Characterization, properties and applications of gpot-normal distributions

Lila Ricci*
Universidad Nacional de Mar del Plata, Mar del Plata, Argentina - lricci@mdp.edu.ar

Diana Kelmansky
Universidad de Buenos Aires, Buenos Aires, Argentina - dkedkelman@ic.fcen.uba.ar

Abstract

A new family of probability distributions is introduced, the *gpot*-normal. It was motivated as an alternative approach to transformations for microarray intensity, characterized by the presence of zero and negative values, however this family has shown to be useful also for modeling a wider type of data. It is proved that random variables that belong to the *gpot*-normal distribution family, when they are transformed with a suitable transformation become normal or truncated normal distributed. It is also shown that the proposed family of density functions constitute a pseudo-dispersion model, defined by Jørgensen in 1997 and its deviance and unit variance are obtained. An expression is given for the moments and for the quantiles, in terms of the truncated normal density. A combined maximum likelihood method is proposed to estimate the model parameters, and it is applied to microarray and chemical concentration data.

Keywords: Gpot-normal; Pseudo-dispersion models; Microarrays; Truncated normal.

1. Introduction

Box-Cox transformations [1] have been widely used to normalize asymmetric data; however they can not be applied when there are zero and negative values. This occurs, for example, in chemical concentration and microarray data. In this last special case, the following transformations that approximately normalize the data have been proposed.

- $log_2(Y)$: it allows working with log-ratios which have a simple and intuitive meaning for biologists (see Speed [11] and Smyth, Yang and Speed [10]) and it usually works well for high values but not for zero, negative and low values.
- Generalized logarithm, allows negative values. It usually works well for low values but it is too severe for high values [11]; this transformation was introduced by Durbin *et al.* [2] and by Huber *et al.* [4].
- The family of generalized power transformation was defined for real supported data by Kelmansky *et. al.* [6] as:

$$gpot\,(Y;p) = \begin{cases} \frac{\left(Y+\sqrt{Y^2+1}\right)^p - 1}{p} & p \neq 0 \\ \ln\left(Y+\sqrt{Y^2+1}\right) & p = 0 \end{cases} \tag{1}$$

  it includes *glog* as a special case in the same continuous way as the Box-Cox transformations family includes the natural logarithm.

Instead of looking for the best transformation, the alternative standpoint taken in this work was the search of probabilistic models that fit the observed data in its original scale; this approach has the advantage of keeping that scale scale, facilitating a direct interpretation of the results. With this focus Freeman and Modarres [3] introduced the power-normal distribution in relation to the Box-Cox transformation and Leiva *et al.* [7] proposed the glog-normal distribution family in relation to the *glog* transformation.
The main idea of the present article is to generalize the above results, considering those models that become truncated normals after a gpot transformation, such as occurs for some positive distributions with the Box-Cox transformation family. In Section 1 gpot-normal models are defined and its main properties are demonstrated; their relation with pseudo-dispersion models is studied in Section 2; expressions for the quantiles are obtained in Section 3. Then, in Section 4 a method is described to obtain estimators of the

parameters and real data applications of these methods are given in Section 5. Finally the conclusions and a discussion are presented in Section 6.

1 Gpot-normal distribution

A new family of distributions will be defined and it will be proved that their image under a gpot transformation is a truncated normal random variable.

Definition 1.1. A random variable Y has a gpot-normal distribution if for some $\mu \in \mathbf{R}$, $\sigma > 0$, $p \in \mathbf{R}$ its density function is given by

$$f_Y\left(y;\mu,\sigma^2,p\right) = \frac{1}{K\sqrt{2\pi\sigma^2}} \frac{\left(y+\sqrt{y^2+1}\right)^p}{\sqrt{y^2+1}} \exp\left(-\frac{1}{2\sigma^2} d_p\left(y,\mu\right)\right), \quad y \in \mathbf{R} \tag{2}$$

where

$$K = \begin{cases} 1 - \Phi\left(\frac{-1/p - gpot(\mu,p)}{\sigma}\right) & \text{if } p > 0 \\ 1 & \text{if } p = 0 \\ \Phi\left(\frac{-1/p - gpot(\mu,p)}{\sigma}\right) & \text{if } p < 0 \end{cases} \tag{3}$$

is the normalizing constant, Φ is the cumulative distribution function of the standard normal distribution and $d_p\left(y,\mu\right) = \left(gpot(y;p) - gpot(\mu,p)\right)^2$ is the deviance. This will be denoted as $Y \sim gpotN(\mu,\sigma,p)$.

The next theorem proves the main property of gpot-normal variables: that they become truncated normals after a gpot transformation.

Theorem 1.2. *Let $Y \sim gpotN(\mu,\sigma,p)$, then the transformed variable $X = gpot\left(Y;p\right)$ has a truncated normal distribution $TN\left(\mu_X,\sigma^2,-1/p,\infty\right)$ if $p > 0$, $TN\left(\mu_X,\sigma^2,-\infty,-1/p\right)$ if $p < 0$ and normal distribution $N\left(\mu_X,\sigma^2\right)$ if $p = 0$ with $\mu_X = gpot(\mu,p)$.*

2 Relationship between gpot-normal models and pseudo-dispersion models

Pseudo-dispersion models were defined by Jorgensen [5]. In this Section it is proved that the densities defined by (2) belong to this family. Expressions for their deviance and unit variance functions are also obtained.
The definition can be extended, to pseudo-dispersion models. A remarkable property of these models is that they are asymptotically normal (see [5]).
It is proved that *gpot-normal* models are pseudo-dispersion models.

Theorem 2.1. *Models given in definition 1.1 are pseudo-dispersion models.*

Besides, note that it is enough to define

$$s(y)^T = (gpot^2(y;p), -\sqrt{2}gpot(y;p), 1)$$
$$\alpha(\mu)^T = (1, \sqrt{2}gpot(\mu,p), gpot^2(\mu,p)).$$

to be able to express $d_p(y,\mu)$ as $\mathbf{s}(y)^T\alpha(\mu)$. In particular, if $p = 0$ the variance function is $V\left(\mu\right) = \mu^2 + 1$ and it coincides with the variance function of a generalized hyperbolic secant, that corresponds to a Morris model [8].
In Figure 1 graphic representations of these densities are exhibited for various values of p and σ, always with $\mu = 0$.

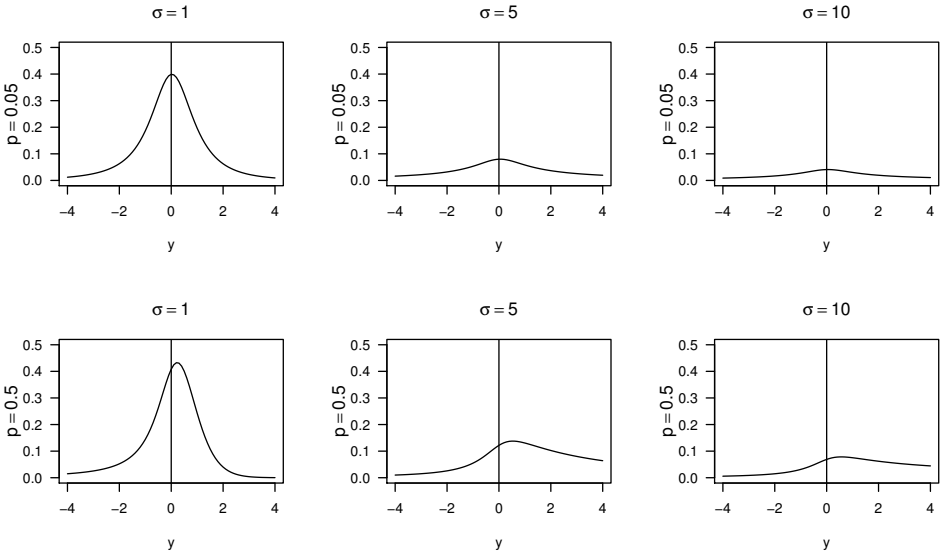

Figure 1: Gpot-normal densities for some values of p and σ, always with $\mu = 0$.

3 Quantiles

In order to avoid the difficulties in moment calculations (i.e. the mean and variance), a straightforward alternative is considering quantiles such as the median and quartiles instead. Further, quantiles enable the graphical examination of the model fit to a data set using quantile-quantile plots.
Let Y be a gpot-normal random variable and $X = gpot(Y;p)$ distributed as $TN\left(\mu_X, \sigma^2, -1/p, \infty\right)$ the transformed variable. Let x_α be the α-quantil for X, it is proved that $x_\alpha = \sigma\Phi^{-1}\left(K_X\left(\alpha - 1\right) + 1\right) + \mu_X$. Also

$$\alpha = P\left(X \leq x_\alpha\right) = P\left(gpot\left(Y\right) \leq x_\alpha\right) = P\left(Y \leq gpot^{-1}\left(x_\alpha\right)\right)$$

then the α-quantile for Y is $y_\alpha = gpot^{-1}\left(x_\alpha\right)$ an its value can be obtained from the standard normal distribution.

4 Parameter estimation

Gpot-normal models have three parameters to be estimated. They are related to the corresponding TN model parameters by the following expressions: $\mu = gpot^{-1}\left(\mu_x\right)$; $\sigma^2 = \sigma_x^2$ and the power p being $-1/T$, where T is the truncation point.
We propose a combined profile likelihood and maximum likelihood approach to estimate the parameters, that is described in detail in the following paragraphs. Given a data set represented by vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, to obtain a profile likelihood for the power p we consider a grid of values $p_0, p_1, \ldots, p_k$ and, for each p_j, $1 \leq j \leq k$ the transformed data $\mathbf{x}_{p_j}$ are calculated as

$$\mathbf{x}_{p_j} = gpot\left(\mathbf{y}, p_j\right).$$

Then, for each p_j, the corresponding μ_{p_j} and σ_{p_j} are estimated, maximizing the likelihood function of the truncated normal variable. Then p_j, μ_{p_j} and σ_{p_j} are used to obtain the log-likelihood function of $\mathbf{y}$ whose density was given by (2):

$$\ln f_{\mathbf{y}}\left(\mathbf{y}; \mu, \sigma^2, p\right) = \quad n \ln \frac{1}{K\sqrt{2\pi\sigma_{p_j}^2}} + \sum_{i=1}^{n}\left(\ln p\left(y_i + \sqrt{y_i^2 + 1}\right) - \ln\sqrt{y_i^2 + 1}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} d_p\left(y_i, \mu_{p_j}\right).$$

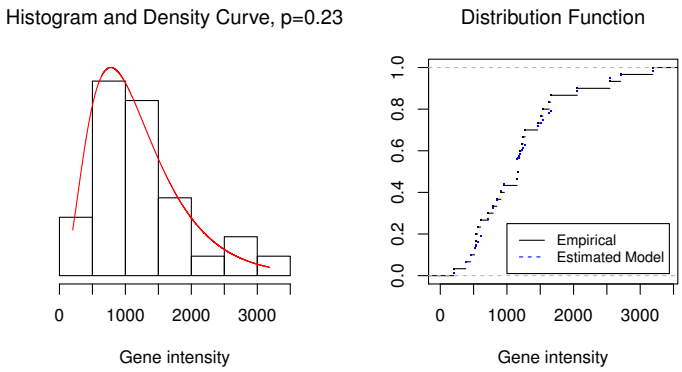

Figure 2: Overlap of the histogram (left) and the empirical distribution curve (right) with the adjusted model density curve and distribution function respectively for gene expression data.

Finally, p is chosen as the one that maximize the log-likelihood in the grid:

$$\widehat{p} = \max_{1 \leq j \leq k} f_{\mathbf{y}} \left(\mathbf{y}; \mu_{p_j}, \sigma^2_{p_j}, p_j \right)$$

The computer program was written in R language [9] and it can be requested to the authors.

5 Real data applications

As it was mentioned in the Introduction the proposed density family was originally motivated by the modeling of microarray intensities, but its application is more general. Here we present two examples: the first one corresponds to intensities of microarray data and the second one to concentrations of ammonia. For these examples the parameter estimators are obtained by the method described in Section 4. Then, the data fit to the corresponding estimated model is shown in two ways. First by the overlap of the data histograms with the density curves. Second by the overlap of the empirical distribution curve with the adjusted model cumulative distribution function. Q-Q plots are also displayed. Only the first example will be detailed.

Example 5.1. The first set of data corresponds to intensities of one gene from the H25K_2: Yale University MAQC project. As can be seen in Figures 2 and 3 it is well fitted by a gpot-normal model with $p = 0.23$.

6 Conclusions

A new family of distributions named gpot-normal indexed by $p \in \mathbf{R}$ has been defined. Its name comes from the fact that variables whose distribution belongs to this family become normal or truncated normal when a gpot transformation is applied. This kind of transformations, unlike the Box-Cox transformation, can be used with data having both negative and positive values, thus the gpot-normal family can model data that include non positive values.
To obtain estimators of the model parameters a combined maximum likelihood method is proposed and successfully applied to real data.
It has been proved that the gpot-normal family is a special case of pseudo-dispersion models.
Obtaining quantiles is straightforward for all values of p. From these quantiles, position and scale measures (i.e. the median and the interquartile range) can be defined. Also the specific quantile-quantile plot is a useful tool for a visual evaluation of real data fit to the estimated models.

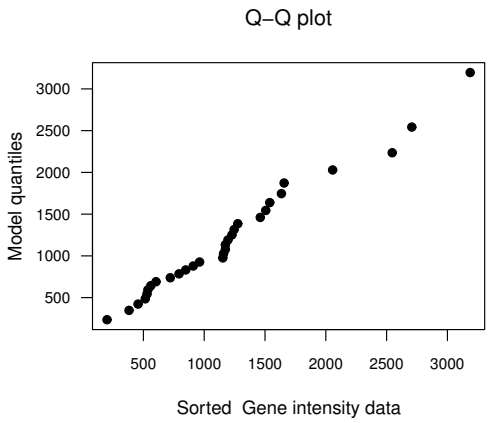

Figure 3: QQ-plot for gene expression data.

References

[1] Box, G. E. P. and Cox, D. R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[2] B. Durbin, J. Hardin, D. Hawkins, and D. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:247–252, 2002.

[3] J. Freeman and S. Modarres. Inverse box-cox: the power-normal distribution. *Statistics and Probability Letters*, 76:S105–S110, 2006.

[4] W. H. Huber, A. Sueltmann H., Poustka, A., and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2, 2003.

[5] B. Jorgensen. *The Theory of Dispersion Models*. Chapman and Hall, 1997.

[6] D. M. Kelmansky, E. J. Martnez, and V. Leiva. A new variance stabilizing transformation for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 2013. To appear.

[7] V. Leiva, A. Sanhueza, D. Kelmansky, and E. Martinez. On the glog-normal distribution and its association with the gene expression problem. *Computational Statistics and Data Analysis*, 53:1613–1621, 2009.

[8] C. N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10:65–80, 1982.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[10] G. Smyth, Y. Yang, and T. Speed. Statistical issues in cdna microarray data analysis. *Methods in Molecular Biology*, 224:111–136, 2003.

[11] T. Speed. *Statistical Analysis of Gene Expression Data*. Chapman and Hall, 2003.