5-1-2013

A Note on α-Curvature of the Manifolds of the Length-biased Lognormal and Gamma Distributions in View of Related Applications in Data Analysis

Makarand V. Ratnaparkhi
Wright State University, makarand.ratnaparkhi@wright.edu

Uttara V. Naik-Nimbalkar
Pune University, Pune, India

Brief Report

A Note on α-Curvature of the Manifolds of the Length-biased Lognormal and Gamma Distributions in View of Related Applications in Data Analysis

Makarand V. Ratnaparkhi
Wright State University,
Dayton, OH

Uttara V. Naik-Nimbalkar
Pune University,
Pune, India

The α-curvature tensors of the statistical manifolds of the length-biased versions of the log-normal and gamma distributions are derived and discussed. This study was designed to investigate observations related to the parameter estimation for the length-biased lognormal distribution as a model for the length-biased data from oil field exploration.

Key words: Statistical manifolds, α-curvature tensor, duality and symmetry, length-biased lognormal distribution, estimation.

Introduction

Data are referred to as length-biased when the probability of inclusion of the observation in the sample is proportional to its magnitude. It has been documented in a number of research articles that the presence of length-biasedness in data arising in fields such as ecology, economics, geology, bioinformatics and medical science. Some examples of such data sets can be found in Ratnaparkhi and Naik-Nimbalkar (2012). The applications of concepts of differential geometry, such as the curvature tensor of the statistical manifolds, have been investigated with reference to statistical inference (Amari, 1985). Recently, Ratnaparkhi and Naik-Nimalkar (2012) observed that certain structural properties of parameter estimators of the length-biased lognormal distribution are significantly different from the corresponding properties of the original lognormal distribution. This article investigates the properties of the manifold of the length-biased lognormal distribution for tracking changes in structural properties, such as the α-curvature of the original manifold, of a usual lognormal distribution.

If $X \sim f(x;\theta)$ where $E[X] < \infty$ and θ is a scalar or a vector of parameters, then, the length-biased version of the pdf $f(x;\theta)$ (referred to as the original distribution) is defined as:

$$g(x;\theta) = xf(x;\theta) / E[X]. \qquad (1)$$

The generalization of (1) (referred to as the weighted distribution) is given by:

$$(x;\theta) = w(x)f(x;\theta) / E[w(X)],$$
$$w(x) > 0, E[w(X)] < \infty. \qquad (1a)$$

The special cases of (1) and (1a) for commonly used statistical distributions such as Gamma, Poisson, etc., are useful in data analysis. For example, the survival or failure time data, the probability of inclusion of the observation is proportional to the magnitude, for example x, of the observation. The resulting data is referred to as length-biased data. For this type of data the pdf (1) is used for modeling the data where

Makarand V. Ratnaparkhi is a Professor of Statistics at Wright State University in Dayton, Ohio, USA. Email him at: makarand.ratnaparkhi@wright.edu. Uttara V. Naik-Nimbalkar is a Professor of Statistics at University of Pune, Pune, India. Email her at: uvnaik@stats.unipune.ac.in.

$f(x;\theta)$ is the model that would have been used if the data were not length-biased. Because, for the estimation of θ (as a parameter of the original distribution), the length-biased version (1) is substituted for the original distribution it is necessary to study the pdf $g(x;\theta)$ for the desired properties of the estimator of θ.

In many situations, the pdf $g(x;\theta)$ behaves as expected; hence the estimation problem is solvable without complexities. However, this may not be the case for every pdf $f(x;\theta)$; therefore, it may be necessary to study the properties of $g(x;\theta)$ in further details. For example, in a recent length-biased data analysis study arising in oil field exploration (Ratnaparkhi & Naik-Nimbalkar, 2012) it was observed that the properties of the estimators of the parameters of the original lognormal distribution - when replaced by its length-biased version - undergo a considerable change. Therefore, a need for further detailed study of the family of length-biased lognormal distributions was identified; in particular, studying the properties of the manifold of the length-biased lognormal distributions is likely to lead to the root cause of the above mentioned somewhat curious situation.

The differential geometric properties, such as the symmetry and duality of α–curvature tensor of a statistical manifold (Amari, 1985), are known to be related to the statistical inference for the parameters of any distribution belonging to the manifold. In particular, in this article the α- curvature tensors for the original lognormal and gamma distributions and their corresponding length-biased versions are derived and their properties are compared in view of related applications in the analysis of length-biased data.

Methodology

A random variable X is said to have a lognormal (LN) distribution with parameters (μ, σ), if its pdf is

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\log x - \mu)^2],$$

where, $x > 0$, $-\infty < \mu < \infty$, $\sigma > 0$.

(2)

Using (1) the pdf of the length-biased lognormal (LBLN) distribution is given by

$$g(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(\log x - (\mu + \sigma^2))^2],$$

where, $x > 0$, $-\infty < \mu < \infty$, $\sigma > 0$.

(3)

The Fisher information matrices for the LN (μ, σ) and LBLN (μ, σ) distributions are provided in Table 1. Note that the information matrices have different structures. These matrices play an important role in deriving the differential geometrical properties of the corresponding manifolds.

Table 1: Fisher Information of LN (μ, σ) and LBLN (μ, σ)

	Fisher Information
LN(μ, σ)	$\begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$
LBLN(μ, σ)	$\begin{bmatrix} 1/\sigma^2 & 2/\sigma \\ 2/\sigma & 4 + 2/\sigma^2 \end{bmatrix}$

The Presence of Curvature Component in the Manifold of LBLN (μ, σ)

Because, the derivation for pdf (3) is straightforward the detailed proof is omitted here for the sake of brevity. Let $U_1 \sim$ LN $(\mu, \sigma/\sqrt{2})$ and $V_1 \sim$ L N $(\sigma^2, \sigma/\sqrt{2})$. If U_1 and V_1 are independent, then, $X = (U_1 \, V_1) \sim$ LBLN (μ, σ) with pdf (3). However, it is notable that the distribution of V_1 belongs to the curved family of distributions, hence, the statistical manifold of the LBLN distributions must be investigated for the nature of its curvature. Note that, in the derivation, using the logarithmic transformation on U_1 and V_1 results in the distribution of $Y = \log(X)$ as a convolution of two normal random variables with respective distributions. Clearly, the manifold of the distributions of Y has the corresponding curvature property.

Statistical Manifolds

A family of probability distributions $\mathcal{F}_\theta = \{ f(x, \underset{\sim}{\theta}) \}$ with p-dimensional parameter vector $\underset{\sim}{\theta}$ having values over some open subset of R^p is a manifold (denoted by $\mathcal{M}$), in which the pdf $f(x; \underset{\sim}{\theta})'s$ are the points and the parameters $\underset{\sim}{\theta}$ provide the necessary coordinate functions (for details, refer to Amari, 1985).

In this case, let $\mathcal{M}_1 = \{ f(x; \mu, \sigma) \}$ denote the manifold of LN (μ, σ) and let $\mathcal{M}_2 = \{ g(x; \mu, \sigma) \}$ denote the manifold of LBLN (μ, σ). The regularity conditions (Amari, 1985) needed for deriving the differential geometric properties of $\mathcal{M}_1$ and $\mathcal{M}_2$ are satisfied and hence the derivation of the α-curvature tensors for $\mathcal{M}_1$ and $\mathcal{M}_2$ can proceed. For these derivations the following are needed:

1. The Fisher information matrix provides a Riemannian metric for the manifold $\mathcal{M}$ of pdfs.

2. The Fisher score functions are the vectors that span the tangent space at a point in $\mathcal{M}$.

3. For application of differential geometric concepts in statistics what are known as the α-connections are necessary. Such

connections for $\mathcal{M} = \{ f(x; \mu, \sigma) \}$ are defined by

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk}^{(1)} + \frac{(1-\alpha)}{2} T_{ijk}$$

where

$$\Gamma_{ijk}^{(1)} = \mathrm{E}[\partial_i \partial_j l(x; \mu, \sigma) \partial_k l(x; \mu, \sigma)]$$

and

$$T_{ijk} = \mathrm{E}[\partial_i l(x; \mu, \sigma) \partial_j l(x; \mu, \sigma) \partial_k l(x; \mu, \sigma)]$$

$$(4)$$

where $l(x; \mu, \sigma)$ is the log-likelihood function. For this study, $l(x; \mu, \sigma)$ is the likelihood function of $f(x; \mu, \sigma)$ or $g(x; \mu, \sigma)$ and the subscripts $i = 1, 2, j = 1, 2$ and $k = 1, 2$ correspond to the derivatives with respect to μ and σ respectively. Thus, there are eight α-connections for each manifold. The α–curvature tensor, using the Einstein summation notation, in terms of the α-connections is given by the formula

$$R_{ijkm}^{(\alpha)} =$$
$$(\partial_i \Gamma_{jk}^{(\alpha)s} - \partial_j \Gamma_{ik}^{(\alpha)s}) h_{sm} + (\Gamma_{irm}^{(\alpha)} \Gamma_{jk}^{(\alpha)r} - \Gamma_{jrm}^{(\alpha)} \Gamma_{ik}^{(\alpha)r}),$$

$$(5)$$

where h_{sm} is the $(s, m)^{th}$ element of the inverse of the Fisher information matrix and using Einstein summation notation:

$$\Gamma_{jk}^{(\alpha)s} = h^{sm} \Gamma_{jkm}^{(\alpha)},$$

where h^{sm} is the $(s, m)^{th}$ element of the inverse of Fisher information matrix. This formula is used for other subscripts and superscripts in (5). The values of the subscripts (i, j, k, m) have the same meaning as in (4), thus, there are sixteen components of α-curvatures for each manifold.

Statistical Applications of α-Curvature

Three properties of a manifold related to the concepts of α-connections and α-curvature relevant to this study are:

1. If all components of $R_{ijkm}^{(\alpha)} = 0$, then the manifold is said to be flat as opposed to curved. The efficiency of the parameter estimators depend on the curvature of the manifold. (See Amari (1985) for a discussion of α-connections and the curvature tensors with reference to the exponential family of distributions.)

2. If the curvature vanishes at α = ± 1 then the α-affine coordinate system exists for α = ± 1. In such cases 1-affine coordinate system is called the natural parameter system and the −1-affine coordinate system is called the expectation parameter system. For example, for the Gaussian manifold it can be shown that such coordinate systems exist. For α=0 the curvature becomes the Riemannian curvature.

3. If R denotes the curvature tensor, and if $R^{\alpha} = R^{-\alpha}$ for all real values of α, then the manifold is said to be conjugate symmetric. This is considered to be one of the desirable properties in statistical applications of differential geometry, for, example, the Gaussian manifold is conjugate symmetric.

Results

The maximum likelihood estimators ($\hat{\mu}, \hat{\sigma}^2$) of the parameters μ and σ^2 of the LBLN distribution and their properties are discussed in Ratnaparkhi and Naik-Nimbalkar (2012). These estimators, for reference, are:

$$\hat{\mu} = (\sum \log x_i / n) - \hat{\sigma}^2$$

$$\hat{\sigma}^2 = \frac{1}{n}[\sum (\log x_i)^2 - (\sum \log x_i)^2 / n]$$

(6)

Table 2 shows that the α-connections for LN and LBLN do not have the same structure and that the α-curvature tensors are expected to be different for these distributions, however, interest lies in the nature of the structure of the difference between the curvatures.

There are total of sixteen components of the α-curvature tensor for each of the manifolds of the LN and LBLN distributions. The derivation of all such components is lengthy and tedious; for the objective of this article - and to draw conclusions - all sixteen components are unnecessary. For brevity, only some of the components are reported and serve the purpose of illustrating the difference between the behaviors of these two manifolds.

The α-curvature of $\mathcal{M}_1$ is given by

$$R_{1212}^{(\alpha)} = (1 - \alpha^2) / \sigma^4,$$

(7)

and the Riemannian curvature, at α = 0 is

$$\kappa = -R_{1212}^{(\alpha)} \cdot h^{11} h^{22} / 2 = -1/2$$

(8)

For $\mathcal{M}_2$ the scalar curvature κ is a function of σ^2 (i.e., not constant as in case of $\mathcal{M}_1$).

The α-Curvature of the Manifolds of Gamma (μ, β) and Length-Biased Gamma (μ, β)

The gamma distribution is sometimes considered an alternate model for the lognormal distribution. Let G (μ, β) and LBG (μ, β) denote the gamma and its length-biased version respectively where parameter μ is the mean of the gamma distribution. The pdf of G (μ, β) distribution is given by

$$f_2(x; \mu, \beta) = (\beta / \mu)^\beta x^{\beta-1} \exp(-x\beta / \mu) / \Gamma(\beta),$$
$$x > 0, \mu > 0, \beta > 0.$$

(9)

The pdf of LBG (μ, β) is given by

$$g_3(x; \mu, \beta) = (\beta / \mu)^{\beta+1} x^{\beta+1-1} \exp(-x\beta / \mu) / \Gamma(\beta+1),$$
$$x > 0, \mu > 0, \beta > 0.$$

(10)

Using formulae (4) and (5), the α-curvature of the manifold of G (μ, β) distributions is given by

$$R_{1212}^{(\alpha)} = (\alpha^2 - 1)[\phi(\beta) + \beta\phi'(\beta)] / 4\mu^2 \beta\phi(\beta)$$

(11)

Table 2: α-Connections $\Gamma_{ijk}^{(\alpha)}$ for the LN (μ, σ) and LBLN (μ, σ) Distributions

$\Gamma_{ijk}^{(\alpha)}$	LN(μ,σ)	LBLN(μ,σ)
$\Gamma_{111}^{(\alpha)}$	0	0
$\Gamma_{112}^{(\alpha)}$	$(1-\alpha)/\sigma^3$	$(1-\alpha)/\sigma^3$
$\Gamma_{121}^{(\alpha)}$	$-(1+\alpha)/\sigma^3$	$-(1+\alpha)/\sigma^3$
$\Gamma_{122}^{(\alpha)}$	0	$-4\alpha/\sigma^2$
$\Gamma_{211}^{(\alpha)}$	$-(1+\alpha)/\sigma^3$	$-(1+\alpha)/\sigma^3$
$\Gamma_{212}^{(\alpha)}$	0	$-4\alpha/\sigma^2$
$\Gamma_{221}^{(\alpha)}$	0	$-2(1+2\alpha)/\sigma^2$
$\Gamma_{222}^{(\alpha)}$	$-2(1+2\alpha)/\sigma^3$	$-2(1+2\alpha+6\alpha\sigma^2)/\sigma^3$

The α-curvature of the manifold of LBG (μ, β) is of the same form as (10) where β is replaced by (β+1).

Conclusion

Previous research investigated estimation of the parameters (μ, σ) of the original lognormal distribution (LN) using the length-biased lognormal (LBLN) distribution based on the known length-biased nature of data arising from oil field exploration (Ratnaparkhi & Naik-Nimbalkar, 2012). In this estimation process, the MLE ($\hat{\mu}, \hat{\sigma}^2$) shown in (6) was obtained. Examination of (6) shows that the properties of these estimators are different from the properties of the corresponding estimators of the original lognormal distribution, thus, to better understand the structure of LBLN distribution the differential geometric properties of the manifolds of the LN and LBLN distributions were studied. In addition, this research was undertaken because the respective pdf's (1) and (2) of the LN and LBLN distributions are considered to be of the same form. Although pdf's (1) and (2) appear similar it was thought that there must be some other criterion for referring to these two pdf's as of the same form. Finally, for comparison purposes the differential geometric properties of the gamma and length-biased distributions were derived.

Based on the above results the following conclusions are put forth:

1. It is known that all information about the curvature tensor of the manifold of the LN distribution can be extracted from the single component $R_{1212}^{(\alpha)}$ given in equations (7) and (8). This component vanishes at $\alpha = \pm 1$, hence the manifold has the three properties previously noted. Equation (8) shows that the scalar curvature and in particular the Riemannian curvature for $\mathcal{M}_1$ is constant. This property does not hold for $\mathcal{M}_2$, the manifold of LBLN distributions. Thus, the

curvature is not constant and in particular is a function of σ^2. The properties of the estimators the parameters are known to be associated with the curvature of the manifold (Amari, 1985). Therefore, if the LBLN is the model that is used for the estimation of the parameters of the original LN distribution the properties of the estimators will be distorted non-uniformly depending on σ. In fact, the manifold of the LBLN distributions is a unique example that we have identified in real data analysis. The systematic theoretical analysis of the LBLN manifold is needed. The details are beyond the scope of this brief report and hence are not included here. However, this observation indicates that due caution is necessary if the LBLN distribution is to be considered as a model for length-biased data.

2. The gamma distribution can be used in certain data analysis problems in lieu of the lognormal distribution investigated herein, the differential geometrical properties of the gamma manifold. Equation (11) verifies that $R_{1212}^{(\alpha)}$ vanishes at α = ± 1 for the manifolds of the gamma and the length-biased version

of gamma distribution and, therefore, the three properties presented herein hold equally for both manifolds. Hence, the parameter estimators of the original gamma distribution based on its length-biased version should not pose problems either locally or globally.

Thus it may be concluded that, in data analysis, a researcher should either use the LBLN distribution with due caution considering the unusual behavior of the LBLN manifold or, whenever possible, consider the LBG distribution as an alternative to the LBLN distribution.

References

Amari, S. I. (1985). *Differential-geometrical methods in statistics*. New York: Springer-Verlag.

Ratnaparkhi, M. V., & Naik-Nimbalkar, U. V. (2012). Length-biased lognormal distribution and its application in the analysis of data from the oil field exploration studies. *Journal of Modern Applied Statistical Methods*, *11*(*1*), 255-260.