On the Characterization and Estimation of Glog-Normal Scale Mixture Distributions and Their Application in Genetics

Filidor Vilca[a], Mariana Rodrigues-Motta[a] and Víctor Leiva[b]

[a] *Departamento de Estatística, Universidade Estadual de Campinas, São Paulo, Brazil,* [b] *Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Valparaíso, Chile*

Abstract

Microarray data studies allow to generate expression of thousands of genes yielding valuable information for biologists. However, this kind of data does usually not agree with the assumption of constant variance in which classic statistical methodology rely. In order to avoid the transformation of data that conducts to the stabilization of variance and provide a robust methodology in presence of atypical data, we propose a model based on a class of symmetric distributions that solves these difficulties. Specifically, we describe structural aspects of this new distribution considering its density, distribution and quantile functions, properties, moments and parameter estimation. Finally, the usefulness of the proposed distribution for modeling gene expression data is shown by means of a real numerical example.

Key words: Gene expression; Johnson's system distributions; MA plots; Microarrays; Non-normality; Transformations.

1 Introduction

An important part of the classic statistical methodology relies on assumptions of normally and constant variance. When these assumptions are violated, a logarithmic transformation of the data is usually employed for fulfilling such assumptions and so utilizing a traditional methodology.

Microarray data studies allow to simultaneously measure the expression of thousands of genes yielding valuable information for biologists. However, this kind of data does usually not agree with the above mentioned assumptions; see Rocke and Durbin (2003). Since the logarithmic transformation of gene expression data does not always stabilize the variance of such observations

in presence of non-positive values, a transformation that can be used for this purpose is given by

$$\operatorname{glog}(x) = \operatorname{arcsinh}(x) = \log\left(x + \sqrt{x^2+1}\right), x \in \mathbb{R}. \tag{1}$$

This transformation is known as the generalized logarithm (glog); for more details about how the glog transformation stabilizes the variance of gene expression data, see Durbin et al. (2002).

Although a practitioner could choose to use the glog transformation for gene expression data in order to use a classic methodology, some problems such as reduction the power of the study and difficulties of interpretation could be presented; see Huang and Qu (2006). An alternative way that one could choose is developing a new statistical methodology useful for modeling this kind of data, which should be mathematically treatable and available for users.

Johnson (1949) used the translation method to generate statistical distributions covering a wide variety of shapes by the random variable (r.v.)

$$Z = \gamma + \delta f\left(\tfrac{Y-\xi}{\lambda}\right), \tag{2}$$

where $Z \sim \mathrm{N}(0,1)$ and $f(\cdot)$ is a simple monotone function of Y. These distributions are known as Johnson's system and have four parameters, such as is the case of the well-know Pearson's system. These parameters are denoted by γ and δ (shape), ξ (location), and λ (scale). Without loss of generality, $f(\cdot)$ can be assumed as a non-decreasing function and δ and λ as positive values. Based on Eq. (2), note that (i) if $f(x) = \log(x)$, $x > 0$, we have Johnson's S_{L} model and (ii) if $f(x) = \operatorname{glog}(x)$, $x \in \mathbb{R}$, we have Johnson's S_{U} model. Observe that the r.v. given in (i) is related to the well-known normal distribution. On the other hand, the r.v. given in (ii) is related to the glog-normal (GLN) distribution; see Leiva et el. (2009). In this way, by the use of these two functions, we are confronting the classic methodology with a new one. Thus, Johnson's system of distributions can provide a wide avenue for characterizing complicated data sets such as is the case of microarray data; see George (2007).

The normal scale mixture models are a family of symmetric distributions that admits an interesting stochastic representation, which conducts to attractive properties, such as the robust parameter estimation, easy number generation, and efficient computation of the ML estimates via the EM-algorithm. (Of course the normal distribution is a particular case of such a family.) For more details about this class of models, see Gneiting (1997). Specifically, the normal scale mixture models are related to the normal distribution through the stochastic representation

$$Z = \mu + U^{-1/2} Z_0, \tag{3}$$

where $Z_0 \sim \mathrm{N}(0, \sigma^2)$, U is a positive r.v. independent of Z_0 with cumulative distribution function (cdf) $H(\cdot)$ indexed by a scalar or vector parameter $\boldsymbol{\nu}$. The multivariate version of (3) is the family of normal/independent (NI) distributions discussed by Lange and Sinheimer (1993). Note, when H is degenerate, with $U = 1$, we obtain the normal distribution. We base the present study on these class of normal scale mixture distributions. An r.v. Z has a NI distribution with location and scale parameters, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, respectively, iff its probability density function (pdf) is of the form

$$\phi_{\mathrm{NI}}(z) = \int_0^\infty \phi\big(z; \mu, \sigma^2/u\big)\, dH(u), \tag{4}$$

where $\phi(\cdot\,; \mu, \sigma^2/u)$ is the pdf of the normal distribution with mean μ and variance σ^2/u and $H(u)$ is the cdf of U introduced in (3). For an r.v. Z with pdf as given in (4), the notation $Z \sim \mathrm{NI}(\mu, \sigma^2; H)$ is used. Now, when $\mu = 0$ and $\sigma^2 = 1$, we use the simpler notation $Z \sim \mathrm{NI}(H)$. When the EM algorithm is used in the ML estimation of parameters of the NI distribution, we obtain similar expressions to the normal case and so the procedure here proposed generalizes that developed for the normal distribution.

The aims of the article are (i) to introduce and characterize a distribution based on the generalized logarithm and the normal/independent model, which we call denoted by GLNI, (ii) to provide a robust parameter estimation procedure based on the GLNI distribution for analyzing gene expression data, and (iii) to show the utility of such a distribution in the modeling of this kind of genetics data.

We organize this study as follows. In Section 2, we introduce and characterize the GLNI distribution and find several of its structural aspects such as its pdf, cdf, quantile function (qf), properties and moments. In Section 3, we estimate parameters of this distribution by using the maximum likelihood (ML) method via an EM-type algorithm. In Section 4, we show the usefulness of the new distribution in the modeling of gene expression data by means of a real numerical example. Finally, in Section 5, we draw some conclusions.

2 Characterization of the model

Based on (1), (2) and (3), an r.v. Y follows a GLNI distribution with shape parameters $\gamma \in \mathbb{R}$ and $\delta > 0$, location parameter $\xi \in \mathbb{R}$, and scale parameter $\lambda > 0$, which is denoted by $Y \sim \mathrm{GLNI}(\gamma, \delta, \xi, \lambda; H)$, if this r.v. can be stochastically represented by

$$Y = \xi + \lambda \sinh\left(\frac{Z - \gamma}{\delta}\right) = \xi + \lambda \sinh\left(\frac{Z_0 - \sqrt{U}\,\gamma}{\sqrt{U}\,\delta}\right),$$

where $Z = U^{-1/2} Z_0 \sim \mathrm{NI}(H)$, with $Z_0 \sim \mathrm{N}(0, 1)$ independent of U.

2.1 Shape analysis

Theorem 1 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then, the pdf of Y is*

$$f_Y(y) = \phi_{\text{NI}}(a_y)\, A_y, \quad y \in \mathbb{R}, \gamma \in \mathbb{R}, \delta > 0, \xi \in \mathbb{R}, \lambda > 0, \tag{5}$$

where $\phi_{\text{NI}}(\cdot)$ is the pdf of a standard NI distribution given in (4) and $a_y = a_y(\gamma, \delta, \xi, \lambda) = \gamma + \delta \operatorname{arcsinh}\Big([y-\xi]/\lambda\Big)$ and $A_y = A_y(\gamma, \delta, \xi, \lambda) = \frac{\delta}{\lambda}\left[\left\{\frac{y-\xi}{\lambda}\right\}^2 + 1\right]^{-\frac{1}{2}}$

Theorem 2 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then the r.v. Y given $U = u$, which is denoted by $Y|(U = u)$, follows the GLN distribution with parameters γ, δ, ξ, and λ, i.e.,*

$$Y|(U = u) \sim \text{GLN}(\sqrt{u}\gamma, \sqrt{u}\delta, \xi, \lambda).$$

Remark 1 *Note that if $U = 1$, we have that the GLNI distribution reduces the GLN model.*

Corollary 1 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H))$. Then,*

(i) The pdf of the r.v. $U|(Y = y)$ is given by

$$f_{U|Y}(u|y) = \frac{\phi(a_y; 0, 1/u)h(u)}{\phi_{\text{NI}}(a_y)}, u > 0;$$

(ii)

$$\mathbb{E}\Big(U|y\Big) = \frac{1}{\phi_{\text{NI}}(a_y)} \int_0^\infty u\phi\Big(a_y; 0, 1/u\Big) dH(u).$$

2.2 Distribution and quantile functions

We find here the cdf and quantile function (qf) of the GLNI distribution.

Theorem 3 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then, the cdf of Y is*

$$F_Y(y) = \Phi_{\text{NI}}(a_y); \quad y \in \mathbb{R}, \delta > 0, \xi \in \mathbb{R}, \lambda > 0,$$

where $\Phi_{\text{NI}}(\cdot)$ denotes the cdf of a NI distribution.

Corollary 2 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then, the qf of Y is*

$$y(q) = F_Y^{-1}(q) = \xi + \lambda \sinh\left(\frac{z(q) - \gamma}{\delta}\right),$$

where $z(q)$ is the q-th quantile of the normal/Independent distribution $NI(H)$ and $F_Y^{-1}(\cdot)$ is the inverse function of $F_Y(\cdot)$.

The following theorems provide some properties and the moments of the GLNI distribution.

Theorem 4 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then, the following holds:*

(i) $a + bY \sim \text{GLNI}(\gamma, \delta, a + b\xi, |b|\lambda; H)$, *where* $a \in \mathbb{R}$, $b \in \mathbb{R} - \{0\}$, *and*
(ii) $V = \left[\gamma + \delta \operatorname{arcsinh}([Y - \xi]/\lambda)\right]^2 \sim G\chi^2(H)$, *where* $G\chi^2(H)$ *denotes the generalized chi-square distribution*

Remark 2 *(i) From Theorem 4(i) we have that: The distribution of any linear combination of a r.v. with GLN distribution is also in the same class. For example, if* $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda,)$, *then* $W = [Y - \xi]/\lambda \sim \text{GLNI}(\gamma, \delta, 0, 1; H)$, *which can be called the standard GLNI distribution.*
(ii) The result in Theorem 4(ii) about the distribution of V, *enables us to check the model in practice, as we will see it in Section 5.*

Theorem 5 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then the k-th moment of Y is*

$$\mathbb{E}\left[Y^k\right] = \sum_{i=0}^{k}\sum_{j=0}^{i} \binom{k}{i}\binom{i}{j}\frac{\lambda^i}{2^i}\exp\left(-\frac{\gamma[2j-i]}{\delta}\right) + \mathbb{E}_U\left[\exp\left(\frac{[2j-i]^2 U}{2\delta^2}\right)\right]\xi^{k-i}\,[-1]^{i-j},$$

which depend on the moments of $\exp(bU)$, with $b > 0$. Note that $\mathbb{E}_U[\,\cdot\,]$ is taken with respect to the r.v. U.

Corollary 3 *Let $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$. Then, the mean and variance of Y are*

$$\mathbb{E}[Y] = \xi - \lambda \sinh\left(\frac{\gamma}{\delta}\right)\mathbb{E}_U\left[U_\delta\right] \quad \textit{and}$$

$$\text{Var}[Y] = \frac{\lambda^2}{2}\,\mathbb{E}\left[U_\delta - 1\right]\left[\cosh\left(\frac{2\gamma}{\delta}\right) + 1\right] + \lambda^2 \sinh^2\left(\frac{\gamma}{\delta}\right)\text{Var}_U\left[U_\delta\right],$$

respectively, where $\text{Var}_U(U_\delta) = \mathbb{E}_U[U_\delta^2] - \mathbb{E}_U^2[U_\delta]$, with $U_\delta = \exp\left(\frac{U}{2\delta^2}\right)$.

2.4 Special cases of the GLNI family

2.4.1 The GL-contaminated normal distribution

Consider the case when $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$, with H being the cdf of the r.v. U, which has a pdf of the form

$$h_U(u) = \nu_1 \mathbb{I}_{\{\nu_2\}}(u) + [1 - \nu_1]\mathbb{I}_{\{1\}}(u), 0 < \nu_1 < 1, 0 < \nu_2 < 1, \tag{6}$$

where $\mathbb{I}_{\{A\}}(\cdot)$ denotes the indicator function of the set A. Then, from equations (4) and (5), we have the pdf of the r.v. Y to be

$$f_Y(y) = [\nu_1\sqrt{\nu_2}\phi(\sqrt{\nu_2}a_y) + (1-\nu_1)\phi(a_y)]\, A_y, y > 0, \tag{7}$$

where $\phi(\cdot)$ is the standard normal pdf. The model with the pdf given as in equation (7) is the GL-CN distribution. In this case, the pdf of $U|(Y=y)$ is given by $h_{U|Y}(u|y) = \nu_1 p(y,u)\mathbb{I}_{\{\nu_2\}}(u) + [1-\nu_1]p(y,u)\mathbb{I}_{\{1\}}(u)$, where

$$p(y,u) = \frac{\sqrt{u}\exp\left(-\frac{ua_y^2}{2}\right)}{\nu_1\sqrt{\nu_2}\exp\left(-\frac{\nu_2 a_y^2}{2}\right) + (1-\nu_1)\exp\left(-\frac{a_y^2}{2}\right)}.$$

Thus,

$$\mathbb{E}[U|(Y=y)] = \frac{1-\nu_1+\nu_1\nu_2^{3/2}\exp\left(\frac{[1-\nu_2]a_y^2}{2}\right)}{1-\nu_1+\nu_1\sqrt{\nu_2}\exp\left(\frac{[1-\nu_2]a_y^2}{2}\right)}.$$

From Corollary 2, we have the mean and variance of Y depend on $\mathbb{E}_U[U_\delta^k]$, $k=1,2$ that in this case them are given by

$$\mathbb{E}_U[U_\delta^k] = \nu_1\left(\exp\left(\frac{k\nu_2}{2\delta^2}\right) - \exp\left(\frac{k}{2\delta^2}\right)\right) + \exp\left(\frac{k}{2\delta^2}\right), \quad k=1,2.$$

2.4.2 The GL-slash distribution

Consider the case when $Y \sim \text{GLNI}(\gamma,\delta,\xi,\lambda;H)$, with H being the cdf of the r.v. $U \sim Beta(\nu,1)$, which has a pdf of the form

$$h_U(u) = \nu u^{\nu-1}\mathbb{I}_{[0,1]}(u),\ \nu > 0. \tag{8}$$

Then, from equations (4) and (5), we have the pdf of the r.v. Y to be

$$f_Y(y) = \left[\nu\int_0^1 u^{\nu-1}\phi(a_y;0,\frac{1}{u})du\right] A_y, y > 0. \tag{9}$$

The model with the pdf given as in equation (8) is the GL-SL distribution. In this case, $U|(Y=y) \sim Gamma(1/2+\nu, a_y^2/2)$ truncated at $[0,1]$. Thus,

$$\mathbb{E}[U|(Y=y)] = \left[\frac{1+2\nu}{a_y^2}\right]\frac{P_1\left(\frac{3}{2}+\nu,\frac{a_y^2}{2}\right)}{P_1\left(\frac{1}{2}+\nu,\frac{a_y^2}{2}\right)},$$

where $P_x(a,b)$ denotes the cdf of the Gamma distribution of parameters a and b evaluated at x according to the parametrization established in the pdf given

in equation (7). To obtain the mean and variance of Y, we need of $\mathbb{E}_U[U_\delta^k]$, $k = 1, 2$ which are given by

$$\mathbb{E}_U[U_\delta^k] = \nu \int_0^1 u^{\nu-1} \exp\left(\frac{ku}{2\delta}\right) du, \ k = 1, 2.$$

2.5 The GL-St distribution

Consider the case when $Y \sim \text{GLNI}(\gamma, \delta, \xi, \lambda; H)$, with H being the cdf of the r.v. $U \sim Gamma(\nu/2, \nu/2)$, which has a pdf of the form

$$h_U(u) = \frac{[\frac{\nu}{2}]^{\frac{\nu}{2}} u^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})} \exp\left(-\frac{\nu u}{2}\right), u > 0, \nu > 0. \tag{10}$$

Then, from equations (4) and (5), we have the pdf of the r.v. Y to be

$$f_Y(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\sqrt{\tau}\Gamma(\frac{\nu}{2})} \left[1 + \frac{1}{\nu} a_y^2\right]^{-\frac{(\nu+1)}{2}} A_y, y > 0. \tag{11}$$

The model with the pdf given as in equation (10) is the GL-St distribution. In this case, $U|(Y = y) \sim Gamma([\nu + 1]/2, [\nu + a_y^2]/2)$. Thus,

$$\mathbb{E}[U|(Y = y)] = \frac{\nu + 1}{\nu + a_y^2}.$$

Moreover

$$\mathbb{E}_U[U_\delta] = \left(\frac{\nu\delta^2}{\nu\delta^2 - 1}\right)^{\nu/2}, \ \nu\delta^2 > 1 \text{ and } \mathbb{E}_U[U_\delta^2] = \left(\frac{\nu\delta^2}{\nu\delta^2 - 2}\right)^{\nu/2}, \ \nu\delta^2 > 2,$$

3 Estimation of the model

Next, we discuss ML estimation of the parameter $\boldsymbol{\theta} = (\gamma, \delta, \xi, \lambda)^\top$ for the GLNI distribution. We assume that the parameter vector $\boldsymbol{\nu}$ that indexes the pdf $h_U(\cdot)$ is known, and from now on the parameter of the model is $\boldsymbol{\theta}$. Thus, the observed-data log-likelihood function for $\boldsymbol{\theta}$ based on observed data set $\mathbf{y} = (y_1, \ldots, y_n)^\top$, is given by $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$, where

$$\ell_i(\boldsymbol{\theta}) = \log[\phi_{\text{NI}}(a_{yi})] + \log[A_{yi}], \tag{12}$$

with a_{yi} and A_{yi} are as in Theorem 1.

3.1 The observed information matrix

Letting $I_i^{\phi}(w) = (1/\sqrt{2\pi})E_U\left[U^w e^{-Ua_{yi}^2/2}\right]$, $i = 1, .., n$, we have the score function given by $U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta})$, where $U_i(\boldsymbol{\theta})$ has elements given by

$$U_{i\eta}(\boldsymbol{\theta}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \eta} = \frac{1}{\phi_{NI}\left(a_{yi}\right)} \frac{\partial \phi_{NI}\left(a_{yi}\right)}{\partial \eta} + \frac{1}{A_{yi}} \frac{\partial A_{yi}}{\partial \eta}, \ \eta = \gamma, \delta, \xi, \lambda, \tag{13}$$

where $\dfrac{\partial \phi_{NI}\left(a_{yi}\right)}{\partial \eta} = -I_i^{\phi}\left(3/2\right) a_{yi} \dfrac{\partial a_{yi}}{\partial \eta}$. And the observed information matrix is given by

$$\mathbf{J}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \left(\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \eta \partial \tau} \right), \ \ \eta, \tau = \gamma, \delta, \xi, \lambda \tag{14}$$

where

$$\begin{aligned} \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \eta \partial \tau} = & -\frac{1}{\left(\phi_{NI}\left(a_{yi}\right)\right)^2} \frac{\partial \phi_{NI}\left(a_{yi}\right)}{\partial \eta} \frac{\partial \phi_{NI}\left(a_{yi}\right)}{\partial \tau} + \frac{1}{\phi_{NI}\left(a_{yi}\right)} \frac{\partial^2 \phi_{NI}\left(a_{yi}\right)}{\partial \eta \partial \tau} \\ & -\frac{1}{(A_{yi})^2} \frac{\partial A_{yi}}{\partial \eta} \frac{\partial A_{yi}}{\partial \tau} + \frac{1}{A_{yi}} \frac{\partial^2 A_{yi}}{\partial \eta \partial \tau} \end{aligned}$$

with $\dfrac{\partial^2 \phi_{NI}\left(a_{yi}\right)}{\partial \eta \partial \tau} = I_i^{\phi}\left(\dfrac{5}{2}\right) a_{yi}^2 \dfrac{\partial a_{yi}}{\partial \eta} \dfrac{\partial a_{yi}}{\partial \tau} - I_i^{\phi}\left(\dfrac{3}{2}\right) \left(\dfrac{\partial a_{yi}}{\partial \eta} \dfrac{\partial a_{yi}}{\partial \tau} + a_{yi} \dfrac{\partial^2 a_{yi}}{\partial \eta \partial \tau} \right)$. The derivatives of a_{yi} and A_{yi} involve standard algebraic manipulations which are given in the Appendix A.

3.2 ML estimation via EM-algorithm

In this section we develop an EM-type algorithm for maximum likelihood estimation of the parameters of for the GLNI distribution. In order to do this, we first represent the GLNI model in an incomplete data framework using the stochastic representation given in Theorem 2. Thus, we consider the following hierarchical representation for Y_i

$$Y_i | U_i = u_i \overset{\text{ind}}{\sim} \text{GLN}(\sqrt{u_i}\gamma, \sqrt{u_i}\delta, \xi, \lambda), \tag{15}$$

$$U_i \overset{\text{ind}}{\sim} H(.; \boldsymbol{\nu}) \tag{16}$$

We assume that the parameter vector $\boldsymbol{\nu}$ that indexes the pdf $h_U(\cdot)$ is known Specifically, the EM-type algorithm that we produce for ML estimation in GLNI model considers $\boldsymbol{u} = [u_1, \ldots, u_n]^\top$ are missing data augmented with the observed data set $\boldsymbol{y}$ and so the complete data set is $\boldsymbol{y}_c = [\boldsymbol{y}^\top, \boldsymbol{u}^\top]^\top$. Hence, under the hierarchical representation given in Eqs. (15) and (16), the EM-type algorithm is applied to the complete-data log-likelihood function that can be written as (without the additive constant) $\ell(\boldsymbol{\theta}|\boldsymbol{y}_c) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}|y_i, u_i)$, where

$$\ell_i(\boldsymbol{\theta}|y_i, u_i) = \log(\delta) - \log(\lambda) - \frac{u_i}{2}\, a_{yi}^2 - \frac{1}{2}\log\left[\left(\frac{y_i - \xi}{\lambda}\right)^2 + 1\right]$$

with a_{yi} defined as in Eq. (12), for $i = 1, \ldots, n$.

The E-step of the EM-type algorithm requires the evaluation of $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \mathrm{E}[\ell(\boldsymbol{\theta}|\boldsymbol{y}_c)|\boldsymbol{y}, \hat{\boldsymbol{\theta}}]$, where the expectation is taken with respect to the conditional distribution of U given $\boldsymbol{Y} = \boldsymbol{y}$ and $\hat{\boldsymbol{\theta}}$. Considering the estimate of $\boldsymbol{\theta}$ at the rth iteration, say $\hat{\boldsymbol{\theta}}^{(r)} = (\hat{\gamma}^{(r)}, \hat{\delta}^{(r)}, \hat{\xi}^{(r)}, \hat{\lambda}^{(r)})^\top$, and letting $\hat{u}_i^{(r)} = \mathrm{E}[U_i|y_i, \hat{\boldsymbol{\theta}}^{(r)}]$, we obtain the $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ has the form

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \mathrm{E}\left[\ell(\boldsymbol{\theta}|\boldsymbol{y}_c)|\boldsymbol{y}, \hat{\boldsymbol{\theta}}\right] = \sum_{i=1}^n Q_i(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}), \tag{17}$$

where $Q_i(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \log(\delta) - \log(\lambda) - \frac{1}{2}\hat{u}_i\, a_{yi}^2 - \frac{1}{2}\log\left[\left(\frac{y_i-\xi}{\lambda}\right)^2 + 1\right]$.

We then have the following EM-type algorithm:

E-step. Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(r)}$, compute $\hat{u}_i{}^{(r)}$, for $i = 1, \ldots, n$.

CM-step 1. Fix $\hat{\xi}^{(r)}$ and $\hat{\lambda}^{(r)}$, and update $\hat{\gamma}^{(r)}$ and $\hat{\delta}^{(r)}$ as

$$\begin{aligned}
\widehat{\gamma}^{(r+1)} &= -\frac{\hat{\delta}^{(r+1)}}{\overline{u}^{(r)}}\left[\frac{1}{n}\sum_{i=1}^n \hat{u}_i^{(r)} \operatorname{arcsinh}\left(\frac{y_i - \hat{\xi}^{(r)}}{\hat{\lambda}^{(r)}}\right)\right], \\
\hat{\delta}^{2\,(r+1)} &= \frac{\overline{u}^{(r)}}{\overline{u}^{(r)}\,\frac{1}{n}\sum_{i=1}^n \hat{u}_i^{(r)}\left[\operatorname{arcsinh}\left(\frac{y_i-\hat{\xi}^{(r)}}{\hat{\lambda}^{(r)}}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^n \hat{u}_i^{(r)} \operatorname{arcsinh}\left(\frac{y_i-\hat{\xi}^{(r)}}{\hat{\lambda}^{(r)}}\right)\right]^2},
\end{aligned}$$

where $\overline{u}^{(r)} = \frac{1}{n}\sum_{i=1}^n \hat{u}_i^{(r)}, \quad i = 1, \ldots, n$.

CM-step 2. Fix $\hat{\gamma}^{(r+1)}$ and $\hat{\delta}^{(r+1)}$, and update $\hat{\xi}^{(r)}$ and $\hat{\lambda}^{(r)}$ as

$$(\hat{\xi}^{(r+1)}, \hat{\lambda}^{(r+1)}) = \underset{\xi,\lambda}{\operatorname{argmax}}\, Q(\hat{\gamma}^{(r+1)}, \hat{\delta}^{(r+1)}, \xi, \lambda|\hat{\boldsymbol{\theta}}^{(r)}).$$

Remark 3 *We note that when* $\gamma = 0$ *CM-step 1 of the EM-type algorithm presented above reduce to*

$$\hat{\delta^2}^{(r+1)} = \frac{n}{\sum_{i=1}^{n} \hat{u}_i^{(r)} \left[\text{arcsinh}\left(\frac{y_i - \hat{\xi}^{(r)}}{\hat{\lambda}^{(r)}}\right)\right]^2}.$$

On the other hand, if $U = 1$ *in the EM-algorithm proposed (i.e., if the r.v.* U *is degenerate), the EM-algorithm here presented might be to supplement of the recent work presented by Leiva et el. (2009) in the estimation context of the parameter.*

4 Application of the model in Genetics

In this section, we obtain results for the GLNI distribution to a real data set from gene expression. First, an implementation in `R` code of the proposed methodology is discussed. Second, the problem upon analysis is discussed. Third, an exploratory data analysis (EDA) is produced. Finally, analysis based on the GLN and GLNI distribution are carried out first using ML methods for estimating the parameters of these distributions and then model checking tools.

4.1 Implementation in R code

We have developed a new `R` package named `glni` to analyze data from the GLNI model, which is available upon request. This package contains diverse probabilistic indicators and allows practitioners to compute ML estimates of the parameters of the GLNI distribution and the standard likelihood ratio (LR) test, which is useful for checking the suitability of the GLNI distribution with respect to nested models inside it.

4.2 Description of the problem

Microarrays are solid supports onto which nucleotide probes, such as DNA, are immobilized. The probes are chosen such that they bind to specific sample molecules; for DNA arrays, this is ensured by the sequence-specificity of the hybridization reaction between complementary DNA strands. The unknown DNA from one or two biological samples is prepared in solution, labelled with fluorescent dyes and allowed to bind to the array, for one-color or two-color

arrays, respectively. The abundance of sample DNA molecules among different conditions can then be studied by comparing the fluorescence intensities at the matching probe sites. We implement our method for analyzing a subset of the two-color (green and red representing two channels) array data from the H25K experiment. These data are included in the Microarray Quality Control (MAQC) data sets and have been publicly available since October 2006. The experiment that allowed the H25K data used here to be generated was carried out by using the MAQC guidance with its H25K Human Genome Microarrays. Specifically, each microarray was hybridized with RNA tissues (samples) A and B, which were mixed at TeleChem ArrayIt to produce mixture C. Two colors were used for labelling the RNA of each type of tissue. Thus, the H25K two-color sample-pairs were generated using the convention: `A` (`A/B`, `Cy3/Cy5`), `B` (`B/A`, `Cy3/Cy5`), `sA` (`A/A`, `Cy3/Cy5`) and `sB` (`B/B`, `Cy3/Cy5`), where `A` means that the tissues A and B were hybridized to the green (Cy3) and red (Cy5) channels, respectively; `B` means that the tissues B and A were hybridized to the green (Cy3) and red (Cy5) channels, respectively; `sA` means that the tissues A and B were hybridized to the green (Cy3) and red (Cy5) channels, respectively; and `sB` means that the tissues B and A were hybridized to the green (Cy3) and red (Cy5) channels, respectively; for more details about this experiment see Patterson et al. (2006) and http://arrayit.com/Products/Microarrays/H25K/h25k.html.

4.3 Exploratory data analysis

As it is well known, due to variations in experimental factors, such as amount of sample mRNA or labelling and hybridization efficiencies, the intensities cannot be directly compared and therefore calibrated intensities must be used. The gene expression intensities for the a subset of the H25K data, which we will call H2KS, were calibrated using the Huber's method (Huber et at., 2003) by means of the `R` package (R Development Core Team, 2008) named `vsn` available from http://www.bioconductor.org and CRAN (http://CRAN.R-project.org/). Table 1 presents a descriptive summary of 30 tissues while Fig. 3 (left side) shows the histogram and boxplot of these data. An EDA of the H25KS calibrated intensities based on Table 1 and the histogram in Fig. 4 shows a negatively skewed distribution with high kurtosis and variability degrees. The GLNI distribution considers the degrees of variability, skewness and kurtosis presented in the data. We propose the GLNI distribution for modeling these data.

Table 1
Descriptive statistics for H25KS calibrated intensities

Median	Mean	SD	CV	CS	CK	Range	Min.	Max.	n
1116.5	1127.6	3236.9	2.9	−1.0	4.0	18523.4	−10623.3	7900.1	30

5 Results

The maximum likelihood estimates of $\boldsymbol{\theta} = (\xi, \delta, \lambda, \gamma)'$ using the H25KS data set have been calculated via the EM-algorithm proposed in Section 3.2 by beginning with ML estimates of $\boldsymbol{\theta}$ in the GLN model reported in Leiva et al. (2009). Results are displayed in Table 2. For the GL-St, GL-SL and GL-CN models we have chosen $\boldsymbol{\nu}$ that maximizes the likelihood function and then established as ML estimates of $\boldsymbol{\theta}$ those associated with the maximum likelihood function. We have estimated $\boldsymbol{\theta}$ for the GLN model using the GL-St model with $\nu_1 = 100$, assuming that ν_1 is large enough to assure convergency to the GLN model.

Table 2
ML estimation results for fitting the GLN, GL-St, GL-SL and GL-CN models for the data set. SDs are the estimated asymptotic standard deviations based on the observed information matrix.

	GLN	GL-St	GL-SL	GL-CN
	(SD)	(SD)	(SD)	(SD)
ξ	1128.592	1128.592	1128.592	1128.592
	(1123.27)	(1428.21)	(1474.48)	(776.45)
δ	1.488	1.779	2.051	2.333
	(0.21)	(0.45)	(0.44)	(0.83)
λ	3360.834	3360.834	3360.834	3360.834
	(235.46)	(769.71)	(490.6)	(1402.4)
γ	-0.025	-0.026	-0.055	0.024
	(0.47)	(0.74)	(0.87)	(0.57)
ν_1	100	6	2	0.4
ν_2	-	-	-	0.2

All models have produced estimates equal to 1128.592 and 3360.834 for the location ($\hat{\xi}$) and scale ($\hat{\lambda}$) parameters, respectively. Estimates of the parameter γ (shape) are close to zero under all models, suggesting the distributions are symmetric. The GLN model estimated the parameter δ (shape) close to 1.5, and the other models close to 2. The standard deviations of the estimators of the parameters have been produced and are dispalyed in Table 2. None of the models have been absolute in producing the smallest standard deviations for all parameters simultaneously. Graphics on the top of Figure 1 present the weights used by the EM procedure for estimating $\boldsymbol{\theta}$; the weights for the GLN distribution are indicated as a horizontal line. In particular, we notice that the GL-NI models attribute smaller weight to observation number nine than

the GLN model. In order to detect outlying observations, the Mahalanobis distance has been considered. Graphics on the bottom of Figure 1 display such distances for the GL-NI fitted models. The cutoff lines corresponds to the quantile of 0.95. In these graphics, we highlight observation 9 as possible outlier in all models. Replacing the MLE estimates of $\boldsymbol{\theta}$ in the Mahalanobis distance δ_k, in Figure 2 we present simulated envelopes (lines represent the 5^{th} percentile, the mean, and the 95^{th} percentile of the 100 simulated points for each observation). These plots suggest evidence that the GL-NI distributions provide a better fit to the data set, in particular to observation nine, than the GLN distribution. The GL-St model seems to accomodate observation number 9 better, by visually inspecting the envelope plots.

We have calculated the relative change (RC) in percentage of each parameter estimate defined by: $RC_j = \left|\frac{\hat{\theta}_j - \hat{\theta}_{j[-i]}}{\hat{\theta}_j}\right| \times 100\%$, where $\hat{\theta}_{j[-i]}$ is the ML estimate of $\boldsymbol{\theta}_j$ with the i^{th} observation deleted and $\theta_1 = \xi$, $\theta_2 = \delta$, $\theta_3 = \lambda$ and $\theta_4 = \gamma$. RCs are displayed in Table 3. The smallest RCs have been detected on the MLE of δ for the GL-St model, and on the MLE of γ for the GL-SL model; for the remaining parameters, the RCs are zero for all models.

We assess how much the ML estimates of $\boldsymbol{\theta}$ are influenced by a change of ϵ units in a single observation by recording the relative change in the estimates, defined as $RC(\epsilon) = \left|\frac{\hat{\theta}_j - \hat{\theta}(\epsilon)_j}{\hat{\theta}_j}\right| \times 100\%$, where $\hat{\theta}(\epsilon)_j$ is the ML estimate of $\boldsymbol{\theta}_j$ with the i^{th} observation contaminated as $y_i + \epsilon$ and $\theta_1 = \xi$, $\theta_2 = \delta$, $\theta_3 = \lambda$ and $\theta_4 = \gamma$. In Figure 3 we present the results of relative changes of the estimates of ξ, δ, λ and γ by contaminating observation 22 of the H25KS data with an addition of an ϵ that varies from -50 to 50, by steps of size 10. The estimates of the parameters by the GL-St, GL-SL and GL-CN models are less affected by variations of ϵ than the GLN model. In particular, the GL-St model had the smallest variation.

In order to verify the fit of the glog distributions to H25KS, we have used the invariance property of the MLEs for estimating the GL-NI pdfs, which are shown in Figure 4 on the histogram and empirical pdf of the data, respectively. The results presented here show the good agreement between the GLN and GL-NI distributions with the H25KS data. However, we can observe that the GL-NI models tend to accomodate the tails better than the GLN model.

6 Concluding remarks

In this paper, we have discussed the GL-NI distributions and its extension based on the glog-normal model. These distributions are very flexible on kurtosis and skewness. This is because the GL-NI distributions have greater and

Table 3
RCs for the parameters of the GLN, GL-St, GL-SL and GL-CN models.

Observation	Distribution	RC_ξ	RC_δ	RC_λ	RC_γ
#9	GLN	0	16.1	0	459.3
	GL-St	0	4.7	0	297.5
	GL-SL	0	12.1	0	169.3
	GL-CN	0	26.6	0	684.7

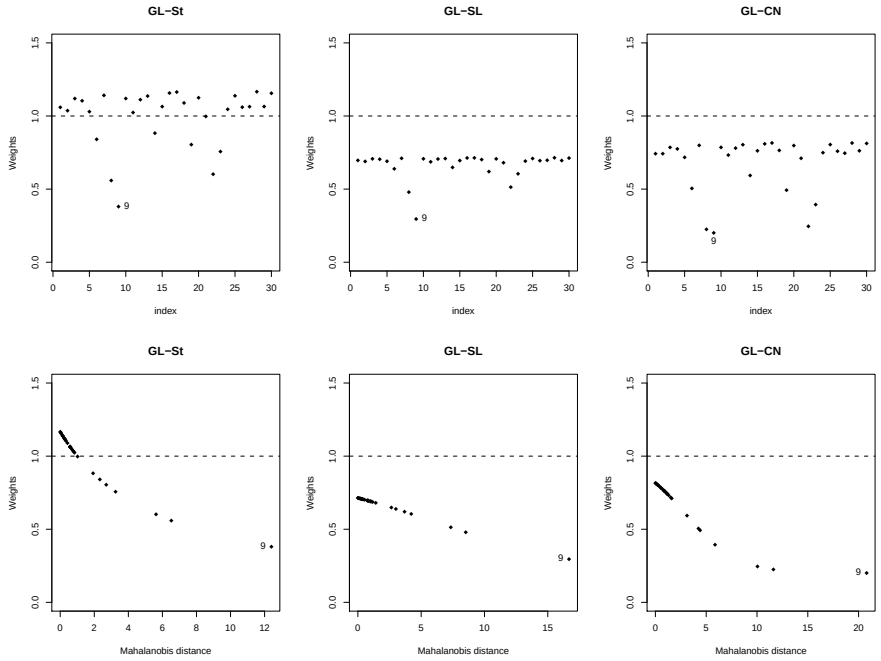

Fig. 1. Graphics of the EM estimated weights (top) and of the Mahalanobis distances (bottom) for the GLN, GL-St, GL-SL and GL-CN models using the H25KS data.

lesser kurtosis than the GLN model, allowing positive and negative skewness as well as symmetry. We have derived the GL-NI density and have carried out a shape analysis of the distribution in order to see how its parameters influence the shape and form of the pdf. Thus, we have presented the GL-NI model as a statistical distribution that can be useful for modeling gene expression data without transforming them as has been considered by using the logarithmic or glog transformations. An application to real gene expression data showed that the GL-NI models accomodate the data better than the normal-glog model for producing more robust parameter estimates. In particular, the GL-St distribution showed to be more appropriate for fitting the the data set.

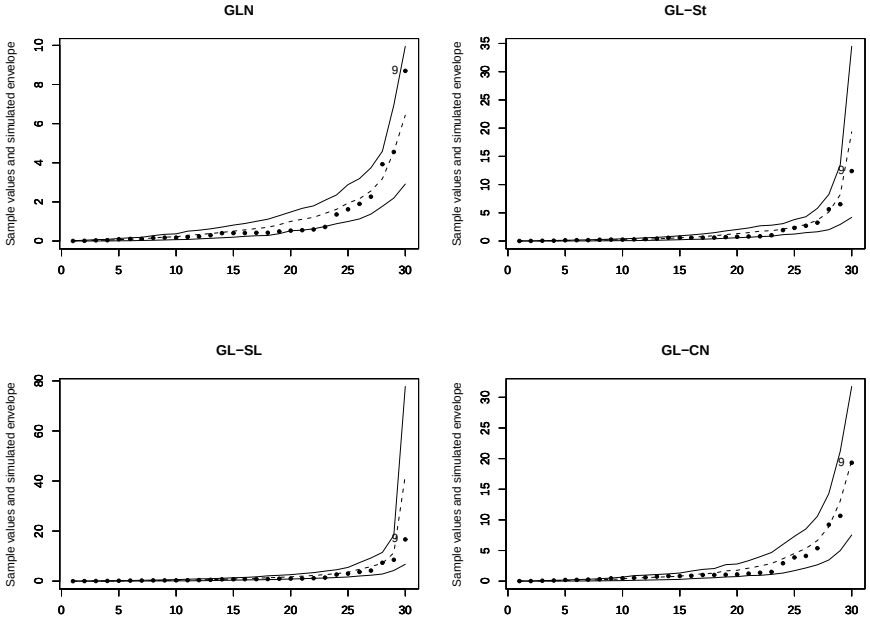

Fig. 2. Simulated envelopes using the GLN, GL-St, GL-SL and GL-CN distributions.

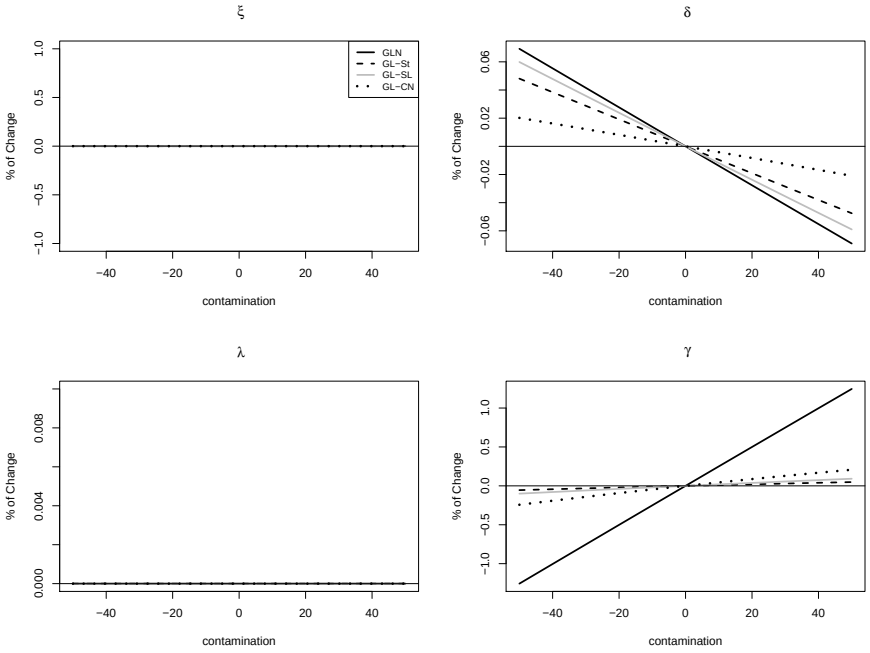

Fig. 3. Relative changes in the ML estimates of ξ, δ, λ and γ of fitting a GLN, GL-St, GL-SL and GL-CN distributions for different contaminations of observations #22 in the H25KS data.

References

Durbin, B.P., Hardin, J.S., Hawkins, D.M., Rocke, D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. Bioinfor-

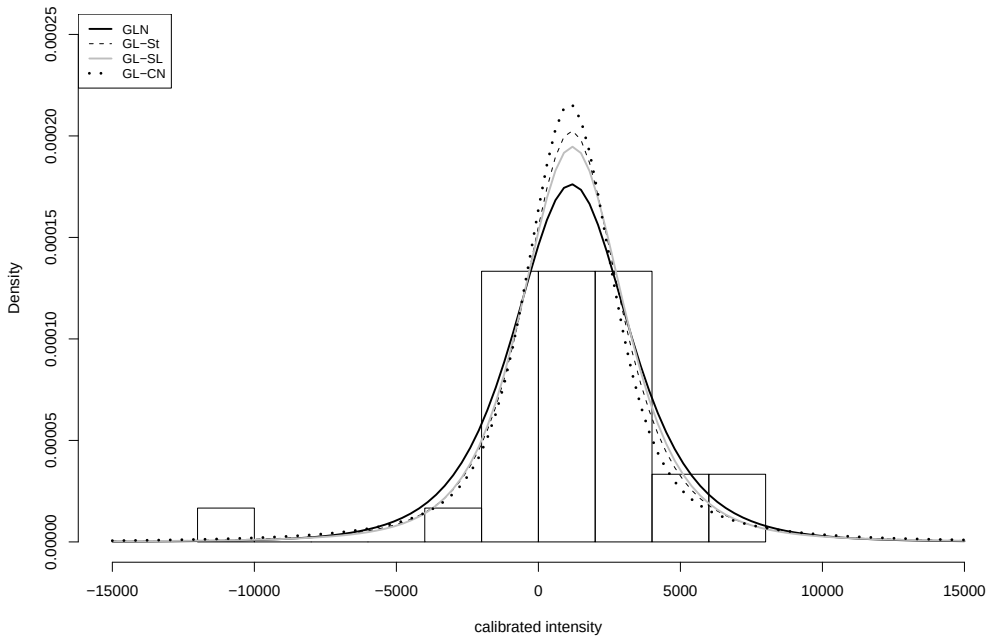

Fig. 4. Histogram of the H25KS data with estimated pdfs using the GLN, GL-St, GL-St and GL-CN distributions.

matics, 18, S105-S110.

George, F. (2007) Johnson's System of Distributions and Microarray Data Analysis. Unpublished PhD dissertation. Department of Mathematics College of Arts and Sciences, University of South Florida, USA.

Gneiting, T. (1997) Normal scale mixtures and dual probability densities. J. Stat. Comp. Simul., 59, 375-384.

Huber, W., Heydebreck, A., Sültmann, H. Poustka, A. Vingron, M. 2003. Parameter estimation for the calibration and variance stabilization of microarray data. In "Statistical Applications in Genetics and Molecular Biology". Vol. 2(1), Article 3. The Berkeley Electronic Press, Berkeley.

Huang, S, Qu, Y. (2006) The loss in power when the test of differential expression is performed under a wrong scale. J. Comput. Biol. 13, 786-797.

Johnson, N.L. (1949) Systems of frequency curves generated by methods of translation. Biometrika, 36, 149-176.

Lange, K. and Sinsheimer, J.S. (1993) Normal/independent distributions and their applications in robust regression. J. Comp. Graph. Stat., 2, 175-198.

Leiva, V., Sanhueza, A. Kelmansky, S., Martinez, E. (2009) On the glog-normal distribution and its association with the gene expression problem. Comp. Stat. Data Anal., 53, 1613-1621.

R Development Core Team. (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rocke, D.M., Durbin, B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics, 19, 966-972.