized linear models using the Poisson family ($GM(0,12)$ has an insignificant improvement of deviance. Once its coefficients have been calculated, we test that they are significant for both sexes in Table 3.

We should point out that for making the results of Poisson and Gamma models comparable, we have evaluated the inverse of the Gamma model predictions. Thus, the results shown in Table 2 have been calculated from the Poisson Likelihood obtained with these inverses.

Figure 2 shows the graphic comparison of the $GM(0,s)$ models, from $s = 7$ to $s = 12$ for each sex. In order to make the results obtained with all models and functions comparable, the above comparison is made in terms of q_x in place of the fitted μ_x.

Table 3: *Coefficients of models* $GM(0,11)$

	MEN[a]				**WOMEN**[b]			
	coef	**std error**	**p-value**	**t-value**	**coef**	**std error**	**t-value**	**p-value**
const.	-5.439e+00	1.046e-01	-51.985	< 2e-16	-5.369e+00	8.434e-02	-63.655	< 2e-16
age	-1.874e+00	1.613e-01	-11.614	< 2e-16	-1.902e+00	1.313e-01	-14.481	< 2e-16
age^2	3.165e-01	3.776e-02	8.383	8.85e-13	3.430e-01	3.137e-02	10.935	< 2e-16
age^3	-2.428e-02	3.788e-03	-6.409	7.55e-09	-2.917e-02	3.174e-03	-9.192	2.00e-14
age^4	1.061e-03	2.081e-04	5.098	2.02e-06	1.418e-03	1.742e-04	8.139	2.77e-12
age^5	-2.883e-05	6.910e-06	-4.172	7.20e-05	-4.257e-05	5.743e-06	-7.412	8.04e-11
age^6	5.060e-07	1.448e-07	3.494	0.000755	8.187e-07	1.191e-07	6.875	9.34e-10
age^7	-5.756e-09	1.929e-09	-2.984	0.003703	-1.012e-08	1.566e-09	-6.461	5.99e-09
age^8	4.106e-11	1.584e-11	2.592	0.011210	7.776e-11	1.268e-11	6.134	2.54e-08
age^9	-1.672e-13	7.316e-14	-2.285	0.024762	-3.386e-13	5.767e-14	-5.871	7.94e-08
age^{10}	2.968e-16	1.454e-16	2.041	0.044275	6.384e-16	1.128e-16	5.657	1.98e-07

a. deviance= 165.85 on 86 d. f.; over-dispersion parameter $\phi = 1.977649$

b. deviance= 113.46 on 86 df; over-dispersion parameter $\phi = 1.296974$

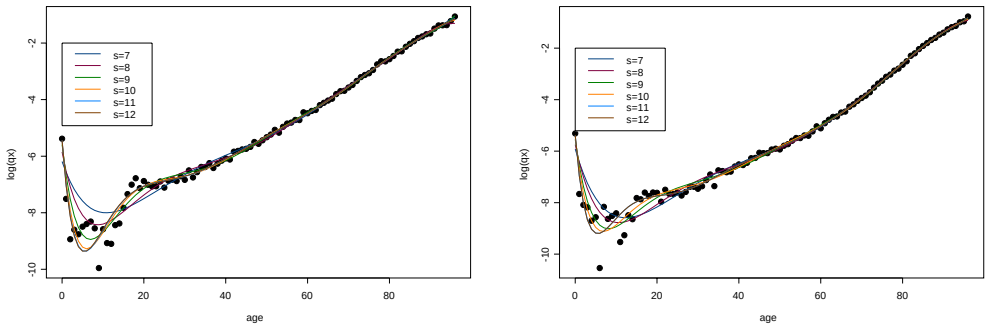

Figure 2: *Comparison of the q_x corresponding to the models GM(0,s) for men and women.*

3.2 Modelling q_x

The modelling q_x has been done through the functions $LGM(0, s)$, $s = 2, \ldots, 12$, using generalized linear models of the binomial family, and through Heligman and Pollard's second law for whose estimation we have used weighted least squares. Table 4 summarizes the results obtained. The first column of the table contains the initial reference values corresponding to the null model for the deviance and to the saturated model for the log-likelihood. The observed values indicate that the best model for both sexes, taking into consideration the commitment between goodness of fit and its complexity, is $LGM(0, 11)$. The coefficients of these models for both sexes are shown in Table 5.

Figure 3 shows the graphic comparison of the models from $s = 7$ to $s = 12$ for each sex. Both are presented in *logit* scale.

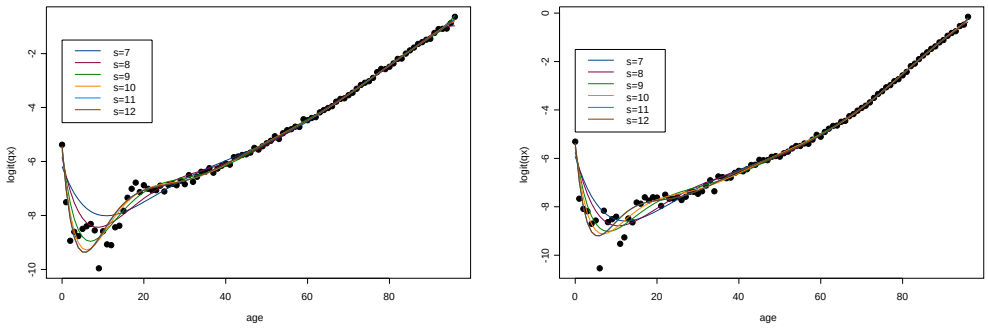

Figure 3: *Comparison of models LGM (0,s) for men and women*

	MEN				WOMEN			
		LGM(0,2)	**LGM(0,3)**	**LGM(0,4)**		**LGM(0,2)**	**LGM(0,3)**	**LGM(0,4)**
deviance	102248	2332.71	890.30	794.34	172974.9	5080.83	886.03	806.22
d.f.	96	95	94	93	96	95	94	93
log-likelihood	-169352.1	-170518.5	-16797.3	-16749.3	-190038.2	-192578.6	-190481.2	-190441.3
χ^2		14214.66	2042.38	1226.10		121027.5	1839.7	1195.73
	LGM(0,5)	**LGM(0,6)**	**LGM(0,7)**	**LGM(0,8)**	**LGM(0,5)**	**LGM(0,6)**	**LGM(0,7)**	**LGM(0,8)**
deviance	769.87	635.95	517.09	375.22	792.99	408.35	323.94	279.36
d.f.	92	91	90	89	92	91	90	89
log-likelihood	-169737.1	-169670.1	-169610.7	-169539.7	-190434.7	-190242.3	-190200.1	-190177.8
χ^2	1000.90	683.45	527.72	371.51	1047.88	411.94	317.108	268.03
	LGM(0,9)	**LGM(0,10)**	**LGM(0,11)**	**LGM(0,12)**	**LGM(0,9)**	**LGM(0,10)**	**LGM(0,11)**	**LGM(0,12)**
deviance	225.16	173.95	166.14	165.59	186.84	155.37	114.16	114.10
d.f.	88	87	86	85	88	87	86	85
log-likelihood	-169464.7	-169439.1	-169435.2	-169434.9	-190131.6	-190115.8	-190095.2	-190095.2
χ^2	221.32	175.54	169.22	168.31	179.15	151.64	111.77	111.82

***Table 4**: Goodness of fit measures for the $LGM(0,s)$ of q_x*

Table 5: *Coefficients of models* $LGM(0, 11)$

	MEN[a]				WOMEN[b]			
	coef	**std error**	**t-value**	**p-value**	**coef**	**std error**	**t-value**	**p-value**
const.	-5.438e+00	1.049e-01	-51.825	< 2e-16	-5.367e+00	8.476e-02	-63.316	< 2e-16
age	-1.875e+00	1.623e-01	-11.549	< 2e-16	-1.913e+00	1.328e-01	-14.401	< 2e-16
age^2	3.167e-01	3.814e-02	8.305	1.27e-12	3.467e-01	3.190e-02	10.868	< 2e-16
age^3	-2.430e-02	3.842e-03	-6.326	1.09e-08	-2.962e-02	3.244e-03	-9.132	2.65e-14
age^4	1.063e-03	2.119e-04	5.015	2.82e-06	1.446e-03	1.789e-04	8.081	3.63e-12
age^5	-2.890e-05	7.063e-06	-4.092	9.63e-05	-4.360e-05	5.928e-06	-7.355	1.05e-10
age^6	5.078e-07	1.486e-07	3.417	0.000968	8.421e-07	1.235e-07	6.817	1.21e-09
age^7	-5.783e-09	1.986e-09	-2.911	0.004583	-1.045e-08	1.632e-09	-6.402	7.79e-09
age^8	4.131e-11	1.637e-11	2.523	0.013468	8.059e-11	1.327e-11	6.072	3.32e-08
age^9	-1.684e-13	7.586e-14	-2.220	0.029045	-3.522e-13	6.066e-14	-5.806	1.05e-07
age^{10}	2.994e-16	1.513e-16	1.980	0.050947	6.662e-16	1.192e-16	5.588	2.66e-07

a. deviance= 166.14 on 86 d. f.; over-dispersion parameter $\phi = 1.980545$

b. deviance= 114.16 on 86 d. f.; over-dispersion parameter $\phi = 1.304967$

The graduation results for Heligman and Pollard's second law are presented graphically in Figure 4. The criterion used for weighting the square difference was the first in 5), the choice being based on the number of relative deviations greater than 2 and 3 and the value of the χ^2 for the goodness of fit.

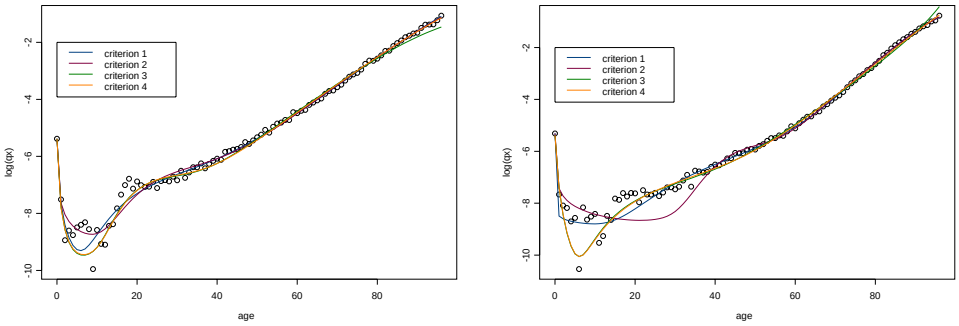

Figure 4: *Comparison of Heligman and Pollard's models for men and women*

The coefficients corresponding to the Heligman-Pollard model have not presented great difficulty for men. This was not the case for women because they do not present the accident hump. The Spanish female population has high mortality spread over many more years (Felipe and Guillén, 1999). The problem was solved by fixing the parameter $F = 96$. This technique of fixing the values of some parameters and fitting the rest was used by Congdon (1993). In order not to fall into the problem pointed out by Congdon (1993), we have also carried out a study of the meaningfulness of the parameters. Some problems related to the singularity of the matrix of the coefficients were found

in this study, and have been overcome through the use of generalized non-linear least squares. The parameter estimates are shown in Table 6, some of them not significant, in particular, *A*, *B*, *C*, *D* and *E* for men, and *B* and *C* for women.

Table 6: *Coefficients of Heligman and Pollard models*

	MEN				**WOMEN**			
	coef	**std error**	**t-value**	**p-value**	**coef**	**std error**	**t-value**	**p-value**
A	0.00054	0.00035	1.5670	0.1207	0.000335	0.0000746	4.4886	<0.0001
B	0.12921	0.21037	0.6142	0.5407	0	0.0000030	0.1073	0.9148
C	0.16301	0.09059	1.7993	0.0754	0.027444	0.0153661	1.7860	0.0775
D	0.00138	0.00065	2.1289	0.0361	0.002757	0.0003135	8.7939	<0.0001
E	0.74764	0.44753	1.6706	0.0984	1.140099	0.1308293	8.7144	<0.0001
F	63.03293	37.14621	1.6969	0.0933	96	—	—	—
G	0.00002	0	3.4307	0.0009	0.000001	0.0000001	4.5996	<.0001
H	1.11313	0.00425	262.0285	< .0001	1.159430	0.0031811	364.4753	<.0001
K	0.91755	0.26233	3.4977	0.0007	1.108379	0.0822712	13.4723	<.0001

4 Comparison of the models

We have compared the different models by choosing the best fitting model for each one of them. Specifically, we have compared the $GM(0,11)$ for μ_x, the $LGM(0,11)$ for q_x and Heligman and Pollard's second law (HP) for q_x. In order to make the first one comparable to the other two, the values of μ_x have been transformed through the relation

$$q_x = 1 - \exp(-\mu_{x+\frac{1}{2}}).$$

The comparison is carried out by applying the tests proposed by Forfar *et al.* (1988), which Navarro (1991) and Navarro *et al.* (1995) also used in their work. In order to obtain an expected number of deaths not inferior to 5, we have had to aggregate data for ages between 4 and 10, with the consequent decrease in the number of degrees of freedom.

We have also obtained the values of the mean absolute percentage error (MAPE) and R^2 that Felipe and Guillén (1999) used in their work. The value of R^2 has been obtained as 1 minus the proportion of the variance that remains unexplained, because if we calculate it directly as a percentage of explained variance, in some cases it exceeded 1. This can happen when the models are not linear.

Table 7 presents the results of the tests for the three models. Figure 5 shows the autocorrelations of standarized residuals for all the models. In all the cases there are a few isolated correlated values out of the Heligman and Pollard model adjusted for women. This agrees with the worst behaviour of this adjustment.

***Table 7**: Comparison of the three best fitted parametric models.*

		GM(0,11) for μ_x		**LGM(0,11) for q_x**		**HP for q_x**	
		Men	Women	Men	Women	Men	Women
Relative Desv. [a]	> 2	8	3	8	4	6	10
	> 3	3	0	3	0	6	2
Signs test	pos.(neg.)	46 (48)	53 (42)	46 (48)	53 (42)	50 (44)	47 (50)
	p-value	0.4589	0.8909	0.4589	0.8909	0.7647	0.4196
Runs test	runs	44	50	44	50	41	35
	p-value	0.4319	0.5372	0.4319	0.5372	0.3839	0.2731
K-S test[b]	K-S	0.0433	0.0316	0.0426	0.0316	0.0532	0.0825
	p-value	1	1	0.9994	1	1	0.8987
χ^2 **test** [c]	χ^2	164.32	102.44	164.18	101.07	224.31	165.56
	d.f.	83	84	83	84	85	88
	p(χ^2)	2.69e-07	0.0836	2.80e-07	0.0989	1.66e-04	1.11e-06
$\mathbf{R^2}$		0.9972	0.9991	0.9972	0.9991	0.9967	0.9981
MAPE		16.34	16.44	16.35	16.44	15.13	21.17

a. standarized residuals

b. Kolmogorov-Smirnov test

c. χ^2 statistic, sum of squared standarized residuals

5 Conclusions

From Table 7 and Figure 5, we can conclude that

1. Heligman and Pollard's models fit worse than the other two,
2. Women provide a better fitting in the three models, and
3. The model $LGM(0, 11)$ provides the most acceptable results for both sexes.

In relation to the work of other authors, we should highlight two distinctive features of the methodology presented here:

- The first one is the possibility of comparing the different models, as all of them end up producing estimates of q_x and are susceptible to having their goodness of fit measured with the same criteria.
- The second one is that all the models have been fitted for the full range of ages without the need to recur to a division into sections of that range. In this respect, it is interesting to compare our best model, the $LGM(0, 11)$, with that obtained by other authors for data of the same origin (Navarro et al, 1995). This comparison can be seen in Debón *et al.* (2003). They obtain a slightly better fit, but the resulting function presents irregularities (peaks) in the junction points between the sections due to the restrictions imposed on the functions to be fitted in each section. Moreover, the fitting of a single function entails a great saving of time.

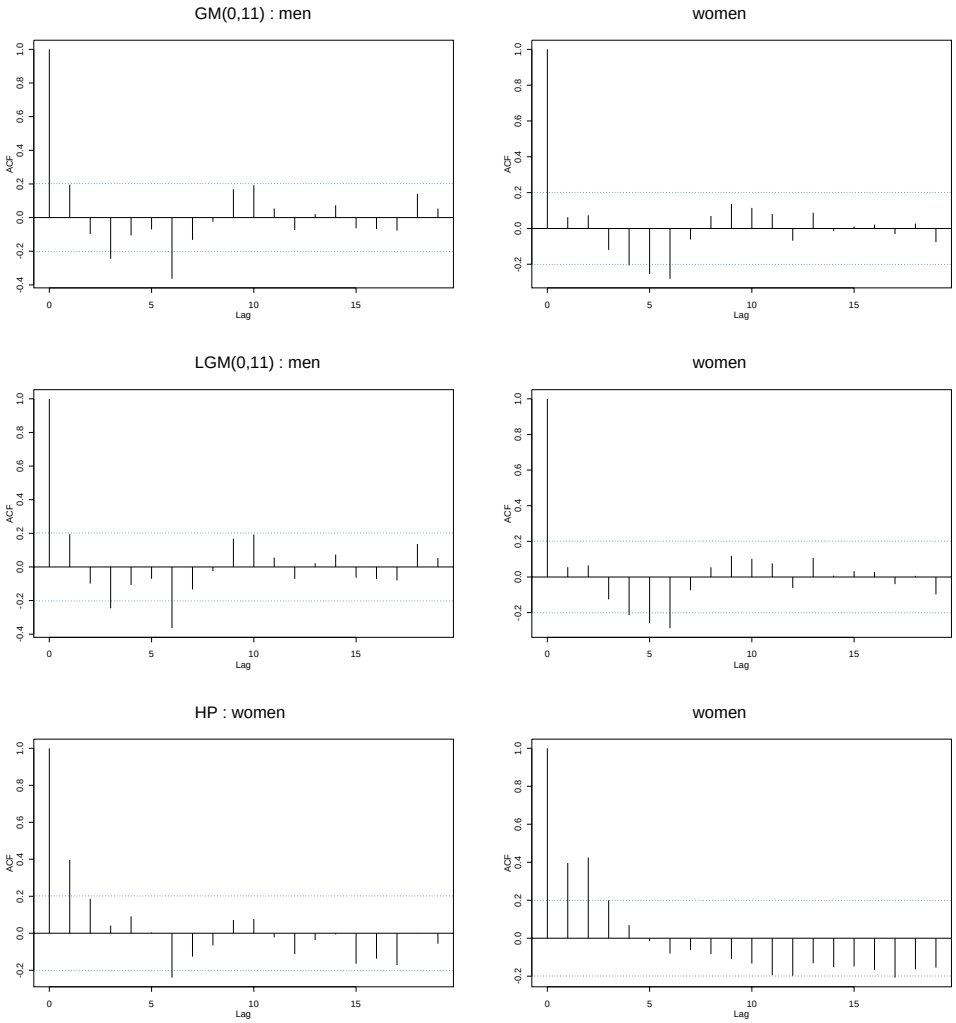

Figure 5*: *Autocorrelations of standarized residuals

We should point out that all the models present problems for younger ages due to the irregular profile of crude mortality rates. We can observe a greater distance between the values predicted by the models and the observations for the lower ages in Figures 2, 3 and 4. This is a well-known problem when graduating mortality data. Many authors achieve better fits by eliminating this group of ages, which they justify by arguing that the actuarial operations begin at a more advanced age.

Contrary to this criterion, we have decided to include the young ages groups for two reasons. The first one is that it enables as to compare our results with those obtained by Navarro *et al.* (1995), who graduate mortality data for the Valencia Region for the years 1990-92 for the complete range of ages. As far as we know, that is the only study that covers the same geographical area as ours and a comparison is vital. The second argument in favour of our approach is to remember that the double exponential which appears in Heligman and Pollard's laws was introduced specifically to deal with the difficulty of adjusting young age groups.

Finally, in line with the work of Butt and Haberman (2004), we conclude that the GLM method has a stronger theoretical justification and yields models with more favourable properties than the classical non-linear least squares method.

References

Benjamin, B. and Pollard, J. (1992). *The Analysis of Mortality and Other Actuarial Statistics*, 6th edition. London: Butterworth-Heinemann.

Butt, Z. and Haberman, S. (2004). Application of frailty-based mortality models using generalized linear models. *Astin Bulletin*, 34, 175-197.

Congdon, P. (1993). Statistical graduation in local demographic analysis and projection. *Journal of the Royal Statistical Society A*, 156, 237-270.

Debón, A., Montes, F., and Sala, R. (2003). Graduación de datos de mortalidad. In *Actas del 27 Congreso Nacional de Estadística e Investigación Operativa*, Lleida, España. Universitat de Lleida, 562-578.

Dellaportas, P., Smith, A., and Stavropoulos, P. (2001). Bayesian analysis of mortality data. *Journal of the Royal Statistical Society A*, 164, 275-291.

Felipe, M. and Guillén, M. (1999). *Evolución y Predicción de las Tablas de Mortalidad Dinámicas para la Población Española*. Cuadernos de la Fundación, Fundación Mapfre Estudios.

Forfar, D., McCutcheon, J., and Wilkie, A. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, 115, 1-149.

Gavin, J., Haberman, S., and Verrall, R. (1993). Moving weighted average graduation using kernel estimation. *Insurance: Mathematics and Economics*, 12, 113-126.

Gavin, J., Haberman, S., and Verrall, R. (1995). Graduation by kernel and adaptive kernel methods with a boundary correction. *Transactions. Society of Actuaries*, XLVII, 173-209.

Gerber, H. (1997). *Life Insurance Mathematics*. Berlin: Springer-Verlag.

Gompertz, B. (1825). On the nature of the function of the law of human mortality and on a new mode of determining the value of life contingencies. *Philosophical Transactions of The Royal Society*, 115, 513-585.

Haberman, S. and Renshaw, A. (1996). Generalized linear models and actuarial science. *The Statistician*, 45, 407-436.

Heligman, L. and Pollard, J. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 49-80.

INE (1997). *Evolución de la población de España entre los censos de 1981 y 1991*. Madrid: Instituto Nacional de Estadística.

INE (2001). *Evolución de la población de España entre los censos de 1991 y 2001*. Madrid: Instituto Nacional de Estadística.

Lambert, J. (1772). Anmerkungen über die Sterblichkeit, Todtenlisten, Gerburthen und Ehen. *in Beyträge*, 3, 475-599.

London, D. (1985). *Graduation: the Revision of Estimates*. Actex Publication, Winsted, Cunnecticud.

Makeham, W. (1860). On the law of mortality. *Journal of the Institute of Actuaries*, 13, 325-358.

Navarro, E. (1991). *Tablas de Mortalidad de la Población Española 1982. Metodología y Fuentes*. Madrid: Mapfre.

Navarro, E., Ferrer, R., Gonzalez, C., and Nave, J. (1995). *Tablas de Mortalidad de la Comunidad Valenciana 1990-91. Censos de Població i Habitatges*, volume I. IVE, Valencia.

Renshaw, A. (1991). Actuarial graduation practice and generalised linear models. *Journal of the Institute of Actuaries*, 118, 295-312.

Renshaw, A., Haberman, S., and Hatzopoulos, P. (1997). On the duality of assumptions underpinning the construction of life tables. *Astin Bulletin*, 27, 5-22.

Renshaw, A. and Hatzopoulos, P. (1996). On the graduation of amounts. *British Actuarial Journal*, 2, 185-205.

Thiele, P. (1972). On a mathematical formula to express the rate of mortality throughout the whole of life. *Journal of the Institute of Actuaries*, 16, 313-329.

Verrall, R. (1996). A unified framework for graduation. *Actuarial Research Paper*, 91, 2-25.

Book reviews

MODELS FOR DISCRETE LONGITUDINAL DATA

G. Molenberghs and G. Verbeke

Springer, New York, 2005
ISBN 0-387-25144-8
pp. 687 + XXII, 61 illustrations, hardcover

This book covers a wide variety of statistical techniques for longitudinal data analysis. The authors, Geert Molenberghs and Geert Verbeke –both well known in this field– have extended their previous textbook (Verbeke and Molenberghs, 1997), mainly focused on linear mixed model for continuous data, to the non-Gaussian setting, including binary, ordinal, and counts repeated measures.

The book has 32 chapters divided in six main sections. It starts (Section I: Chapters 1 to 5) by providing a general perspective of generalised linear models and extensions to linear mixed models for Gaussian longitudinal data. Following sections are focussed on the special non-linear models, showing and examining differences between the classes of marginal (Section II: Chapters 6 to 10), conditional (Section III: Chapters 11 and 12) and subject-specific (Section IV: Chapters 13 to 16) models. In these sections, approximate numerical methods are shown for each model, including a description of its advantages and limitations. Many practical examples are provided in these sections. In addition, Section V presents a set of case-studies (Chapters 17 to 20) showing how different problems, mainly from the pharmaceutical and medical research field, may call for different models presented in previous chapters. Finally, the authors show how to deal with missing data in longitudinal studies (Section IV, Chapters 26 to 32).

The book is clearly written, and the theoretical bases for the different models described in all sections are comprehensively treated. However, the authors prefer to emphasize practice rather than mathematical enhancements. For this reason, a large number of practical examples using SAS procedures, such as MIXED, GENMOD GLIMMIX, NLMIXED, MI, and MIANALYZE, are presented as illustrations. A limited number of examples are also analysed using other statistical software, such as S-Plus or MLwiN. Selected programs, macros and datasets used as examples in the book are available at the authors' web site (http://www.censtat.be/research/software.asp).

In summary, this book provides very useful guidance and advice on practical issues when dealing with longitudinal studies, specially for non-Gaussian data, and is an essential and highly recommendable reference for applied statisticians and other researchers in this field.

References

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models for Longitudinal Data*. Springer: New York.

Aurelio Tobías
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.es) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a LaTeX 2_ε.

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name _____
_____
Organisation _____
_____
Street Address _____
_____
Zip/Postal code _____ City _____
State/Country _____ Tel. _____
Fax _____ NIF/VAT Registration Number _____
E-mail _____
Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2005 (volume 29)

Annual subscription rates:
– Spain: €22 (VAT included)
– Other countries: €25 (VAT included)

Price for individual issues (current and back issues):
– Spain: €9/issue (VAT included)
– Other countries: €11/issue (VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number
☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT *(Statistics and Operations Research Transactions)*

The undersigned _____

authorises Bank/Financial institution _____

located at (Street Address) _____

Zip/postal code _____ City _____

Country _____

to draft the subscription to **SORT *(Statistics and Operations Research Transactions)*** from my account

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date _____

Signature

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Quatre modalitats de subscripció al DOGC

(Diari Oficial de la Generalitat de Catalunya)

Generalitat de Catalunya

Imprès, edició diària

Base de dades, actualització diària

DVD, edició semestral

A la carta, servei diari personalitzat

A més, per als subscriptors de l'edició impresa i del DVD, tramesa gratuïta d'un CD-ROM trimestral que conté les pàgines en format PDF (DOGC en imatges)

L'Administració més a prop

EADOP • Informació i subscripcions • Rocafort, 120 - Calàbria, 147 • 08015 Barcelona
Tel. 93.292.54.17 • Fax 93.292.54.18 • subsdogc@gencat.net • www.gencat.net/eadop

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009

29002
9 771696 228009