Outliers in the power exponential model

M. Ibazizen[1*] and M. Mehiri[2]

[1] *Université de Poitiers, Laboratoire de Mathématiques et Applications, France; mohamed.ibazizen@univ-poitiers.fr*
[2] *Université Mouloud Mammeri, Tizi-Ouzou, Algérie; m2mehiri@yahoo.fr*
[*] *Presenting author*

Keywords. *Kurtosis; Outliers; Power Exponential Model.*

Outliers are observations that do not follow the pattern of the majority of the data and show extremeness relative to some basic model.

For the $p-$ variate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ as a model, where $\mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{pxp}$, Σ positive definite, Becker and Gather (1997) gave the general concept of an α outlier : an α outlier with respect to $\mathcal{N}_p(\mu, \Sigma)$ is defined as an element of the set

$$out(\alpha, \mu, \Sigma) = \{x \in \mathbb{R}^p : (x-\mu)'\Sigma^{-1}(x-\mu) > \chi^2_{p,1-\alpha}\} \tag{1}$$

for $\alpha \in]0,1[$, with $\chi^2_{p,1-\alpha}$ denoting the $(1-\alpha)$ quantile of the χ^2_p distribution. We can write

$$P[X \in out(\alpha, \mu, \Sigma)] = \alpha \quad \text{for} \quad X \sim \mathcal{N}_p(\mu, \Sigma).$$

For usual choices of α ($\alpha = 0.05$, $\alpha = 0.10$), this reflects the idea of an outlier being an observation that is rather unlikely under the assumed model and also situated "outside the main mass of the distribution".
Our objective is to extend this approach for a more general distribution having higher or lower tails than those of the normal distribution, namely "the Multivariate Power Exponential distribution". The definition and properties of this distribution are given in Gömez & al. (1998).
A random vector $X = (X_1, ..., X_p)'$, with $p \geq 1$, has a p-dimensional power exponential distribution, with μ, Σ and β parameters ($\beta > 0$) if its density function is

$$f(x; \mu, \Sigma, \beta) = k|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{\beta}\right\}, \tag{2}$$

with $k = \dfrac{p\Gamma(\frac{p}{2})}{\pi^{\frac{p}{2}}\Gamma(1+\frac{p}{2\beta})2^{1+\frac{p}{2\beta}}}$.

This distribution, denoted by $PE_p(\mu, \Sigma, \beta)$, is a member of the family of elliptically symmetric distributions. Its mean, covariance matrix and kurtosis are :

$$E(X) = \mu, \quad Var(X) = \frac{2^{\frac{1}{\beta}}\Gamma(\frac{p+2}{2\beta})}{p\Gamma(\frac{p}{2\beta})}\Sigma, \quad \gamma_2(X) = \frac{p^2\Gamma(\frac{p}{2\beta})\Gamma(\frac{p+4}{2\beta})}{\Gamma^2(\frac{p+2}{2\beta})} - p(p+2) \qquad (3)$$

The parameter β indicates, in terms of kurtosis, the disparity from (2) to the normal case, which corresponds to $\beta = 1$.

Consider the quadratic form : $R(x, \mu, \Sigma) = (x - \mu)'\Sigma^{-1}(x - \mu)$.
The distribution of the r.v. R^β is $\Gamma(\frac{1}{2}, \frac{p}{2\beta})$ (see Gömez & al. (1998)).
By analogy with the multivariate normal case, let us define the set

$$out(\alpha, \mu, \Sigma, \beta) = \left\{ x \in \mathbb{R}^p : R(x, \mu, \Sigma) > \left[\gamma_{(\frac{1}{2}, \frac{p}{2\beta})}(1-\alpha)\right]^{1/\beta} \right\}$$

Likewise, $\gamma_{(\frac{1}{2}, \frac{p}{2\beta})}(1-\alpha)$ is here the $(1-\alpha)$ quantile of the Gamma distribution with parameters $\frac{1}{2}$ and $\frac{p}{2\beta}$.
An α outlier with respect to the power exponential distribution $PE_p(\mu, \Sigma, \beta)$ is then an element of the set $out(\alpha, \mu, \Sigma, \beta)$.

Recall that the parameter β in the $PE_p(\mu, \Sigma, \beta)$ model is related to the kurtosis which reflects the tails of the distribution.

In this note, we perform a numerical study to confirm that an observation may be an outlier for a given model but not be for yet another neighbor of the first model, such that the same identification rule led to different conclusions in the two models considered to be relatively close. The most interesting fact is that the number of outliers depends heavily on the model assumption.

References

Barnett, T., Lewis, T. (1994). Outliers in Statistical Data. J. Wiley & Sons, New York.

Becker, C., Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *JASA*, **94(447)**, 947–55.

Gömez, E., Gömez-Villegas, M.A., Marïn, J.M. (1998). A Multivariate Generalization of the Power Exponential Family of Distributions. *Communications in Statistics-Theory and Methods* **27(3)**, 589–600.

Lopuhaä, H. & Rousseeuw, P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, **19**, 229–248.

Maronna, R., Martin, D. & Yohai, V. (2006). Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester.