The Tukey g -and- h distribution

The Tukey *g*-and-*h* distribution

Yuan Yan *and* ***Marc G. Genton*** *explain this boosted log-normal distribution, which can be used to model wind speed data and stock market returns, among other things*

What is the Tukey *g*-and-*h* distribution?

Gloria Harvey wakes up around 6am. She listens to the radio while getting ready for work. The weather forecast suggests a mild temperature with a moderate breeze; nothing to be concerned about. Over breakfast, she scrolls through her smartphone, checking the performance of her stock portfolio. Her shares are up a small amount; nothing remarkable. It looks to be a thoroughly average day, and this is how life goes – most of the time.

On other days, Gloria might wake to find the wind blowing a gale or, while browsing her stocks, she might discover that a financial storm has wiped out weeks of modest gains and her shares are now worth less than they were when she bought them. Unpleasant though these experiences are, they are not too far from the ordinary.

Wind speed and stock returns are examples of phenomena whose data exhibit skewness and heavy tails when modelled as a probability distribution. We can see examples of these in Figure 1a and 1b, showing (in red) a right- and left-skewed distribution, respectively, against a normal (or Gaussian) distribution (black dashed line) for comparison. Wind speed data are usually right skewed, meaning that speeds tend to cluster around a range of low to medium values, but with a heavy tail of higher speeds occurring with less frequency. Log-returns of the stock market are the opposite – they are left, or "negatively", skewed due to people's overreactions to bad financial news; extreme losses exist in financial data when computing value-at-risk.

In settings such as these, the Tukey *g*-and-*h* (TGH) family of parametric distributions can accommodate these non-Gaussian features to better model the data. As the family name suggests, there are two parameters involved in this probabilistic model: a real number, g, and a non-negative real number, $h \geq 0$, that control the skewness and tail-heaviness of the distribution, respectively. Location and scale parameters can be added (see box).

What does it look like?

Figure 1 depicts the probability density function of the TGH distribution with different values for g and h. The TGH family includes as special cases the Gaussian distribution when $g = h = 0$, the shifted log-normal distribution[1] when $h = 0$, and Pareto-like distributions[2] when $g = 0$.

BOX: The TGH distribution in detail

The random variable T, obtained after transforming a standard normal random variable Z with the monotonic TGH transformation $\tau_{g,h}$, follows a TGH distribution:

$$T = \tau_{g,h}(Z) = \frac{e^{gZ}-1}{g}\, e^{\frac{hZ^2}{2}}, \quad Z \sim N(0,1).$$

Then, $\xi + \omega T$ is the location-scale version with the location parameter ξ, a real number, and the scale parameter $\omega > 0$.

Who discovered it?

John W. Tukey invented the *g*-and-*h* distribution[3] because he was interested in the behavior of tails in data distributions. His main idea was to model quantiles directly rather than modeling the density function. He achieved this goal by transforming a standard normal random variable by the monotonic TGH transformation, which is equivalent to applying the inverse TGH transformation to 'Gaussianize' the data.

The inverse transformation with the skewness parameter g and the tail-heaviness parameter $h \geq 0$ is more flexible than the logarithm or Box-Cox transformations commonly used to 'Gaussianize' data. Martinez and Iglewicz[4] and Hoaglin[5] further studied properties of this family in detail; see also Yan[6] and references therein for a recent account of additional properties.

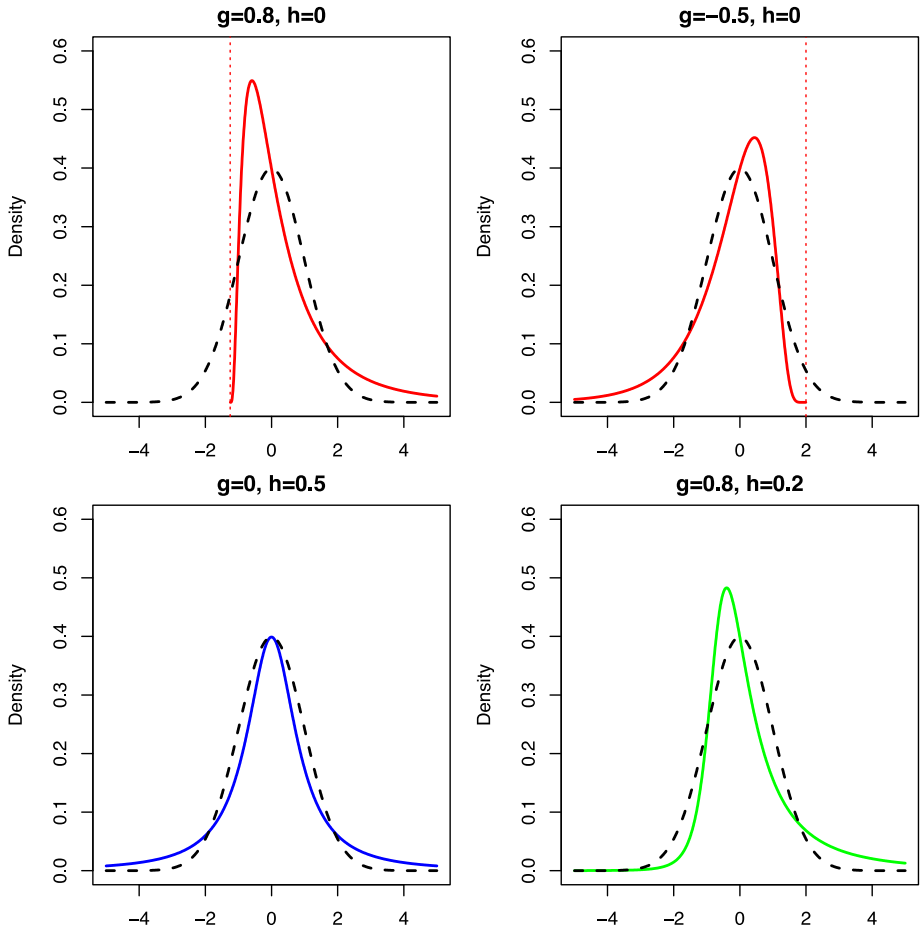

FIGURE 1 Probability density of the Tukey *g*-and-*h* distribution (colored curves) with different values of the parameters *g* and *h*, when $\xi = 0$ and $\omega = 1$. The dotted vertical red lines in the two top plots indicate the lower and upper bound ($-1/g$ when $h = 0$), respectively. The density is symmetric when $g = 0$. The standard normal density ($g = h = 0$) is also shown in each plot by a dashed black curve for comparison.

When should it be used?

The TGH distribution has been applied to data collected in environmental science, economics and finance, to name a few areas of application. The flexibility of the TGH family allows us to model continuous data with different levels of skewness and tail-heaviness.

Parameter estimation from data was originally proposed by matching empirical quantiles with the theoretical quantiles obtained through the TGH transformation of Gaussian quantiles. Although maximum likelihood inference was formerly deemed to be computationally too expensive given that the exact likelihood involves the inverse TGH transformation – which does not have a closed form – Xu and Genton[7] proposed the maximum approximated likelihood method, in which they used the piecewise linearized inverse function instead of the exact one. On current computers, this method is both fast and accurate. If the TGH distribution fits the data well, then the inverse transformed data with the estimated parameters should be approximately Gaussian.[BT1]

Parameter estimation from data can be carried out by quantile matching or maximum approximated likelihood methods as described by Xu and Genton[7]. If the TGH distribution fits the data well, then the inverse transformed data with the estimated parameters should be approximately Gaussian.[BT2]

When should it not be used?

The TGH transformation preserves the unimodality (single peak) of the normal distribution and can make tails only heavier, not lighter. Therefore, if the data's empirical distribution displays multimodality (more than one peak) or light tails, then the TGH family of distributions will not be able to model these features.

Keep in mind . . .

When applying the TGH distribution to spatial and/or temporal data, it is advisable to consider the correlation structure as well as the marginal skewness and heavy-tail features simultaneously for better estimation performance. Finally, if data exhibit non-Gaussian features, then do not ignore them; rather, model them with the Tukey *g*-and-*h* distribution, for instance!

References

1. Limpert, E. and Stahel, W. A. (2017) The log-normal distribution. *Significance*, **14**(1), 8-9.
2. Newman, M. (2017) Power-law distribution. *Significance*, **14**(4), 10-11.

~~Field, C. and Genton, M. G. (2006) The multivariate *g*-and-*h* distribution. *Technometrics*, **48**, 104-111.~~

~~Xu, G. and Genton, M. G. (2016) Tukey max-stable processes for spatial extremes. *Spatial Statistics*, **18**, 431-443.~~

~~Xu, G. and Genton, M. G. (2017) Tukey *g*-and-*h* random fields. *Journal of the American Statistical Association*, **112**, 1236-1249.~~

~~Yan, Y. and Genton, M. G. (2019) Non-Gaussian autoregressive processes with Tukey *g*-and-*h* transformations. *Environmetrics*, **30**, e2503.~~

3. Tukey, J. W. (1977) Modern techniques in data analysis, NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.
4. Martinez, J. and Iglewicz, B. (1984) Some properties of the Tukey g and h family of distributions. *Communications in Statistics: Theory and Methods*, **13**(3), 353-369.
5. Hoaglin, D. C. (1985) Summarizing shape numerically: the *g*-and-*h* distributions. In: *Data Analysis for Tables, Trends, and Shapes*, Hoaglin, D. C., Mosteller, F., Tukey, J. W. (eds.), Wiley: New York, pp. 461-513.
6. Yan, Y. (2018) *Spatio-Temporal Data Analysis by Transformed Gaussian Processes*, Ph.D. Thesis, Statistics Program, King Abdullah University of Science and Technology, Saudi Arabia.
7. Xu, G. and Genton, M. G. (2015) Efficient maximum approximated likelihood inference for Tukey's *g*-and-*h* distribution. *Computational Statistics & Data Analysis*, **91**, 78-91.

Yuan Yan is an environmental statistician. Her Ph.D. dissertation led her to study the use of the Tukey *g*-and-*h* transformation on stochastic processes.

Marc G. Genton is a Distinguished Professor of Statistics at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia.