International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2018): 5.426

Compound Exponential Power Distribution and its Application to Microarray data

Bindu Punathumparambath

Assistant Professor, Department of Statistics, Govt. Arts and Science College, Calicut, Kerala, India

Abstract: ***The compound distributions have applications in many fields including insurance. The present paper introduces a new family of Exponential power distribution called compound Exponential power distribution, which is obtained by compounding the exponential power distribution with the gamma distribution. This family of distribution includes the Laplace and normal distributions as special cases. We discuss various properties of this distribution. The maximum likelihood estimation procedure is employed to estimate the parameters of the proposed distribution and an algorithm in R package is developed to carry out the estimation. Simulation studies for various choices of parameter values are performed to validate the algorithm. Finally, we illustrate the application using microarray gene expression data.***

Keywords: Compound exponential power distribution, exponential power distribution, gene expression, microarray

1. Introduction

The microarray technology introduced in 1990's is a powerful tool for simultaneous study of the expression levels of thousands of genes. After normalization, gene expression distribution generally have heavier tails than Gaussian distribution and have asymmetry of varying degrees with a sharp peak, due to the bulk of the mass at the middle.

The gene expression distribution has been modelled using several densities, Kuznetsov [22] used different classes of skewed probability functions such as Poisson, exponential, logarithmic series and Pareto-like distribution. In Hoyle et al. [10] the error distribution is modelled by two distributions: a log-normal in the bulk of microarray spot intensities and a power law in the tails. In Khondoker et al. [17] the distribution of gene expression is modelled using a Cauchy distribution as a part of a statistical model for estimating gene expression using data from multiple-laser scans. Various authors suggested error distribution for gene expression data, Asymmetric Laplace distribution (Purdom & Holmes [11]), asymmetric type II compound Laplace (Bindu et al.[4]), slash distribution with normal kernel (Bindu [7], asymmetric slash Laplace (Bindu [6]), skew slash t (Bindu [5]), Laplace mixture (Bindu and Kannan [8], slash distribution with Cauchy kernel (Bindu [3]), Double Lomax (Bindu and Sangita [2]), etc. In the present study we introduce the compound exponential power distribution as an error distribution for cDNA microarray gene expression data.

The compound distributions have applications in the study of production/inventory problems, since it provides a flexible description of the stochastic properties of the system. These distributions play importantl role in insurance and other areas of applied probability modeling such as queuing theory. Also compound distributions also found applications in genomic studies, Bindu et al. [4] introduced the asymmetric type II compound Laplace and found that it is suitable for modelling impulsiveness and skewness required for microarray gene expression data. In present paper we introduce generalization of exponential power distribution.

This article is organized as follows. Brief introduction of exponential power distribution (Subottin [18]) is given in section 2. In section 3, the compound exponential power distribution is derived, and various properties explored. In section 4 we describe the maximum likelihood estimation of parameters using the BFGS algorithm of optim function in R (R Core Team [19]). The application of the compound exponential power distribution distribution is illustrated in section 5 and we conclude in section 6.

2. Exponential Power Distribution

2.1 Exponential Power Distribution

The Exponential power (EP) distribution introduced by Subottin [18], with scale parameter $\sigma > 0$, a shape parameter $p > 0$ and a location parameter $\mu \in R$.

The probability density function of exponential power distribution is given by

$$f(x,p,\mu,\sigma) = \frac{1}{2\sigma\Gamma\left(\frac{1}{p}\right)p^{\left(\frac{1}{p}-1\right)}} exp\left(-\frac{1}{p}\left|\frac{x-\mu}{\sigma}\right|^{p}\right), (1)$$

Where $-\infty < x < \infty, p > 0, \sigma > 0, \mu \in \mathbb{R}$.

If we put $\mu = 0$ and $\sigma = 1$ we get the standard exponential power distribution. For p < 2 the distribution has heavier tails, smaller values of p corresponds to fatter tails. The EP distribution is leptokurtic for $0 < p < 2$ and platikurtic for $p > 2$. Also for $p = 1$ we get the Laplace density, the normal density for $p = 2$ and the uniform distribution as $p \to \infty$.

Maximum likelihood estimation of the EP distribution are discussed in Agro' [12]. Software Tool to compute the density function, the distribution function and the quantiles and to generate pseudorandom numbers from the EP distribution is developed by Mineo and Ruggieri [1]. Rachev and Mittnik [21] and Nelson [9] found applications of EP models for financial data modeling.

Now we introduce the compound exponential power distribution, a family of exponential power distributions

which is obtained by compounding the exponential power density with the gamma distribution.

3. Compound Exponential Power Distribution

3.1 Compound Exponential Power

The compound exponential power (CPE) distribution, a family of exponential power distributions which is obtained by compounding the exponential power density with the gamma distribution. Now we derive the probability density function of compound exponential power distribution from the exponential power distribution.

Let X follow a exponential power distribution given **s** with density given by

$$f(x,p,\mu|s) = \frac{s}{2\Gamma\left(\frac{1}{p}\right)p^{\left(\frac{1}{p}-1\right)}} exp\left(-\frac{s^p}{p}|(x-\mu|^p\right),$$

and let s follow a Gamma(α,β) distribution with density

$$f(s;\alpha,\beta) = \frac{s^{\alpha-1}e^{-s/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \alpha > 0, \beta > 0, s > 0. \quad (2)$$

Then the unconditional distribution of X is the compound exponential power distribution with parameters (μ, p, α, β), denoted by X ~CEP(μ, p, α, β) and the density function is given by

$$f(x;\mu,p,\alpha,\beta) = \frac{\beta\Gamma\left(\alpha^p+\frac{1}{p}\right)}{2\,\Gamma(\alpha^p)\Gamma\left(\frac{1}{p}\right)p^{\frac{1}{p}-1}}\left[1+\frac{1}{p}|(x-\mu)\beta|\right]^{-\left(\alpha^p+\frac{1}{p}\right)}, (3)$$

$-\infty < x < \infty, p > 0, \alpha > 0, \beta > 0, \mu \epsilon \mathbb{R}.$

For p=1 the density reduce to

$$f(x;\mu,\alpha,\beta) = \frac{\alpha\beta}{2}[1+|(x-\mu)\beta|]^{-(\alpha+1)}.$$

Which is the density of the compound Laplace studied in Bindu et al. 43].

Remark 1: If X has a compound exponential power distribution with parameters α, β, p and μ then $\alpha \to \infty$, $\beta \to \infty$ and αβ= σ the compound exponential power density f(x) converges to the exponential power density.

Figure 1 shows density plots of compound exponential power distribution (for various values of α, β, p). From figure 1 we can see that peakedness of CEP density increases when increasing the values of α. From figure 2 and 3 we can see that peakedness of the pdf decrease when value of p increases.

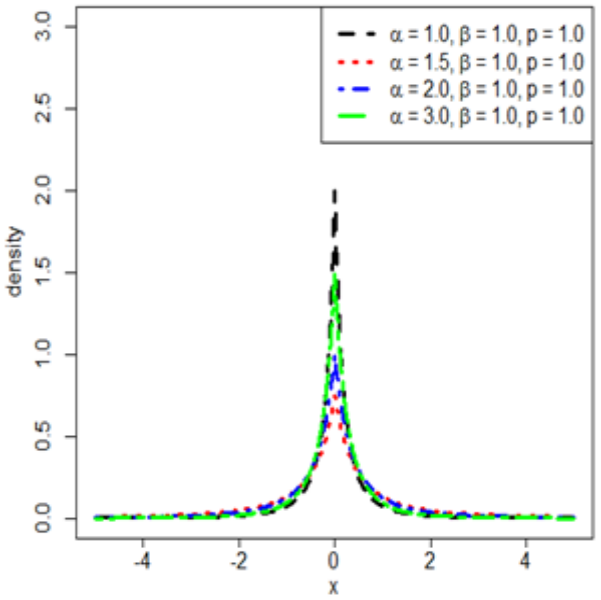

Figure 1: Plot of CEP of various values of α and for μ=0, p=1and β=1.

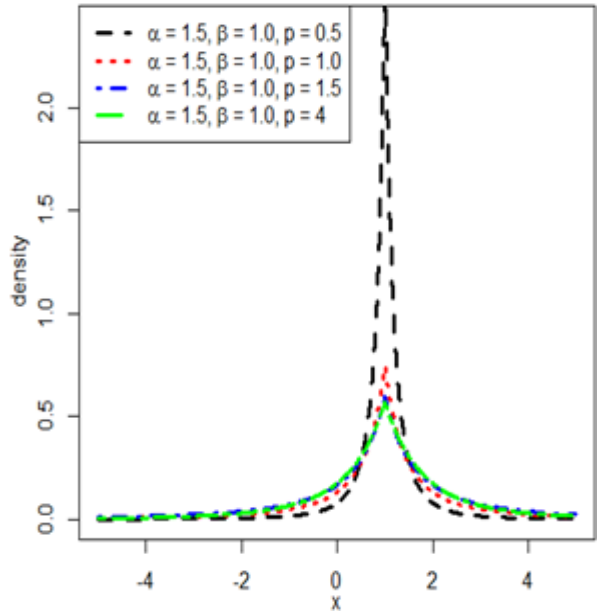

Figure 2: Plot of CEP of various values of p and for μ=1, α=1.5 and β=1.

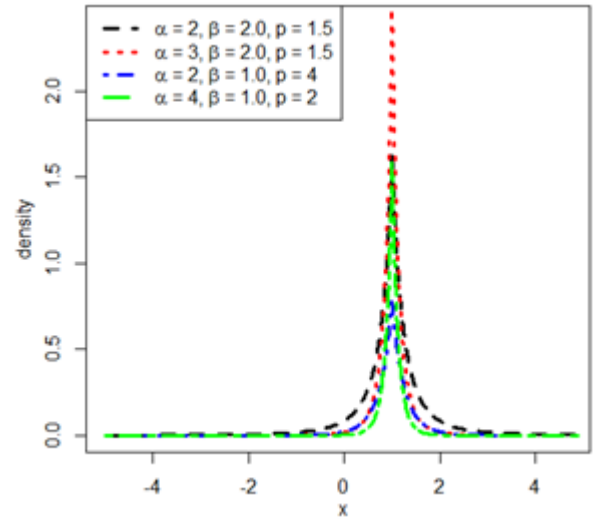

Figure 3: Plot of CEP of various values of p, α, β and for μ=1.

3.2 Properties

The compound exponential power density has the following interesting properties.

P1. The compound exponential power distribution is symmetric and unimodal around the location parameter μ.
P2. The compound exponential power distribution has finite mean if $(\alpha^p+1/p)> 2$ and has finite variance if $(\alpha^p+1/p) > 3$.
P3. The compound exponential power distribution is heavy tailed than exponential power, Laplace and normal distributions. Also more area is concentrated towards the center (mode). Note that the tail probability of the compound exponential power density is, $F(x) \sim cx^{-(\alpha^p p+1/p)}$; *as* $x \to \pm\infty$.
P4. The compound exponential power distribution is completely monotone on (μ, ∞) and absolutely monotonic on $(-\infty, \mu)$. Hence the compound exponential power distribution with μ=0 is completely monotone on $(0, \infty)$ [and absolutely monotone on $(-\infty, 0)$]. As noted by Dreier [15], every symmetric density on $(-\infty, \infty)$, which is completely monotone on $(0, \infty)$, is a scale mixture of Laplace distributions.

4. Estimation of Parameters

In this section we study the problem of estimating four unknown parameters θ= (μ, p, α, β), of compound exponential power distribution. To estimate the parameter μ we use the quantile estimation and maximum likelihood estimation is used to estimate other three parameters.

Let X $=(X_1, X_2, \ldots, X_n)$ be independent and identically distributed samples from compound exponential power distribution with parameters θ. To estimate θ maximum likelihood estimation procedure is used where the likelihood function is maximized to estimate the unknown parameters and is describe below.

The log-likelihood function of the data X takes the form,
Log L(θ, X) = n log β- n log 2- n log B $\left(\alpha^p, \frac{1}{p}\right)$ –n $\left(\frac{1}{p}-1\right)$ logp $-\left(\alpha^p+\frac{1}{p}\right) S(\mu, p, \beta)$,
Where

$$S(\mu, p, \beta) = \sum_{i=1}^{n} \log\left[1+\frac{1}{p}|(x_i-\mu)\beta|\right]$$

Existence, uniqueness and asymptotic normality of maximum likelihood estimators (MLEs) can be derived on the same lines as described in detail for exponential power distribution in Subottin [18], and compound Laplace distribution Bindu et al. [4].

The MLEs of (α, β, p) for given $\mu = \hat{\mu}$ are obtained by solving the score equations for α, β and p.

In our illustrations, the maximization of the likelihood is implemented using the optim function of the R statistical software, applying the BFGS algorithm (R Development Core Team [19]). Estimates of the standard errors were obtained by inverting the numerically differentiated information matrix at the maximum likelihood estimates.

5. Applications

In this section, we present applications of the compound exponential power distribution. We use microarray gene expression dataset from published microarray experiments. The dataset is the cDNA dual dye microarray dataset (Experiment id-51402) downloaded from the Stanford Microarray Database. Each array chip contains approximately 42000 human cDNA elements, representing over 30000 unique genes. These datasets were normalized using (Lowess) locally weighted linear regression method (Cleveland and Delvin, [23]). This method is capable of removing intensity dependence in log2(Ri/Gi) values and it has been successfully applied to microarray data (Yang et al., [24]), where Ri is the red dye intensity and Gi is the green dye intensity for the ith gene.

We use the maximum likelihood estimation method to estimate the parameters. The maximization of the likelihood is implemented using the optim function of the R statistical software, applying the BFGS algorithm (See R Development Core Team, [19]).

The maximum likelihood estimates of the parameters and standard errors (SE) are reported in Table 1. Figure (4) given below, depicts the histogram of the gene expression data and the fitted probability density function evaluated at the MLEs. We compared the empirical distribution function of the microarray gene expression data with the compound Exponential Power (CEP) red line, Exponential Power (EP) blue dotted line and Gaussian (normal) distribution, green dashed line distributions evaluated at the MLEs. It can be clearly seen that the estimated density of the *CEP* fits the data quite well compare to EP and normal distributions.

Table 1 Application - maximum likelihood estimates and their asymptotical standard deviations for CEP, EP and normal

	CEP	*EP*	*Normal*
μ	0.061(0.003)	0.029(0.002)	0.001(0.002)
σ	-	0.941(0.026)	0.081(0.004)
p	1.071(0.009)	0.961(0.061)	-
α	5.554(0.127)	-	-
β	0.421(0.014)	-	-
AIC	85090	89184	100225
BIC	85125	89212	100243

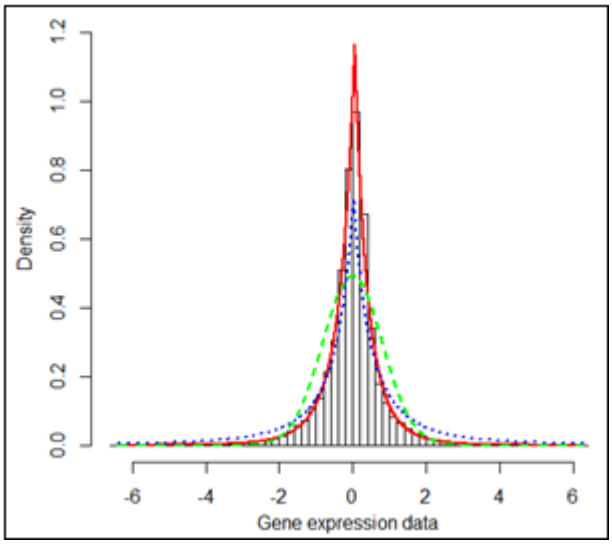

Figure 4: Fitted compound Exponential Power (CEP) red line, Exponential Power (EP) blue dotted line and Gaussian (normal) distribution, green dashed line (evaluated at MLEs) for the microarray data from Experiment id-51402

We used Akaike's Information Criterion (AIC) (Akaike, [14]; Burnham and Anderson, [16]) and Bayesian Information Criterion (BIC) (Schwarz, [13]) to assess the appropriateness of CEP over the EP and normal. The AIC and BIC are given by

AIC= -2 log L + 2 k and BIC= -2 log L + k log n,

Where k is the number of parameters estimated and n is the sample size.

A smaller value of AIC or BIC indicates a better fit. We calculated AIC and BIC for the CEP, EP and normal distributions for the dataset examined. From Table 1we can see that AIC and BIC values are smaller for CEP. Smaller AIC and BIC values indicate better fit and hence CEP fit the data better than EP and normal distributions.

6. Conclusion

In the present paper we introduced a new heavy tailed generalization of exponential power distribution called compound exponential power (CEP) distribution and is useful in analysing datasets that are symmetric, leptokurtic, and deviate considerably from the classical symmetric distributions such as normal, Laplace, etc. These are some of the common features of data in financial modelling and microarray modelling in addition to the heavy tail structure. The CEP introduced in this paper can be useful in analyzing data sets which exhibits heavy tails and peakedness. We found that CEP is suitable for modeling microarray gene expression data, since it is having thick tails and sharp peak in the middle.

7. Acknowledgement

I would like to thank the Department of Science and Technology, Science and Engineering Research Board (SERB), Government of India, New Delhi, for financial assistance under Core Research Grand, Project No. CRG/2018/004468.

References

[1] A. M. Mineo, M. Ruggieri, M., "A Software Tool for the Exponential Power Distribution: The normalp package," Journal of Statistical Software, XII(4), pp. 1-24, 2005.

[2] B. Punathumparambath , K. Sangita, "Double Lomax and its applications," STATISTICA, LXXV(3), pp. 331-342, 2015.

[3] B. Punathumparambath, "A new family of skewed slash distribution generated by Cauchy kernel," Communications in Statistics- Theory and Methods, IVII, pp. 2351-2361, 2013.

[4] B. Punathumparambath, K. Sangita, G.Sebastian," Asymmetric type II compound Laplace distribution and its application to microarray gene expression," Computational Statistics and Data analysis, LVI, pp. 1396- 1404, 2012.

[5] B. Punathumparambath, "The multivariate skew-slash t and skew-slash Cauchy distributions," Model Assisted Statistics and Applications, VIII, pp. 33-40, 2012.

[6] B. Punathumparambath, "The multivariate asymmetric slash Laplace distribution and its applications," STATISTICA, LXII (2), pp. 235-249, 2012.

[7] B. Punathumparambath, "A new family of skewed slash distributions generated by normal kernel," STATISTICA, LXI (3), pp. 345- 353, 2011.

[8] B. Punathumparambath, V. M. Kannan, "Two component mixed Laplace model for microarray gene expression data.," JP Journal of Fundamental and Applied Statistics, II (1), pp. 15-25, 2012.

[9] D. B. Nelson, "Conditional hetroscedasticity in asset returns: A new approach," Econometrica, LIX(2), pp. 347-370, 1991.

[10] D. C. Hoyle, M. Rattray, R. Jupp, A. Brass, "Making sense of microarray data distributions," Bioinformatics XVIII (4), pp. 576–584, 2002.

[11] E. Purdom, S. Holmes, "Error distribution for gene expression data," Statistical Applications in Genetics and Molecular Biology, IV (1), doi:10.2202/1544-6115.1070. Article 16, 2015, Available at: http://www.bepress.com/sagmb/vol4/iss1/art16.

[12] G. Agro', "Maximum Likelihood Estimation for the Exponential Power Function Parameters," Communications in Statistics -Simulation and Computation XXIV, pp. 523-536, 1995.

[13] G. Schwarz , "Estimating the dimension of a model," Annals of Statistics, VI, pp. 461-464, 1978.

[14] H. Akaike, "Information theory and an extension of the maximum likelihood principle", in KOTZ and JOHNSON (eds.), Breakthroughs in Statistics, Vol I. Springer Verlag, New York, pp. 610-624,1973.

[15] I. Dreier, "Inequalities for real characteristic functions and their moments," Ph.D. Dissertation, Technical University of Dresden, Germany, 1999

[16] K. Burnham, D. Anderson, Model selection and Inference, Springer, New York, 1998.

[17] M. R. Khondoker, C. A. Glasbey, B. J. Worton, "Statistical estimation of gene expression using

multiple laser scans of microarrays," Bioinformatics, XXII(2), pp. 215–219, 2006.
[18] M. T. Subottin, "On the law of frequency of error," Mathematicheskii Sbornik, XXXI, pp. 296-300, 1923.
[19] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www. R-project.org/, 2015.
[20] S. Kotz, T.J. Kozubowski, K. Podgorski , The Laplace Distribution and Generalizations: A Revisit with Applications to Communications. Economics, Engineering and Finance. Birkh¨auser, Boston, 2001.
[21] S. Rachev, S. Mittnik, Stable paretian Models in Finance. Wiley, New York, 2000.
[22] V. A. Kuznetsov, "Distribution associated with stochastic processes of gene expression in a single eukaryotic cell," EURASIP Journal of Applied Signal Process. IV, pp. 285–296, 2001.
[23] W.S. Cleveland, S.J. Delvin, "Locally weighted regression: an approach to regression analysis by local fitting," Journal of American Statistical Association, LXXXIII (403), pp. 596–610, 1988.
[24] Yang, Yee Hwa, Dudoit, Sandrine, Luu, Percy, Lin, M. David, Peng, Vivian, Ngai, John, Speed, P. Terence, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," Nuclic Acids Research, XXX(4), e15, 2002.

Author Profile

Bindu Punathumparambath received the Ph. D. in Statistics from Mahatma Gandhi University, Kottayam in 2014 under the Women Scientist Scheme of Department of Science and Technology (DST), Govt. of India. Fellow of society for applied biotechnology, India. Areas of research include Biostatistics, Computational Statistics, Distribution Theory, Reliability, Statistical Genetics, Statistical Inference and Microarray Modelling.