Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 177 a 195

On the Student-t Mixture Inverse Gaussian Model with an Application to Protein Production

Sobre el modelo gaussiano inverso mezclado t-Student y una aplicación a producción de proteínas

Antonio Sanhueza[1,a], Víctor Leiva[2,b], Liliana López-Kleine[3,c]

[1]Departamento de Matemática y Estadística, Universidad de La Frontera, Temuco, Chile

[2]Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Valparaíso, Chile

[3]Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

In this article, we introduce a mixture inverse Gaussian (MIG) model based on the Student-t distribution and apply it to bacterium-based protein production for food industry. This model is mainly useful to describe data that follow positively skewed distributions and accommodate atypical observations in a better way than its classical version. Specifically, we present a characterization of the MIG-t distribution. In addition, we carry out a hazard analysis of this distribution centered mainly on its hazard rate. Furthermore, we discuss the maximum likelihood method, which produces–in this case–robust parameter estimates. Moreover, to evaluate the potential influence of atypical observations, we produce a diagnostic analysis for the model. Finally, we apply the obtained results to novel bacterium-based protein production data and statistically compare two types of protein producers using the likelihood ratio test based on the MIG-t model as an alternative methodology to the procedures available until now. This fact is very important, since the evaluation of protein production using both constructions allows practitioners to choose the most productive one before the bacterial culture is scaled to an industrial level.

***Key words*:** Distribution mixture, Length-biased, Likelihood methods, distributions, R computer language.

[a] Professor. E-mail: asanhueza@ufro.cl

[b] Professor. E-mail: victor.leiva@uv.cl

[c] Assistant professor. E-mail: llopezk@unal.edu.co

Resumen

En este artículo, introducimos un modelo Gaussiano inverso (MIG) mezclado basado en la distribución t-Student y lo aplicamos a la producción de proteínas basada en bacterias para la industria de alimentos. Este modelo es especialmente útil para describir datos que siguen una distribución con sesgo positivo ya que permite acomodar observaciones atípicas de mejor forma que su versión clásica. Específicamente, presentamos una caracterización de la distribución MIG-t y realizamos un análisis de confiabilidad de esta distribución centrado principalmente en la tasa de fallas. También, discutimos el método de verosimilitud máxima, el cual proporciona en este caso estimaciones robustas de los parámetros del modelo. Con el fin de evaluar la influencia potencial de observaciones atípicas, proponemos un análisis de diagnóstico para la distribución. Finalmente, aplicamos los resultados obtenidos al análisis de datos nuevos de producción de proteína basada en bacterias utilizada en la industria de alimentos y comparamos estadísticamente dos tipos de bacterias productoras usando la prueba de razón de verosimilitudes basada en el modelo MIG-t como una metodología alternativa a los procedimientos disponibles a la fecha. Este punto es muy importante, ya que la evaluación de producción de proteínas usando dos construcciones distintas permite a los investigadores escoger el tipo más productivo antes de proceder al cultivo industrial a gran escala.

***Palabras clave*:** distribuciones de largo sesgado, lenguaje de computación R, métodos de verosimilitud, mezcla de distribuciones.

1. Introduction

The normal distribution has been a reference model in statistics for over one hundred years. Its attractive properties are well-known and widely used in statistical theory and practice. However, inference upon normality is vulnerable to atypical data, which are found in several fields. Specifically, the parameter estimators of the normal model obtained with the maximum likelihood (ML) method are sensitive to atypical observations. Lange, Little & Taylor (1989) proposed to use the Student-t distribution for solving this problem of sensitivity, since it has greater kurtosis than the normal distribution. Thus, such atypical cases could be accommodated in a better way by using the t model than the normal model. Moreover, as it can be seen in Figure 1, the degree of kurtosis of the t model is flexible and then it can appropriately model different quantities and magnitudes of atypical data. For these reasons, the t model has been used as an alternative to the normal model to obtain qualitatively robust estimates, which is a first concept of robustness. See Lucas (1997) and Montgomery, Peck & Vining (2001, pp. 381-413). Specifically, robustness studies the sensitivity of the results of a statistical analysis to deviations in the assumptions that validate this analysis.

A random variable (r.v.) X following the t distribution with ν degrees of freedom, denoted by $X \sim t(\nu)$, has probability density function (p.d.f.) and

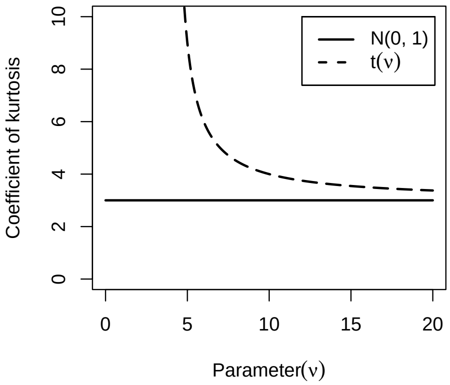

Figure 1: Coefficients of kurtosis of the normal and t distributions.

cumulative distribution function (c.d.f.) respectively given by

$$\phi_t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\,\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\,\left[1+\frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}} \quad \text{and} \quad \Phi_t(x) = \frac{1}{2}\left[1+\mathrm{I}_r\left(\frac{1}{2},\frac{1}{2}\,\nu\right)\right], \;\; x\in\mathbb{R},\, \nu>0$$

where $\mathrm{I}_r(a,b) = [\int_0^x t^{a-1}[1-t]^{b-1}\,\mathrm{d}t]/[\int_0^1 t^{a-1}[1-t]^{b-1}\,\mathrm{d}t]$ is the beta incomplete function ratio, with $r = x^2/[x^2+\nu]$. See Johnson, Kotz & Balakrishnan (1994, pp. 364). Special cases of the t distribution are the Cauchy distribution, when $\nu = 1$, and the normal distribution, when $\nu \to \infty$. The normal and t models are symmetric distributions in the seto of real numbers. However, many phenomena present data whose distributions are asymmetrical, such as occurs frequently in biotechnology and industry data.

A very popular, positively skewed, asymmetric probability model is the inverse Gaussian (IG) distribution, which is also known as the first passage time distribution of the Brownian motion with positive drift. See Schrodinger (1915), Wald (1947) and Tweedie (1957). The inverse Gaussian (IG) and normal distributions are very similar, although these distributions describe different types of data. In fact, Folks (2007) provided a table that contains 42 analogies between these two distributions. The IG distribution has been widely studied. Several books devoted to this distribution have appeared within the last 30 years. See Jorgensen (1982), Chhikara & Folks (1989), Seshadri (1993, 1999) and Johnson et al. (1994, pp. 259-297). Specifically, the IG model is characterized by the mean (μ) and scale (λ) parameters, denoted by $T \sim \mathrm{IG}(\mu,\lambda)$. An r.v. T with IG distribution has p.d.f. and c.d.f. respectively given by

$$f_T(t) = \phi\left(a_t\right)\frac{\sqrt{\lambda}}{\sqrt{t^3}} \quad \text{and} \quad F_T(t) = \Phi\left(a_t\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-b_t\right), \;\; t>0,\, \mu>0,\, \lambda>0$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the N(0,1) p.d.f. and c.d.f. respectively, and

$$a_t(\mu,\lambda) = \frac{\sqrt{\lambda}[t-\mu]}{\mu\sqrt{t}} \quad \text{and} \quad b_t(\mu,\lambda) = \frac{\sqrt{\lambda}[t+\mu]}{\mu\sqrt{t}} \tag{1}$$

According to (1) and for simplicity, we sometimes use the notations $a_t(\mu,\lambda) = a_t$ and $b_t(\mu,\lambda) = b_t$ along the paper. However, when we need to emphasize the dependence of the functions a_t and b_t on μ and λ, we use the notations $a_t(\mu,\lambda)$ and $b_t(\mu,\lambda)$. Some properties of the IG distribution are $c\,T \sim \text{IG}(c\,\mu, c\,\lambda)$, with $c > 0$, and to be a member of the exponential family.

Length-biased distributions are particular cases of the weighted distributions and have interesting properties; see Gupta & Kirmani (1990), Patil (2002) and Leiva, Sanhueza & Angulo (2009). The length-biased inverse Gaussian (LBIG) distribution was presented by Gupta & Akman (1995). However, this result was previously postulated by Jorgensen, Seshadri & Whitmore (1991), although with a distinct denomination, since this was called the distribution of the complementary reciprocal of the IG distribution. Specifically, if $T \sim \text{IG}(\mu,\lambda)$, then the r.v. $L = \mu^2/T$ has a LBIG distribution. In this case, the p.d.f. is given by $f_L(l) = \phi\,(a_l)\,\sqrt{\lambda}/[\mu\sqrt{l}]$, for $l > 0$, $\mu > 0$ and $\lambda > 0$.

Mixture distributions provide powerful and popular tools for generating flexible distributions with attractive statistical and probabilistic properties. See McLachlan & Peel (2000). Specifically, if $0 < p < 1$ is a mixing parameter and $f_{X_1}(x)$ and $f_{X_2}(x)$ are the densities of the variates X_1 and X_2, respectively, then the p.d.f. of the r.v. X expressed by the mixture between X_1 and X_2 is $f_X(x) = [1-p]\,f_{X_1}(x) + p f_{X_2}(x)$, for $x > 0$. Thus, an r.v. M with mixture inverse Gaussian (MIG) distribution obtained from the mixture of the IG and LBIG models has p.d.f. given by

$$f_M(m) = \phi\,(a_m)\,\frac{\sqrt{\lambda}}{\sqrt{m^3}}\left[1 - p + \frac{p\,m}{\mu}\right], \quad m > 0,\, \mu > 0,\, \lambda > 0,\, 0 < p < 1 \tag{2}$$

This is denoted by $M \sim \text{MIG}(\mu,\lambda,p)$. For more details about the MIG distribution and some extensions, see Gupta & Akman (1995), Balakrishnan, Leiva, Sanhueza & Cabrera (2009) and Kotz, Leiva & Sanhueza (2010). We note from (2) that the MIG model is related to the normal model. Thus, by using this relationship, we can define a MIG distribution based on the t model, which we call the MIG-t distribution and should be highly flexible admitting different degrees of kurtosis and asymmetry. In addition, this distribution has parameter estimates that are often non-sensitive to atypical data. Therefore, the MIG-t model can be considered in place of the classic MIG model to produce robust estimation such as occurs with the t and normal models. See Lange et al. (1989). This methodology avoids the use of robust estimation procedures in their classical way, such as Sanhueza, Sen & Leiva (2009) and Leiva, Sanhueza, Sen & Araneda (2010) proposed, by the utilization of the t model in the construction of the MIG distribution.

The IG, LBIG, MIG distributions have been applied in diverse areas, such as actuarial science, agricultural, biotechnology, business and industry, demography, earth sciences, economy and finance, engineering sciences, internet, linguistics, medical sciences, and social and behavior sciences. For more details about these applications, see Chhikara & Folks (1989, pp. 159-184) and Seshadri (1999, pp. 167-316). As mentioned, applications of the IG model in biotechnology and industry have been considered. In this study, we propose a new application to these two fields, which considers bacterium-based protein production for food industry. In general, bacteria are used for the production of proteins with industrial purposes. Simoes-Barbosa, Abreu, Silva-Neto, Gruss & Langella (2004) and Le-Loir, Nouaille, Commissaire, Bretigny, Gruss & Langella (2001) investigated the potential of a lactic acid bacteria called *Lactococcus lactis*, which is a microorganism primarily used in the dairy food industry to produce and secret proteins. Such bacteria can also be used for industrial processes such as meat, wine and dairy. Then, to characterize and compare protein production in different strains and constructions is important by using appropriate statistical distributions and tests. We should explored this aspect because the evaluation of protein production by employing several constructions allows practitioners to choose the most productive one before the bacterial culture is scaled to an industrial level.

The aims of this paper are: (*i*) to introduce the MIG-t distribution as a model that can fit data with high kurtosis, such as it could occur in biotechnology and industry, (*ii*) to carry out a hazard analysis for this distribution centered mainly on the hazard rate (h.r.) and (*iii*) to apply the obtained results to protein production data of *Lactococcus lactis*.

The rest of this article is organized as follows. In Section 2, we provide a probabilistic characterization of the MIG-t distribution and carry out an analysis of its h.r. In Section 3, we estimate the parameters of the MIG-t distribution and make inference about them by using the ML method. Also, in this section, we produce a diagnostic analysis for this distribution to evaluate the potential influence of atypical observations. In Section 4, we apply the obtained results to novel bacterium-based protein production data. In addition, in this section, we statistically compare two types of proteins using the MIG-t distribution by the likelihood ratio (LR) test as an alternative methodology to the classical techniques proposed so far. Finally, in Section 5, we drawn some conclusions.

2. The MIG-t Distribution

In this section, we present and characterize the MIG-t probabilistic model.

2.1. The Probabilistic Model

An r.v. T follows the MIG-t distribution with parameters $\mu > 0$, $\lambda > 0$, $0 < p < 1$ and $\nu > 0$ if and only if its p.d.f. is given by

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\lambda}}{\sqrt{\nu\,\pi}\,\Gamma\left(\frac{\nu}{2}\right)\sqrt{t^3}} \left[1 + \frac{a_t^2}{\nu}\right]^{-\frac{\nu+1}{2}} \left[1 - p + \frac{p\,t}{\mu}\right], \quad t > 0$$

The notation $T \sim$ MIG-$t(\mu, \lambda, p, \nu)$ is used in this case. The following theorem presents some properties of this model.

Theorem 1. *Let* $T \sim$ MIG-$t(\mu, \lambda, p, \nu)$. *Then,*

(*i*) $c\,T \sim$ MIG-$t(c\,\mu, c\,\lambda, p, \nu)$, *with* $c > 0$, *i.e., the MIG-t distribution belongs to the scale family.*

(*ii*) $1/T \sim$ MIG-$t(1/\mu, \lambda/\mu^2, 1 - p, \nu)$, *i.e., the MIG-t distribution belongs to the family closed under reciprocation.*

(*iii*) *The c.d.f. of* T *is* $F_T(t) = \Phi_t(a_t) + [1 - 2p] \int_{b_t}^{\infty} \phi_t\left(\sqrt{u^2 - 4\,\lambda/\mu}\right) du.$

(iv) $U = \frac{\lambda}{\mu}\left[\frac{T}{\mu} + \frac{\mu}{T} - 2\right] \sim \mathcal{F}(1,\nu)$, *i.e., U follows a Fisher distribution with 1 and ν degrees of freedom.*

The following theorems present the mode, denoted by t_m, and the moments of the studied distribution.

Theorem 2. *Let $T \sim \text{MIG-}t(\mu,\lambda,p,\nu)$. Then, the mode of T is given by the solution to*

$$-\frac{\nu+1}{[\nu + a_{t_m}^2]} = \frac{\mu^2 t_m}{\lambda[t_m^2 - \mu^2]}\frac{[3(1-p)\mu + p\,t_m]}{[(1-p)\mu + p\,t_m]}.$$

Theorem 3. *Let $T \sim \text{MIG-}t(\mu,\lambda,p,\nu)$. Then, the first four non-central moments of T are given by*

(i) $\text{E}[T] = \mu + p\,\frac{\mu^2}{\lambda}\,\frac{\nu}{[\nu-2]}$;

(ii) $\text{E}[T^2] = \mu^2 + [1+2p]\,\frac{\mu^3}{\lambda}\,\frac{\nu}{[\nu-2]} + p\,\frac{\mu^3}{\lambda^2}\,\frac{3\nu^2}{[\nu-2][\nu-4]}$;

(iii) $\text{E}[T^3] = \mu^3 + 3[1+p]\,\frac{\mu^4}{\lambda}\,\frac{\nu}{[\nu-2]} + [1+4p]\,\frac{\mu^4}{\lambda^2}\,\frac{3\nu^2}{[\nu-2][\nu-4]} + p\,\frac{\mu^4}{\lambda^3}\,\frac{15\nu^3}{[\nu-2][\nu-4][\nu-6]}$;

(iv) $\text{E}[T^4] = \mu^4 + 2[3+2p]\frac{\mu^5}{\lambda}\frac{\nu}{[\nu-2]} + 5[1+2p]\frac{\mu^5}{\lambda^2}\frac{3\nu^2}{[\nu-2][\nu-4]} + [1+6p]\frac{\mu^5}{\lambda^3}\frac{15\nu^3}{[\nu-2][\nu-4][\nu-6]} + p\frac{\mu^5}{\lambda^4}\frac{105\nu^4}{[\nu-2][\nu-4][\nu-6][\nu-8]}$

***Note* 1.** Observe that if $T \sim \text{MIG-}t(\mu,\lambda,p,\nu)$ and $p = 0, 0.5$ and 1, then we have the IG, Birnbaum-Saunders (BS) and LBIG distributions obtained from the $t(\nu)$ model, respectively. In addition, as mentioned, recall that the standard normal model is obtained as a limiting distribution of the $t(\nu)$ model when $\nu \to \infty$. Thus, the mentioned particular cases correspond to the classical IG, BS and LBIG distributions when $\nu \to \infty$ and $p = 0, 0.5$ and 1, respectively.

2.2. Hazard Analysis and Order Statistics

A hazard is a dangerous event that could conduct to an emergency or disaster. Thus, a hazard is a potential and not an actual possibility, i.e., it can be statistically evaluated, for example, by a useful descriptor known as the hazard rate. This rate for an r.v. $T > 0$ with p.d.f. $f_T(\cdot)$ and c.d.f. $F_T(\cdot)$, is given by

$$h_T(t) = \lim_{\triangle t \to 0}\frac{\mathbb{P}(t < T < \Delta t | T > t)}{\Delta t} = \frac{f_T(t)}{S_T(t)} = -\frac{d\log(S_T(t))}{dt}, \quad t > 0, \tag{3}$$

with $0 < S_T(t) < 1$, where $S_T(t) = \mathbb{P}(T \geq t) = 1 - F_T(t) = \int_t^\infty f_T(u)\,du$, for $t > 0$, is the survival function. In addition, another useful descriptor is the mean residual, which is given by $\mu(x) = \mathbb{E}[T|T > x] = x + [\int_x^\infty S_T(t)\ dt]/S_T(x)$, for $x > 0$ and $S_T(x) > 0$, with $\mu(x) = \mu = \text{E}[T]$, when $x = 0$. For more details about these descriptors, see Johnson, Kotz & Balakrishnan (1995, pp. 640-650), Marshall & Olkin (2007) and Saunders (2007). Note from (3) and below this equation that all of these functions can be expressed by means of the h.r. Therefore, we carry out a hazard analysis based on this rate.

A h.r. function $h_T(t)$ can be increasing, decreasing or constant in t. In particular, if $h_T(t) = \lambda > 0$, for all $t > 0$, then we have that the r.v. T follows an exponential distribution with parameter λ. However, there are distributional families with non-monotone h.r. In this case, an important value for hazard analysis is the change point of the h.r. of T. Within the class of distributions with a non-monotone h.r., we can identify concave or convex hazard rates, i.e., $\cap$-shaped or $\cup$-shaped h.r., respectively. Particularly, for the $\cap$-shaped case, we also have two cases, when the h.r. is initially increasing until its change point and then (i) it decreases to zero, as in the case of the lognormal distribution, or (ii) it decreases until that becomes stabilized in a positive constant, as in the case of the IG and BS distributions. For this reason, when we study distributional families with non-monotone h.r., change point and limit behavior analyses are necessary.

The following theorems provide the MIG-t h.r., its change point, denoted by t_c, and its limiting behavior.

Theorem 4. *Let $T \sim \text{MIG-}t(\mu, \lambda, p, \nu)$. Then, the h.r. of T is given by*

$$h_T(t) = \frac{\sqrt{\lambda}\phi_t(a_t)}{\sqrt{t^3}\left[\Phi_t(-a_t)+(2p-1)J(b_t)\right]}\left[1-p+\frac{p\,t}{\mu}\right], \quad t>0$$

where $J(b_t) = \int_{b_t}^{\infty} \phi_t(\sqrt{u^2 - 4\lambda/\mu})\,du$.

Theorem 5. *Let $T \sim \text{MIG-}t(\mu, \lambda, p, \nu)$. Then, the change point of the h.r. of T is obtained as the solution to*

$$\Phi_t(-a_{t_c}) + [2p-1]J(b_{t_c}) = \frac{\frac{\sqrt{\lambda}}{\sqrt{t^3}}\,\phi_t(a_{t_c})\left[1-p+\frac{p\,t_c}{\mu}\right]}{\frac{\nu+1}{2[\nu+a_{t_c}^2]}\,\frac{\lambda[t_c^2-\mu^2]}{\mu^2 t_c^2}+\frac{3\mu[1-p]+p\,t_c}{2t_c[\mu(1-p)+p\,t_c]}}.$$

Theorem 6. *Let $h_T(t)$ be the h.r. of $T \sim \text{MIG-}t(\mu, \lambda, p, \nu)$, with ν known. Then, $\lim\limits_{t\to\infty} h_T(t) = 0$.*

Order statistics are useful in several statistical procedures. Thus, if $T_1, \ldots, T_n$ are i.i.d. variates, associated order statistics are denoted by $T_{(1)}, \ldots, T_{(j)}, \ldots, T_{(n)}$, where $T_{(1)}$, $T_{(n)}$ and $T_{(j)}$ denote the minimum, maximum and jth order statistic of the variates $T_1, \ldots, T_n$, respectively. For more details about order statistics, see Arnold, Balakrishnan & Nagaraja (1992).

The following theorem provides the p.d.f. of order statistics for the MIG-t distribution.

Theorem 7. *Let $T_1, \ldots, T_n$ be i.i.d. variates, where $T_i \sim \text{MIG-}t(\mu, \lambda, p, \nu)$, for $i = 1, \ldots, n$. Then, for the indicated order statistic, its p.d.f. is given by*

(i) $f_{T_{(1)}}(t) = n\,\phi_t(a_t)\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(-a_t)+(2p-1)\,J(b_t)\right]^{n-1}$, $t>0$

(ii) $f_{T_{(n)}}(t) = n\,\phi_t(a_t)\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(a_t)+(1-2p)\,J(b_t)\right]^{n-1}$, $t>0$

(iii) $f_{T_{(j)}}(t) = \frac{n!\,\phi_t(a_t)}{(j-1)!(n-j)!}\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(a_t)+(1-2p)J(b_t)\right]^{j-1}$
$\times\left[\Phi_t(-a_t)+(2p-1)J(b_t)\right]^{n-j}, t>0$

3. Inference and Diagnostics in the MIG-t Model

In this section, we present estimation, inference and diagnostics useful to estimate the mean protein production and detect the potential influence of atypical data. In problems with this type of data, generally one has enough amount of them to apply asymptotic results. Inference in small samples is not direct, which presents a challenge for a further study.

3.1. ML Estimation, Information Matrix and Inference

Before we find the ML estimators of the MIG-t model parameters, to discuss how one should handle the parameter ν of this model is important. The question is whether ν should be estimated. Several authors treated this topic for the t distribution and models associated with it. See Lange et al. (1989), Lucas (1997), Leiva, Riquelme, Balakrishnan & Sanhueza (2008) and references therein. These authors noticed problems of unbounded and local maximum in the likelihood function, in addition to lack of robustness, when ν is estimated. For this reason, to fix ν is better and assume that it is a known value or, otherwise, acquire information for it from the data. Thus, once the optimum value of ν is found, the parameters μ, λ and p of the MIG-t distribution are estimated as described next.

3.1.1. ML Estimation

The log-likelihood function for $\boldsymbol{\theta} = (\mu,\lambda,p)^\top$, based on a random sample $T_1,\dots,T_n$, where $T_i \sim$ MIG-$t(\mu,\lambda,p,\nu)$, for $i=1,\dots,n$, is expressed as $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$, where

$$\ell_i(\boldsymbol{\theta}) \propto \frac{n}{2}\log(\lambda) - \frac{[\nu+1]}{2}\log\left(\nu+a_{t_i}^2\right) + \log\left(\mu[1-p]+p\,t_i\right) - \log(\mu) \tag{4}$$

The score vector of first derivatives of the log-likelihood function is given by

$$\dot{\ell}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\dot{\ell}_{\theta_1}), \quad \text{with } \theta_1 = \mu, \lambda, \text{ or } p \tag{5}$$

where

$$\begin{aligned}\dot{\ell}_\mu &= \frac{\lambda[\nu+1]}{\mu^3}\sum_{i=1}^n\left\{\frac{t_i-\mu}{\nu+a_{t_i}^2}\right\} + \sum_{i=1}^n\left\{\frac{1-p}{\mu[1-p]+pt_i}\right\} - \frac{n}{\mu},\\ \dot{\ell}_\lambda &= \frac{n}{2\lambda} - \frac{[\nu+1]}{2\lambda}\sum_{i=1}^n\left\{\frac{a_{t_i}^2}{\nu+a_{t_i}^2}\right\} \quad \text{and} \quad \dot{\ell}_p = \sum_{i=1}^n\left\{\frac{t_i-\mu}{\mu[1-p]+p\,t_i}\right\}\end{aligned}$$

The ML estimates of the parameters μ, λ and p are solutions to the equations $\dot{\ell}_\mu = 0$, $\dot{\ell}_\lambda = 0$ and $\dot{\ell}_p = 0$. However, these equations do not provide analytical solutions, so that an iterative numerical method is necessary to find the roots. As starting values for this iterative method, we propose considering the ML estimates of μ, λ and p of the MIG distribution. See Seshadri (1999, pp. 145).

To select the value of ν, we propose looking for the value that maximizes the likelihood function for $\nu \in [1, 100]$ using an optimal search of ν by means of the following algorithm:

(A1) For $\nu = 1$ to $\nu = 100$ by 1:

- **(A1.1)** Estimate the parameters μ, λ and p of the MIG-t model considering the ML estimates of μ, λ and p of the MIG distribution starting values for the numerical iterative procedure;
- **(A1.2)** Compute the corresponding likelihood function;

(A2) Choose the value of ν that maximizes this likelihood function and then consider the ML estimates of μ, λ and p the result.

***Note* 2.** Based on the invariance property of the ML estimators, we can estimate different functions of the parameter $\boldsymbol{\theta}$. For example, the mean protein production can be estimated by using the mean of MIG-t distribution given in Theorem 3 (i).

3.1.2. Information Matrix

The observed information matrix is obtained as $\mathcal{J}(\boldsymbol{\theta}) = -\ddot{\ell}$. Here, $\ddot{\ell}$ is the Hessian matrix of second derivatives of the log-likelihood function given by

$$\ddot{\ell}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} = (\ddot{\ell}_{\theta_1\, \theta_2}), \quad \text{with } \theta_1, \theta_2 = \mu, \lambda, \text{ or } p \tag{6}$$

where

$$\begin{aligned}
\ddot{\ell}_{\mu\mu} &= -\frac{3\lambda[\nu+1]}{\mu^4} \sum_{i=1}^{n} \left\{ \frac{t_i - \mu}{\nu + a_{t_i}^2} \right\} - \frac{\lambda[\nu+1]}{\mu^3} \sum_{i=1}^{n} \left\{ \frac{1}{\nu + a_{t_i}^2} \right\} \\
&\quad - \frac{2\sqrt{\lambda^3}(\nu+1)}{\mu^5} \sum_{i=1}^{n} \left\{ \frac{[t_i - \mu] a_{t_i} \sqrt{t_i}}{[\nu + a_{t_i}^2]^2} \right\} - \sum_{i=1}^{n} \left\{ \frac{1-p}{\mu[1-p] + p\, t_i} \right\}^2 \\
\ddot{\ell}_{\mu\lambda} &= \frac{[\nu+1]}{\mu^3} \sum_{i=1}^{n} \left\{ \frac{t_i - \mu}{\nu + a_{t_i}^2} \right\} - \frac{[\nu+1]}{\mu^3} \sum_{i=1}^{n} \left\{ \frac{[t_i - \mu] a_{t_i}^2}{[\nu + a_{t_i}^2]^2} \right\}, \ \ddot{\ell}_{\lambda p} = 0 \\
\ddot{\ell}_{\mu p} &= -\sum_{i=1}^{n} \left\{ \frac{t_i}{[\mu[1-p] + p\, t_i]^2} \right\}, \ddot{\ell}_{pp} = -\sum_{i=1}^{n} \left\{ \frac{t_i - \mu}{\mu[1-p] + p\, t_i} \right\}^2 \\
\ddot{\ell}_{\lambda\lambda} &= -\frac{n}{2\lambda^2} - \frac{\nu[\nu+1]}{2\lambda^2} \sum_{i=1}^{n} \left\{ \frac{a_{t_i}^2}{[\nu + a_{t_i}^2]^2} \right\} + \frac{[\nu+1]}{2\lambda^2} \sum_{i=1}^{n} \left\{ \frac{a_{t_i}^2}{\nu + a_{t_i}^2} \right\}
\end{aligned}$$

3.1.3. Inference

Inference for $\boldsymbol{\theta}$ can be based on the asymptotic behavior of the ML estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\lambda}, \widehat{p})^\top$ given by $\sqrt{n}\,[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \xrightarrow{\text{d}} \text{N}_3(\mathbf{0}, \boldsymbol{\Sigma}_{\widehat{\theta}})$, where "$\xrightarrow{\text{d}}$" means convergence in distribution. Here, $\widehat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_{\widehat{\theta}}$ is the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$, which can be obtained from the expected information matrix, namely $\mathcal{I}(\boldsymbol{\theta}) = \text{E}[\mathcal{J}(\boldsymbol{\theta})] = -\text{E}[\,\ddot{\ell}\,]$, as $\boldsymbol{\Sigma}_{\widehat{\theta}} = \mathcal{I}(\boldsymbol{\theta})^{-1}$. Thus, the standard errors of the ML estimators can be computed by using the square roots of the diagonal elements of $\mathcal{I}(\boldsymbol{\theta})^{-1}$. Their estimated standard errors can be obtained by evaluating $\boldsymbol{\theta}$ at its ML estimate $\widehat{\boldsymbol{\theta}}$.

***Note* 3.** Instead of the expected information matrix, its observed version could be used to approximate the standard errors of the ML estimators. These errors can be computed by using the square roots of the diagonal elements of $\mathcal{J}^{-1}(\boldsymbol{\theta})$. Once again, their estimated standard errors can be obtained by evaluating $\boldsymbol{\theta}$ at its ML estimate $\widehat{\boldsymbol{\theta}}$. For more details about the use of the observed information matrix instead of its expected value, see Efron & Hinkley (1978).

A confidence region for $\boldsymbol{\theta}$ may be constructed by using the asymptotic normal distribution of $\widehat{\boldsymbol{\theta}}$ above mentioned. Thus, an approximate $(1-\alpha)100\%$ confidence region for $\boldsymbol{\theta}$, with $0 < \alpha < 1$, is given by $\mathcal{R} = \{\boldsymbol{\theta} \in \mathbb{R}^3\text{: } (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}_{\widehat{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_{1-\alpha}(3)\}$, where $\chi^2_{1-\alpha}(3)$ denotes the $(1-\alpha)$th quantile of the χ^2 distribution with three degrees of freedom.

3.2. Influence Diagnostics

Case deletion is a common way to assess the effect of an observation on the estimation procedure. This is a global influence analysis, since the effect of a case is evaluated by dropping it from the data set. Alternatively, local influence is based more on geometric differentiation than the elimination of observations. In this last case, a differential comparison of estimators is used before and after perturbing the data or the model. We implement the local influence method for evaluating possible atypical cases in the protein production data. As in Cook (1986), we use the likelihood displacement to evaluate the local influence. Next, we present global and local influence techniques that may be useful for detecting atypical protein production data and studying the suitability of the MIG-t model to such data.

3.2.1. Global Influence

Cook's distance is an interesting diagnostics technique of the global influence method. See Cook & Weisberg (1982). A generalization of this distance is expressed as $D_i = [(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})]/k$, for $i = 1, \ldots, n$, where k is the number of parameters and $\widehat{\boldsymbol{\Sigma}}_{\widehat{\theta}}$ is an estimator of the covariance matrix of $\widehat{\boldsymbol{\theta}}$, which, as mentioned, can be approximated by $-\ddot{\ell}^{-1}$ evaluated at $\widehat{\boldsymbol{\theta}}$, such that $D_i = [(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top (-\ddot{\ell})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})]/k$. If we use an approximation of first order, we obtain $\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)} \approx [\ddot{\ell}_{(i)}]^{-1} \dot{\ell}_{(i)}$, with $\dot{\ell}_{(i)}$ being the score vector and $\ddot{\ell}_{(i)}$ the Hessian matrix without considering the ith case. Thus, $D_i \approx [(\dot{\ell}_{(i)})^\top(\ddot{\ell}_{(i)})^{-1}(-\ddot{\ell})^{-1}(\ddot{\ell}_{(i)})^{-1}\dot{\ell}_{(i)}]/k$, where a high value for D_i indicates a high impact case on the ML estimator of $\boldsymbol{\theta}$. For the MIG-t model, in D_i, $k = 3$ and $\dot{\ell}_{(i)}$ and $\ddot{\ell}_{(i)}$ are analogously defined as those given in (5) and (6), respectively.

3.2.2. Local Influence

From (4), we can note that the contributions $\ell_i(\boldsymbol{\theta})$ are equally weighted. A perturbed log-likelihood function can be defined by $\ell(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \ell_i(\boldsymbol{\theta})$, with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ being the vector of weights of the contributions from each case to the likelihood function and $\boldsymbol{\omega}_0 = \mathbf{1}_n = (1, \ldots, 1)^\top$ being the non-perturbed point, that is, $\ell(\boldsymbol{\theta} \mid \boldsymbol{\omega}_0) = \ell(\boldsymbol{\theta})$. This scheme of perturbation is useful for evaluating whether the contribution of cases representing to protein production data with different weights influence the ML estimator of $\boldsymbol{\theta}$. Specifically, let $\widehat{\boldsymbol{\theta}}_\omega$ be the ML estimator of $\boldsymbol{\theta}$ obtained from the perturbed likelihood function. The influence of the perturbation $\boldsymbol{\omega}$ on the ML estimator may be checked by means of the likelihood displacement given by $\mathrm{LD}(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\omega)]$. Cook (1986) postulated studying the local behavior of $\mathrm{LD}(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_0$ employing the normal curvature C_l of $\mathrm{LD}(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ and in the direction of some unitary vector $\boldsymbol{l}$. He showed that $C_l = 2 \mid \boldsymbol{l}^\top \boldsymbol{\Delta}^\top \ddot{\ell}^{-1} \boldsymbol{\Delta} \boldsymbol{l} \mid$, with $\|\boldsymbol{l}\| = 1$, where $\ddot{\ell}$ is as defined in (6) and $\boldsymbol{\Delta}$ is a $3 \times n$ perturbation matrix expressed as $\boldsymbol{\Delta} = [\boldsymbol{\Delta}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\Delta}_n(\boldsymbol{\theta})]$, both evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$. For the MIG-t distribution, the elements of $\boldsymbol{\Delta}$ are

$$\boldsymbol{\Delta}_i(\boldsymbol{\theta}) = (\Delta_i(\mu), \Delta_i(\lambda), \Delta_i(p))^\top = \left(\frac{\partial^2 \ell(\mu \mid \boldsymbol{\omega})}{\partial\mu\partial\omega_i}, \frac{\partial^2 \ell(\lambda \mid \boldsymbol{\omega})}{\partial\lambda\partial\omega_i}, \frac{\partial^2 \ell(p \mid \boldsymbol{\omega})}{\partial p\, \partial\omega_i}\right)^\top$$

for $i = 1, \ldots, n$, where

$$\begin{aligned}
\Delta_i(\mu) &= \frac{\lambda\,[\nu+1]\,[t_i - \mu]}{\mu^3\,[\nu + a_{t_i}^2]} + \frac{1-p}{\mu[1-p] + p\,t_i} - \frac{1}{\mu} \\
\Delta_i(\lambda) &= \frac{1}{2\lambda} - \frac{[\nu+1]\,a_{t_i}^2}{2\lambda\,[\nu + a_{t_i}^2]} \quad \text{and} \quad \Delta_i(p) = \frac{t_i - \mu}{\mu[1-p] + p\,t_i}
\end{aligned}$$

Let $\boldsymbol{l}_{\max}$ be the direction of the maximum normal curvature, which corresponds to the perturbation that reaches the greatest local change in $\widehat{\boldsymbol{\theta}}$. The most influential cases in the protein production data may be identified by their large components of the vector $\boldsymbol{l}_{\max}$. In addition, $\boldsymbol{l}_{\max}$ is the eigenvector associated with the largest eigenvalue of $\boldsymbol{B} = \boldsymbol{\Delta}^\top \ddot{\ell}^{-1} \boldsymbol{\Delta}$. Another interesting direction is $\boldsymbol{l} = \boldsymbol{e}_{in}$, which is the ith unitary vector of $\mathbb{R}^n$. In this case, the normal curvature is given by $C_i = 2|b_{ii}|$, with b_{ii} being the ith diagonal element of $\boldsymbol{B}$. Thus, C_i can be useful to detect the total local influence of the ith case of protein production using as benchmark $C_i > 2\,\overline{C}$, where $\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, for indicating whether such a case is potentially influential.

4. Application to Real Data

In this section, for illustrative purposes, we apply some of the obtained results for the MIG-*t* distribution to novel data corresponding to the production of a secreted protein by *Lactococcus lactis*, where initially just one data set is analyzed as follows. First, an implementation in R code of the MIG-*t* model is discussed. Second, the problem upon analysis is described. Third, the data set is provided. Fourth, an exploratory data analysis (EDA) of this set is produced. Fifth, the parameters of the MIG-*t* distribution are estimated by using the ML method. Later, we carry out a brief diagnostic analysis in order to establish the potential influence of some protein production data. Then, goodness-of-fit is presented for studying the suitability of the MIG-*t* distribution to such data. Finally, we compare the production between two different constructions, one of them corresponding to the analyzed data set. The constructions are bacterial strains genetically engineered to produce and secret a protein of interest. In this concluding analysis, by using the invariance property of the ML estimators of the parameters of the MIG-*t* distribution, we estimate the mean of two populations (constructions) and conducted a statistical comparison between these mean values by using the LR test.

4.1. Implementation in R Code

R language is an open-source software package for statistical computing and graphics that can be obtained from `http://www.R-project.org`; see R Development Core Team (2009). Several R packages for analyzing data from different distributions are available and can be downloaded from `http://CRAN.R-project.org`. We have developed an R code to analyze data from the MIG-*t* model. This code contains diverse indicators of the MIG-*t* distribution and allows us to compute ML estimates of its parameters.

4.2. The Problem upon Analysis

Lactococcus lactis is a lactic acid microorganism corresponding to a well-characterized gram-positive bacterium that can be used for food industry. This microorganism can be genetically modified to allow the production of proteins and secretion of proteins into the culture media. These proteins can be purified and used for several purposes in food industry. Depending on the success of the genetic construction, the yield protein will vary among constructs. For this reason, once genetic constructions are finished, *Lactococcus lactis* is reproduced by experiments *in vitro* at a laboratory before produce it at an industrial scale. At this stage, protein production is measured to study its feasibility and stability, and compare production levels among different constructs. Once the production of proteins from this bacterium has reached a level with a small variation among essays, i.e., it has been stabilized, then such proteins can be produced at big scale in a fermenter where cultures of several liters are produced. *Lactococcus lactis* may yield many types of proteins (see Simoes-Barbosa et al. 2004), although this bacterium does not secret an important amount of proteins. Therefore, genetical constructs allowing the production and secretion of proteins of industrial interest in this bacterium have been a major research point. All secreted proteins carry a signal peptide that directs them to the extracellular culture media. Best results have been obtained when a native peptide (belonging to *Lactococcus lactis*) is used even when the produced protein does not belong to this bacterium. See Le-Loir et al. (2001). Specifically, the authors postulated a model for protein secretion based on *Lactococcus lactis* using the staphylococcal nuclease (NucB), a non native protein, and replacing the signal peptide by a native signal peptide (from USp45). The protein production is higher using the *Lactococcus lactis* signal peptides than other peptides, particularly for Usp45 with classical tests for which protein production data do not meet the assumptions. The question that arises here is whether this level of production is transferable to other signal peptides from secreted proteins as YvjB and how this could be addressed with an adequate distribution and an appropriate test for the data. In the application that we make in this study, we analyze data on protein production from secreted NucB possessing the YvjB signal peptide (called "Group 2") and the native signal peptide (called "Group 1"). After analyzing the first data set by the MIG-t distribution, we statistically compare both groups by using this distribution, which may be useful for modeling this kind of protein production data as an alternative procedure to the traditional ones. As mentioned, this fact is very important to determinate the most productive construction before the bacterial culture is scaled to an industrial level.

4.3. The Data Set

As mentioned, the data set corresponds to protein production data (expressed in ng/ml) from *Lactococcus lactis*, which are: 165, 123, 123, 128, 129, 135, 156, 165, 169, 178, 178, 198, 206, 207, 208, 213, 115, 119, 225, 236, 236, 156, 287, 189, 295, 296, 302, 324, 356, 389, and that we simply call `lactis`.

4.4. Exploratory Data Analysis

Table 1 presents a descriptive summary of `lactis`, while Figure 3 (left) shows the histogram and boxplot of these data. An EDA of `lactis` based on Table 1 and Figure 3 shows a positively skewed distribution with an atypical data. We propose the MIG-t model for describing these data.

Table 1: Descriptive statistics for lactis (in ng/ml)

Median	Mean	SD	CV	CS	CK	Range	Min.	Max.	n
193.5	206.867	74.796	36.156%	0.741	2.542	274	115	389	30

4.5. Estimation and Model Checking

To estimate the parameters μ, λ, p and ν of the MIG-t distribution, we use the ML method described in Subsection 3.1.1. As mentioned there, we suggest to fix ν

and assume that it is a known value or, otherwise, get information for it from the data. Thus, to estimate μ, λ and p of the MIG-t_ν model, we fix integer values for ν from 1 to 100 by 1, choosing the value of ν that maximizes the likelihood function. The command `mleMIGt()` has been implemented in the software R to carry out the procedure described in Subsection 3.1.1. The instruction `mleMIGt(lactis)` automatically chooses the value of ν that maximizes the likelihood function and computes the ML estimates of μ, λ and p of the MIG-t model according to (A1)-(A2). The function `optim` is used to solve the corresponding iterative numerical procedure, which is available in the software R.

***Note* 4.** The function `optim` employs the `L-BFGS-B` method developed by Byrd, Lu, Nocedal & Zhu (1995) to carry out the corresponding numerical optimization. This method allows having a "box constraint" and so each variable of the optimization procedure can have a lower or upper bound. The `L-BFGS-B` method uses a limited-memory modification of the quasi-Newton method.

Based on `lactis`, the obtained results for these estimates are $\widehat{\mu} = 204.109$, $\widehat{\lambda} = 1692.646$ and $\widehat{p} = 0.090$, with $-\ell(\widehat{\boldsymbol{\theta}}) = 168.523$ being the negative value of the log-likelihood function evaluated at these estimates.

Next, we detect the effect of potentially influential observations on the ML estimates for `lactis`. These observations are chosen by using the local influence method described in Subsection 3.2.2 by means of the total local influence index plot (Figure 2). From this figure (left), we can note a potential influence of the cases #29 and #30 on the ML estimates of the classical MIG distribution. However, as expected, this potential influence is less pronounced for the MIG-t distribution (see Figure 2, right).

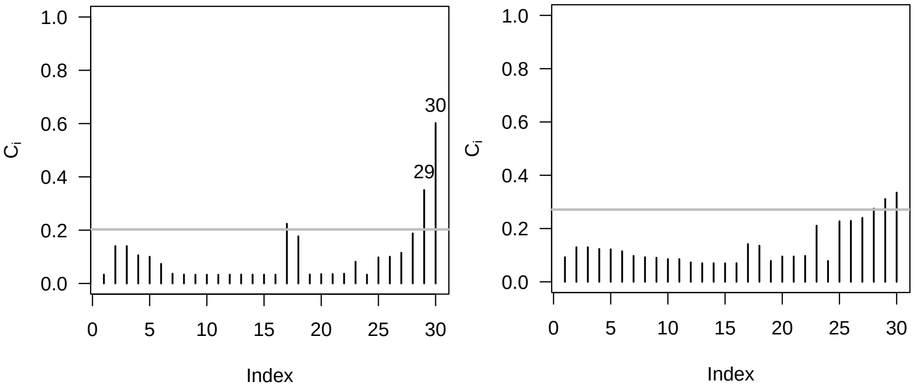

Figure 2: Total local influence index plot for the MIG (left) and MIG-t (right) models.

***Note* 5.** Since the purpose of this article is to illustrate the use of the MIG-t model in the context of protein production data and not to conduct an influence analysis, the information provided by the model checking is sufficient for us. In future studies, a more deeper analysis should be carried out on these atypical cases. Also, comparison of the obtained results in this application with other distributions usually employed in hazard analysis, as well as the analysis of lifetime data by this model, will be addressed in future studies.

Once the MIG-t model parameters have been estimated and the influence analysis conducted, a natural question that arises is how good the fit of the model to `lactis` is. For this purpose, we can calculate the Kolmogorov-Smirnov (KS) distance between the empirical c.d.f. $F_n(\cdot)$, and the MIG-t c.d.f., $F_T(\cdot)$, given by

$$\text{KSD}_i = |F_n(t) - F_T(t)|, \quad i = 1, \dots, n$$

To compute this distance, we replace the parameters in the MIG-t c.d.f. by their respective ML estimates. Once all the n KS distances are calculated, we determine the maximum value of such distances and then compare it to the $(1-\alpha)$th quantile of the KS distribution to evaluate the suitability of the MIG-t model to `lactis`. The p-value of the KS test is 0.910, which strongly supports the hypothesis that the MIG-t distribution fit `lactis` in a very good way. To visually verify this fact, we use the invariance property of the ML estimators for determining the MIG-t p.d.f. and c.d.f., which are shown in Figure 3 on the histogram and empirical c.d.f. of the data, respectively. These graphs show the excellent agreement between the MIG-t model and `lactis`. Other goodness-of-fit tests, such as the Anderson-Darling test or those for normality as the Lillieford and Shapiro-Wilk test adapted to `lactis` could be also applied, but we consider the information provided by the KS test concluding.

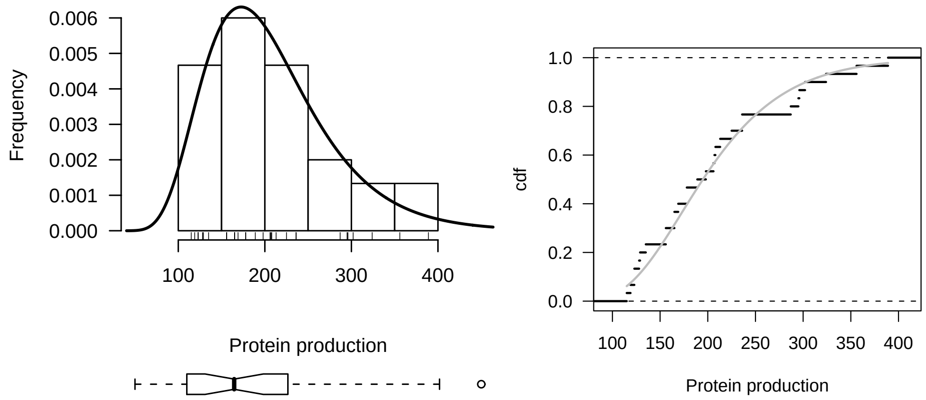

Figure 3: Histogram and boxplot with estimated MIG-t p.d.f. (left) and empirical c.d.f. versus estimated MIG-t c.d.f. (right) for lactis.

4.6. A Comparative Analysis

Once the MIG-t distribution has been chosen for fitting `lactis`, we can estimate the mean of this distribution. To do this, we use the ML estimates of μ, λ and p, the optimum value for ν, the invariance property of these estimators and Theorem 3(i). We estimate the mean $\text{E}[T] = \mu + [p\,\mu^2\,\nu]/[\lambda\,(\nu-2)]$ to detect the protein production based on the MIG-t distribution using `lactis`, which is $\widehat{\mathrm{E}[T]} = 206.370$ ng/ml.

As mentioned, for practitioners, to compare two constructs is important. Let us to denote these constructs as distributions F and G. On the basis of two independent samples $T_{i1}, \ldots, T_{in_i}$, for $i = 1, 2$, each one randomly extracted from its respective population, we assume that $T_{ij} \sim$ MIG-$t(\mu_i, \lambda, p, \nu)$, for $i = 1, 2$. We want to test H_0: $\mathrm{E}[T_1] = \mathrm{E}[T_2]$ against H_1: $\mathrm{E}[T_1] \neq \mathrm{E}[T_2]$, where $\mathrm{E}[T_i]$ is the mean of the ith population. For testing H_0 against H_1, we use the LR test, whose statistic is given by

$$\mathrm{LR} = \prod_{j=1}^{n_1} \left[\frac{1 + \frac{1}{\nu} a^2_{t_{1j}}(\widehat{\mu}, \widehat{\lambda})}{1 + \frac{1}{\nu} a^2_{t_{1j}}(\widehat{\mu_1}, \widehat{\lambda})} \right]^{-\frac{\nu+1}{2}} \prod_{j=1}^{n_2} \left[\frac{1 + \frac{1}{\nu} a^2_{t_{2j}}(\widehat{\mu}, \widehat{\lambda})}{1 + \frac{1}{\nu} a^2_{t_{2j}}(\widehat{\mu_1}, \widehat{\lambda})} \right]^{-\frac{\nu+1}{2}} \tag{7}$$

By using the LR statistic defined in (7), we compare the protein production from *Lactococcus lactis* based on the MIG-t distribution for two groups: NucB (Group 1) and PSYvjB (Group 2), which estimated mean values are $\widehat{\mathrm{E}}[T_1] = 206.370$ ng/ml and $\widehat{\mathrm{E}}[T_2] = 262.167$ ng/ml. The p-value for the LR test is < 0.001, which provides enough evidence for rejecting the hypothesis of equality of means, so that PSYvjB statistically produces a greater amount of proteins than NucB. Therefore, we recommend it as microorganism for producing proteins at big scale in the dairy food industry. The found results agree with those obtained in previous studies, where the native peptide allows providing higher amounts of protein.

5. Concluding Remarks

In this article, we have derived a mixture inverse Gaussian model based on the Student-t distribution and applied it to bacterium-based protein production for food industry. This model is very flexible in kurtosis and skewness, and has a kurtosis levels greater than that of its usual version. The mixture inverse Gaussian-t model is mainly useful to describe data that follow positively skewed distributions and accommodate atypical observations in a better way than its usual version. We have provided several statistical, hazard, probabilistic and computational aspects of the mixture inverse Gaussian-t distribution. Specifically, for this distribution, we have carried out a hazard analysis based on the hazard rate, discussed maximum likelihood estimation and evaluated the potential influence of atypical observations by a diagnostic analysis. Thus, we have introduced a statistical distribution that can be useful for modeling different types of data and, particularly, those of protein production from a lactic acid bacterium called *Lactococcus lactis*, which is a microorganism used primarily in dairy food industry. In problems of bacterium-based protein production, generally one has enough amount of data to apply asymptotic results. Inference in small samples for the mixture inverse Gaussian-t model is not direct so that this presents a challenge for a future study. We have applied the obtained results to novel bacterium-based protein production data and statistically compared two types of protein producers using the proposed distribution by the likelihood ratio test as an alternative methodology to the procedures available so far. This application showed the utility of the mixture inverse Gaussian-t distribution.

Acknowledgements

The authors wish to thank the editors Leonardo Trujillo, Ph.D. and Piedad Urdinola, Ph.D., and anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. This study was partially supported by FONDECYT 1090265 grant from Chile.

[Recibido: octubre de 2010 — Aceptado: marzo de 2011]

References

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (1992), *A First Course in Order Statistics*, John Wiley and Sons, New York.

Balakrishnan, N., Leiva, V., Sanhueza, A. & Cabrera, E. (2009), 'Mixture inverse Gaussian distribution and its transformations, moments and applications', *Statistics* **43**, 91–104.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Chhikara, R. S. & Folks, J. L. (1989), *The Inverse Gaussian Distribution*, Marcel Dekker, New York.

Cook, R. D. (1986), 'Assessment of local influence (with discussion)', *Journal of The Royal Statistical Society Series B–Statistical Methodology* **48**, 133–169.

Cook, R. D. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman & Hall, London.

Efron, B. & Hinkley, D. (1978), 'Assessing the accuracy of the maximum likelihood estimator: Observed vs. expected Fisher information', *Biometrika* **65**, 57–487.

Folks, J. L. (2007), Inverse Gaussian distribution, *in* S. Kotz, C. B. Read, N. Balakrishnan & B. Vidakovic, eds, 'The Encyclopedia of Statistical Sciences', Vol. 6, John Wiley & Sons, New York, pp. 3681–3682.

Gupta, R. C. & Akman, H. O. (1995), 'On the reliability studies of the weighted inverse Gaussian model', *Journal of Statistical Planning and Inference* **48**, 69–83.

Gupta, R. C. & Kirmani, S. (1990), 'The role of weighted distributions in stochastic modeling', *Communications in Statistics: Theory and Methods* **19**, 3147–3162.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 1, John Wiley and Sons, New York.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, John Wiley and Sons, New York.

Jorgensen, B. (1982), *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Springer, Heidelberg.

Jorgensen, B., Seshadri, V. & Whitmore, G. (1991), 'On the mixture of the inverse Gaussian distribution with its complementary reciprocal', *Scandinavian Journal of Statistics* **18**, 77–89.

Kotz, S., Leiva, V. & Sanhueza, A. (2010), 'Two new mixture models related to the inverse Gaussian distribution', *Methodology and Computing in Applied Probability* **12**, 199–212.

Lange, K. L., Little, J. A. & Taylor, M. G. J. (1989), 'Robust statistical modeling using the t distribution', *Journal of the American Statistical Association* **84**, 881–896.

Le-Loir, Y., Nouaille, S., Commissaire, J., Bretigny, L., Gruss, A. & Langella, P. (2001), 'Signal peptide and propeptide optimization for heterologous protein secretion in lactococcus lactis', *Applied and Environmental Microbiology* **67**, 4119–2127.

Leiva, V., Riquelme, M., Balakrishnan, N. & Sanhueza, A. (2008), 'Lifetime analysis based on the generalized Birnbaum-Saunders distribution', *Computational Statistics and Data Analysis* **21**, 2079–2097.

Leiva, V., Sanhueza, A. & Angulo, J. M. (2009), 'A length-biased version of the Birnbaum-Saunders distribution with application in water quality', *Stochastic Environmental Research and Risk Assessment* **23**, 299–307.

Leiva, V., Sanhueza, A., Sen, P. K. & Araneda, N. (2010), 'M-procedures in the general multivariate nonlinear regression model', *Pakistan Journal of Statistics* **26**, 1–13.

Lucas, A. (1997), 'Robustness of the student t based m-estimator', *Communications in Statistics: Theory and Methods* **26**, 1165–1182.

Marshall, A. W. & Olkin, I. (2007), *Life Distributions*, Springer Verlag, New York.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture Models*, John Wiley and Sons, New York.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, third edn, John Wiley and Sons, New York.

Patil, G. P. (2002), Weighted distributions, *in* A. H. El-Shaarawi & W. W. Piegorsch, eds, 'Encyclopedia of Environmetrics', Vol. 4, John Wiley & Sons, Chichester, pp. 2369–2377.

R Development Core Team (2009), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Sanhueza, A., Sen, P. K. & Leiva, V. (2009), 'A robust procedure in nonlinear models for repeated measurements', *Communications in Statistics: Theory and Methods* **38**, 138–155.

Saunders, S. C. (2007), *Reliability, Life Testing and Prediction of Services Lives*, Springer, New York.

Schrodinger, E. (1915), 'Zur theorie der fall-und steigversucheand teilchen mit brownscher bewegung', *Physikalische Zeitschrift* **16**, 289–95.

Seshadri, V. (1993), *The Inverse Gaussian Distribution: A Case Study in Exponential Families*, Clarendon Press, New York.

Seshadri, V. (1999), *The Inverse Gaussian Distribution: Statistical Theory and Applications*, Springer, New York.

Simoes-Barbosa, A., Abreu, H., Silva-Neto, A., Gruss, A. & Langella, P. (2004), 'A food-grade delivery system for lactococcus lactis and evaluation of inducible gene expression', *Applied Microbiology and Biotechnology* **65**, 61–67.

Tweedie, M. C. K. (1957), 'Statistical properties of the inverse Gaussian distribution - I', *Annals of Mathematics Statistical* **28**, 362–377.

Wald, A. (1947), *Sequential Analysis*, John Wiley and Sons, New York.