Discriminating between and within (semi)continuous classes of both Tweedie and geometric Tweedie models

Abid Rahma*
University of Sfax, Laboratory of Probability and Statistics & University Paris-Dauphine Tunis, Department of Mathematics, Tunisia - rahma.abid.ch@gmail.com & rahma.abid@dauphine.tn

Kokonendji Célestin C.
Université Bourgogne Franche-Comté, LMB Besançon, France - celestin.kokonendji@univ-fcomte.fr

Abstract

In both Tweedie and geometric Tweedie models, the common power parameter $p \notin (0,1)$ works as an automatic distribution selection. It separates two subclasses of semicontinuous $(1 < p < 2)$ and positive continuous $(p \geq 2)$ distributions. Our paper centers around exploring diagnostic tools based on the maximum likelihood ratio test and minimum Kolmogorov-Smirnov distance methods in order to discriminate close distributions within each subclass of these two models according to values of p. Grounded on the unique equality of variation indices, we also discriminate the gamma and geometric gamma distributions with $p = 2$ in Tweedie and geometric Tweedie families, respectively. Probabilities of correct selection for several combinations of dispersion parameters, means and sample sizes are examined by simulations. We thus perform a numerical comparison study to assess the discrimination procedures in these subclasses of two families. Semicontinuous $(1 < p \leq 2)$ distributions in the broad sense are significantly more distinguishable than the over-varied continuous $(p > 2)$ ones; and two datasets for illustration purposes are investigated.
Keywords: Kolmogorov-Smirnov distance, Likelihood ratio test, Probability of correct selection.

1. Introduction

Tweedie and geometric Tweedie models provide flexible parametric families of distributions to deal mainly with non-negative right-skewed data and can handle continuous data with probability mass at zero (Tweedie, 1984; Jørgensen and Kokonendji, 2011). The common power parameter $p \notin (0,1)$, so-called the Tweedie parameter which is one-to-one connected to the common (geometric) stability index $\alpha = (2-p)/(1-p)$, plays an intrinsic role in both models. Indeed, p is an index which distinguishes each distribution within one of each family. See, e.g., Kokonendji et al. (2021).

As preliminaries to a discrimination procedure between two distributions, it is necessary that both distributions have common characteristics such as the supports and shapes of densities. More specifically, for both Tweedie and geometric Tweedie families of distributions, we shall also consider zero-mass and variation indices which are recently introduced by Abid et al. (2020) for non-negative random variable Y. Recall that the zero-mass index is defined through $\mathrm{ZM}(Y) := \mathbb{P}(Y \leq y) \in [0,1]$ for $y \to 0$. Thus, $\mathrm{ZM} \to \varrho$ when $y \to 0$ indicates a ZM or semicontinuous distribution if $\varrho > 0$ and an absolutely continuous one if $\varrho = 0$. As for the variation (or Jørgensen) index expressed by $\mathrm{VI(Y)} = \mathrm{VarY}/(\mathbb{E}\mathrm{Y})^2 \in (0,\infty)$, it is defined in relation to the standard exponential distribution. The VI is viewed as the ratio of the variability of Y to its expected exponential variability which is $(\mathbb{E}Y)^2$. The equi-variation implies no discrepancy between both variabilities. As a matter of fact, Y is said to be over- (equi- and under-varied) compared to exponential with mean $\mathbb{E}Y$ if $\mathrm{VI} > 1$ ($\mathrm{VI} = 1$ and $\mathrm{VI} < 1$), respectively. Scrutinizing both phenomena of ZM and VI, there are very close distributions between and within Tweedie and geometric Tweedie families to be discriminated.

Discriminating between two probability distribution functions was studied by Cox (1961, 1962). There are certain methodologies to measure the closeness between two distribution functions. At this stage, we attemp to challenge a new aspect in statistics in the discrimination between and within close distribution classes of models (between and within subclasses of both Tweedie and geometric Tweedie models) through the use of the maximum likelihood ratio test (LRT) and minimum of Kolmogorov-Smirnov distance (KSD) methods. Sections 2 and 3 display some closeness characteristics of the two interested models with the common case of $p = 2$. Section 3 portrays the proposed discrimination procedures and the estimated probability of correct selection (PCS). Section 4 summerizes some numerical results and applications for illustrative purposes.

2. Main properties of the Tweedie family

In this section, some characteristics of continuous and semicontinuous Tweedie models are exhibited. Let X be a random variable distributed as a Tweedie distribution, denoted $Tw_p(m, \phi)$. Its density function can be indicated by

$$f_{Tw_p}(x; m, \phi) = a_p(x; \phi) \exp[\{x\psi_p(m) - K_p(\psi_p(m))\}/\phi] \mathbb{1}_{\mathbb{S}_p}(x), \tag{1}$$

where $\phi > 0$ is the dispersion parameter, $p \in (-\infty, 0] \cup [1, \infty)$ is the Tweedie index determining the distribution, $\mathbb{S}_p$ is the support of distribution, $a_p(x; \phi)$ is the normalizing function to be discussed below, K_p is the cumulant function, ψ_p is the inverse function of the first derivative K_p' and $m = K_p'(\theta)$ is the mean of X. Note that $K_p'(\cdot)$ defines a diffeomorphism between its canonical domain Θ_p and its image $M_p := K_p'(\Theta_p)$ which is its mean domain. Although the Tweedie densities are not known in a closed form, their cumulant functions are simple. Table 1 exhibits some of the subclasses of Tweedie models.

(Geometric) Tweedie models	$\alpha = \alpha(p)$	p	$\mathbb{S}_p$	M_p
(Geometric) Extreme stable	$1 < \alpha < 2$	$p < 0$	$\mathbb{R}$	$(0, \infty)$
(Asymmetric Laplace/) Gaussian	$\alpha = 2$	$p = 0$	$\mathbb{R}$	$\mathbb{R}$
[Do not exist]	$\alpha > 2$	$0 < p < 1$		
(Geometric) Poisson	$\alpha = -\infty$	$p = 1$	$\mathbb{N}$	$(0, \infty)$
(Geometric) Compound-Poisson-gamma	$\alpha < 0$	$1 < p < 2$	$[0, \infty)$	$(0, \infty)$
(*Geometric*) *Non-central gamma*	$\alpha = -1$	$p = 3/2$	$[0, \infty)$	$(0, \infty)$
(Geometric) Gamma	$\alpha = 0$	$p = 2$	$(0, \infty)$	$(0, \infty)$
(Geometric Mittag-Leffler/) Positive stable	$0 < \alpha < 1$	$p > 2$	$(0, \infty)$	$(0, \infty)$
(*Geometric*) *Inverse Gaussian*	$\alpha = 1/2$	$p = 3$	$(0, \infty)$	$(0, \infty)$

Table 1: Summary of Tweedie and geometric Tweedie including their common stability index $\alpha = \alpha(p)$, power p, support $\mathbb{S}_p$ of distributions and mean domain M_p.

Given the expectation m of $X \sim Tw_p(m, \phi)$, its variance is well-known to be ϕm^p. Thus, one has

$$\mathrm{VI}(Tw_p) = \phi m^{p-2} \quad \left(\gtreqless 1 \ \Leftrightarrow \ \phi \gtreqless m^{2-p} \right). \tag{2}$$

Following similar investigations of Abid et al. (2020, Section 4.2 and Figure 1), the dominant behaviors of $\mathrm{VI}(Tw)$ in (2) appear to be over-variations for all $p \notin (0, 1]$ and an equi-variation for $p = 2$. This index is new for Tweedie models. The special case of $\mathrm{VI}(Y) = \phi$ in (2) for the gamma ($p = 2$) distribution does not depend on the mean m.

3. Background of the geometric Tweedie family

Now, we are essentially interested in the continuous and semicontinuous geometric Tweedie models arising from geometric sums of Tweedie variables. Let $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ be the geometric Tweedie variable with power $p \notin (0, 1)$, dispersion $\widetilde{\phi} > 0$ and mean $\widetilde{m}$ parameters. Therefore, one has the following representation:

$$Z = \sum_{j=1}^{G} T_j,$$

where $T_1, T_2, \dots$ are independent and identically distributed (i.i.d.) as a Tweedie distribution $Tw_p(m, \phi)$ and G is a geometric random variable, independent of T_j, with probability mass function $\mathbb{P}(G = g) = q(1 - q)^{g-1}$, for $g = 1, 2, \dots$ and $q \in (0, 1)$. Moreover, the geometric Tweedie family collapses to exponential mixture representation (see, e.g., Abid et al., 2019b, Proposition 2.1) and it is, therefore, expressed by the following hierarchical formulation

$$X \sim \text{Exponential}(1) \quad \text{and} \quad Z|(X = x) \quad \sim Tw_p(x\widetilde{m}, x^{1-p}\widetilde{\phi}).$$

The density function of $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ is deduced from (1) by

$$f_{GTw_p}(z; \widetilde{m}, \widetilde{\phi}, p) = \int_0^{\infty} \exp(-x) f_{Tw_p}(z; x\widetilde{m}, x^{1-p}\widetilde{\phi}) dx, \tag{3}$$

which is however not analytically tractable, apart from special cases corresponding to $p \in \{0, 1, 2, 3\}$. Yet, numerical methods allow the density (3) to be accurately and fast evaluated by simulation.
From the characteristic variance $\widetilde{m}^2 + \widetilde{\phi}\widetilde{m}^p$ of $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$, the variation index is expressed by

$$\mathrm{VI}(GTw_p) = 1 + \widetilde{\phi}\widetilde{m}^{p-2} \quad \left(\gtreqless 1 \;\Leftrightarrow\; \widetilde{\phi} \gtreqless 0 \right). \tag{4}$$

It is noteworthy that just like Tweedie models with $p = 2$ in (2), the Jørgensen (or variation) index $\mathrm{VI}(GTw)$ in (4) for the particular case $p = 2$, corresponding to the geometric gamma distribution, is equal to $1 + \widetilde{\phi}$ and not depending on the mean $\widetilde{m}$. For $p = 2$ and given any $\widetilde{m} = m > 0$, both variation indexes for Tweedie (2) and geometric Tweedie (4) models coincide when their dispersion parameters differ by +1 in the sense of geometric Tweedie. More conventionally, one can write $Tw_2(m, \phi) \approx GTw_2(m, 1 + \phi)$ for $\phi \geq 1$ and $m > 0$.

4. Discrimination procedure

In this section, two techniques are firstly considered involving the maximum LRT and minimum KSD as optimality criteria to diagnose the appropriate fitting model among two given distributions for a dataset. The goal is to compare how the PCSs work for different situations.
Assume that we observe a random sample $Y_1, Y_2, \ldots, Y_n$ that is supposed to belong to one of the parent distributions $f_p(y; m, \phi)$. For fixed $p > 1$, the maximum-likelihood of the mean m and dispersion parameter ϕ are given, respectively, by

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \widehat{\phi} = \arg\max_{\phi > 0} L_p(\widehat{m}, \phi),$$

where $L_p(\widehat{m}, \phi)$ is the profile likelihood function calculated at $\widehat{m}$. The likelihood ratio statistic, also known as the Cox statistic (1961), is defined by

$$LT_{p_j, p_{j'}} = \log\left(\frac{L_{p_j}(\widehat{m}_j, \widehat{\phi}_j)}{L_{p_{j'}}(\widehat{m}_{j'}, \widehat{\phi}_{j'})} \right), \tag{5}$$

The decision rule for discriminating between two distributions having densities f_{p_j} and $f_{p_{j'}}$ refers basically to choosing f_{p_j} if $LT_{p_j, p_{j'}} > 0$, and to rejecting f_{p_j} in favor of $f_{p_{j'}}$ otherwise. Notice that, in contrast to the LRT, the KSD test may consider more than two competitive distributions to describe data. The KSD is defined by

$$KS_{p_j} = \sup_{-\infty < y < \infty} |\widehat{F}_{p_j}(y; \widehat{m}_j, \widehat{\phi}_j) - \widetilde{F}(y)|, \quad j \in \{1, \ldots, \ell\}, \tag{6}$$

with $\ell \geq 2$, $\widehat{F}_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ the distribution function of $f_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ and $\widetilde{F}(\cdot)$ the empirical distribution function calculated directly from data. The model index j_0 with the minimum distance is, therefore, selected as the winning model:

$$j_0 = \arg\min_{j \in \{1, \ldots, \ell\}} KS_{p_j}.$$

The performance of the maximum LRT and minimum KSD methods is investigated by the PCSs based on simulations. In practice, we generate $(Y_n^{(1)}, \ldots, Y_n^{(N)})$, where $Y_n^{(k)}$ are k-random samples of size n that is supposed to belong to f_p. We repeat both procedures, LRT and KSD, for each $Y_n^{(k)}$, $k = 1, \ldots, N$. The PCS, which corresponds to the proportion of times f_p, is chosen as the winner model and can be evaluated by:

$$\widehat{PCS}_p = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}\{Y_n^{(k)} \text{ is correctly classified}\}. \tag{7}$$

4. Simulation studies and applications

Relying upon their similarities based on the variation indices, and resting on supports and shapes, we first discriminate between the gamma $Tw_2(m, \phi)$ and geometric gamma $GTw_2(\widetilde{m}, \widetilde{\phi})$ distributions verifying $\phi = 1 + \widetilde{\phi}$. Next, we distcriminate between the parent distribution Tw_p and the alternative distributions are $Tw_{p+\varepsilon}$, with $\varepsilon > 0$ such that Tw_p and $Tw_{p+\varepsilon}$ have the same type (see Table 1). This part aims to detect the evolution of the discrimination between distributions for each type: $1 < p < 2$ and $p > 2$. Finally, we distcriminate between the parent distribution GTw_p and the alternative distributions are $GTw_{p+\varepsilon}$, with $\varepsilon > 0$ such that GTw_p and $GTw_{p+\varepsilon}$ have the same type.

Two real datasets are analyzed for illustrative purposes. Concerning the first dataset, gamma Tw_2 and geometric gamma GTw_2 distributions are compared. As for the second one, both semicontinuous ($1 < p < 2$) subclasses of Tweedie and geometric Tweedie are considered through suggesting different values of the power parameter p to fit both models.

4.1 Failure times of the air conditioning system Data consist of the failure times of the air conditioning system of an airplane (Linhart and Zucchini, 1986). The maximum likelihood estimates of the parameters of $Tw_2(m, \phi)$ and $GTw_2(\widetilde{m}, \widetilde{\phi})$ distributions are calculated as $\widehat{m} = 59.60$, $\widehat{\phi} = 1.2317$, $\widehat{\widetilde{m}} = 59.60$ and $\widehat{\widetilde{\phi}} = 0.2380$. It is noteworthy that, $\widehat{\widetilde{\phi}} \simeq 1 - \widehat{\phi}$ as expected.

4.2 Time to failure of pumps The second dataset concerns the time to failure of sixty-one cam-driven reciprocating pumps. A significant presence of zeros ($\widehat{\mathrm{ZM}} = 0.1148$) guides us to discriminate among the semicontinuous ($1 < p < 2$) subclasses of both Tweedie and geometric Tweedie families.

Models	$\left(\widehat{\widetilde{\phi}}\right)\widehat{\phi}$	Log-lik	KSD
$(G)Tw_{1.1}$	(1.8000) 5.7238	(−244.7528) −266.0258	(0.0497) 0.1609
$(G)Tw_{1.2}$	(1.5300) 6.1105	(−245.7920) −250.6113	(0.0449) 0.1079
$(G)Tw_{1.3}$	(1.2200) 5.4724	(−250.2459) −246.3995	(0.0449) 0.0791
$(G)Tw_{1.4}$	(2.1000) 4.6439	(−250.5964) −245.9001	(0.0806) 0.0605
$(G)Tw_{1.5}$	(1.2800) 3.8792	(−251.0668) −247.2682	(0.0742) 0.0469
$(G)Tw_{1.6}$	(0.9300) 3.2701	(−252.1867) −250.1159	(0.0672) 0.0379
$(G)Tw_{1.7}$	(0.8600) 2.8560	(−253.9838) −254.8261	(0.0964) 0.0493
$(G)Tw_{1.8}$	(0.6300) 2.7051	(−256.0918) −262.9210	(0.0820) 0.0946
$(G)Tw_{1.9}$	(0.5600) 3.1977	(−260.2073) −280.2501	(0.1076) 0.2092

Table 2: Estimated dispersion parameters, along with the log-likelihood values (Log-lik) and KSDs for both alternatives Tw_p and GTw_p models with $1 < p < 2$. The numbers in the parenthesis represent the results from GTw_p models.

References

Abid R., Kokonendji C.C. & Masmoudi A (2020). Geometric-Tweedie regression models for continuous and semicontinuous data with variation phenomenon. *AStA Advances in Statistical Analysis* **104**, 33-58

Cox D.R. Tests of separate families of hypotheses (1961). Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, Berkeley, University of California Press. 105-123

Cox D.R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society: Series B* **24**, 406-424

Jørgensen B., Kokonendji C.C (2011). Dispersion models for geometric sums. *Brazilian Journal of Probability and Statistics* **25**, 263-293

Kokonendji CC, Bonat WH, Abid R (2021). Tweedie regression models and its geometric sums for (semi-)continuous data. *WIREs Computational Statistics* **13**, e1496.

Tweedie MCK (1984). An index which distinguishes between some important exponential families. In Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference (J. K. Ghosh and J. Roy, eds.), Indian Statistical Institute, Calcutta. 579-604.