A multivariate extension to the Exponentially-modified Gaussian distribution

Anonymous Author(s)
Affiliation
Address
email

Abstract

The exponentially-modified Gaussian (EMG) distribution is a convolution sum of a univariate Gaussian and an exponential distribution. This has been used to model univariate skewed data such as chromatographic peaks' shape, cell population dynamics from single-cell data and reaction times in neuropsychology. Currently, the EMG is only available in its univariate form. In this work, we propose a multivariate extension to the EMG, called *mvEMG*, by using an affine transformation involving rotation, translation and shearing to accommodate for the three moments (mean, variance and skew). We derive statistical properties for mvEMG. Although we demonstrate its performance in synthetic data compared with the multivariate skew normal distribution, we are unable to show its practical applicability, mainly due to lack of efficient sampling strategies and a viable real-world dataset.

1 Introduction and motivation

Generally, datasets are described using statistics involving mean and variance, which through the method of moments, represent the first and second order estimators. Most data-driven approaches assume the data to be generated using a Gaussian distribution where these estimators are used to address simple hypotheses about the datasets. To gain deeper insights into the data, one needs to consider higher-order estimators, such as skewness, and such estimators are rarely used.

In many real world applications, datasets do not generally follow a Gaussian distribution but are rather highly skewed with complex structures. For computational feasibility, one way to handle such data is to log normalise the data rendering it as Gaussian, but by doing so, the inherent skewness is lost. Since the tails of a probability distribution reflect the most extreme data points, it seems plausible that measures based on skewness would be useful for identifying rare events or entities such as transcriptional regulators in gene expression data or highly-variable genes that could be targeted in precision medicine studies. As a specific example, we look at multidimensional single-cell RNA-seq data where multiple genes are profiled per cell. The distribution of gene counts per cell, called library size, is highly right-skewed (Figure 1a upper panel) and log normalising renders the distribution to resemble a Gaussian distribution (Figure 1a lower panel), where the information regarding rare events triggering data skewness is lost. Skewness is important for modelling complex phenomena.

There exists the exponentially-modified Gaussian (EMG) distribution, to handle univariate skewed data. In this work, we present a multivariate extension to the EMG to cater to multidimensional skewed data whilst providing access to the first three distributional moments viz. mean, covariance *and* skew. For this, we make use of the multivariate Gaussian that provides the mean and covariance, and the multivariate exponential distribution that gives the skewness of the data. The multivariate exponential distribution has many forms that can be derived from the univariate exponential distribution (*Esary and Marshall (1974)*). This work aims to provide one way to unify these two useful multivariate distributions into a single viable multivariate form leading to the multivariate EMG (mvEMG).

Under review at ICBINB Workshop @ NeurIPS 2021. Do not distribute.

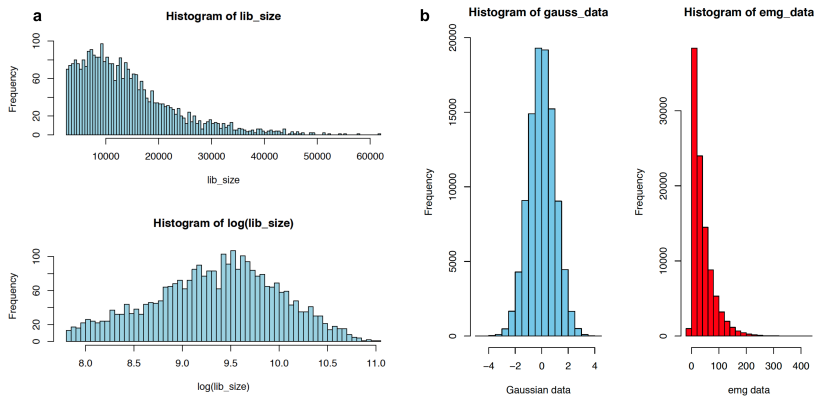

Figure 1: **a: Upper**: Histogram showing the original skewed library-size distribution i.e. the distribution of genes across cells for the *Zeisel, et al. (2015)* single-cell RNAseq dataset for 3500 cells and **a: Lower**: the same library size as above but is log-normalised depicting the shape of a Gaussian distribution. **b: Left**: Histograms showing 10K random variables drawn from a standard Gaussian distribution and **b: Right**: from a univariate EMG $(mean = 0, std = 1, skew = 0.25)$.

2 Univariate EMG

The exponentially-modified Gaussian (EMG) probability distribution is the convolution of a univariate Gaussian distribution and an exponential distribution which are independent of each other. Assume a r.v. $Z = Z_1 + Z_2$ with Z_1 and Z_2 as independent r.vs where $Z_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $Z_2 \sim \exp(\lambda)$ then Z is a convolution of a Gaussian and an Exponential random variable and is said to follow an exponentially-modified Gaussian (EMG) distribution with parameters (μ, σ, λ) i.e. $Z \sim \text{EMG}(\mu, \sigma, \lambda)$ and is given by: $f_Z(z) = \dfrac{\lambda}{2} \exp(\dfrac{\lambda(2\mu + \lambda\sigma^2 - 2z)}{2})\ erfc(\dfrac{\mu + \lambda\sigma^2 - z}{\sqrt(2)\sigma})$ where *erfc* is the complementary error function and $erfc(x) = 1 - erf(x) = 1 - \dfrac{2}{\sqrt(\pi)} \int_x^\infty e^{-t^2} dt$

Figure 1b depicts the histograms of r.vs distributed according to a standard normal (left panel) and according to a univariate EMG with mean=0, std=1 and skew = 0.25 (right panel). One of the key properties of the univariate EMG distribution is its differential behavior in the right and left tails. The distribution exhibits a Gaussian-distributed left tail and an exponentially-distributed right tail. Hence, it is particularly well suited to fit empirical distributions that are right-skewed i.e. exhibit thicker right tails. This distribution has found practical applications in a variety of scientific disciplines such as chromatography (*Naish and Hartwell, 1988; Kong et al., 2005*), cellular biology (*Golubev, 2010; Tyson et al., 2012*) and microarray preprocessing (*Irizzary et al., 2003)*.

The univariate skew normal (SN) distribution is also used to model right-skewed data. Since it has a normal-like right tail, it is thinner as compared to the EMG. Thus the EMG is better suited in modelling fat-tailed phenomenon consisting of outliers or data points residing in the right-tail. Additionally, the SN parameters are mean, scale and shape whereas for the EMG, the parameters are mean, variance and skew. Therefore, to access skewness in fat right-tailed distributions, we pursue building a multivariate extension to the univariate EMG.

3 Derivation of the mvEMG PDF

For each $Z_i \sim EMG(\mu_i, \sigma_i, \lambda_i)$, we have mean $\mathbb{E}(Z_i) = \mu_{Z_i} = \mu_i + \dfrac{1}{\lambda_i}$, var($Z_i$) = $\sigma^2_{Z_i} = \sigma_i^2 + \dfrac{1}{\lambda_i^2}$, and skew($Z_i$) = $\lambda_{Z_i} = \dfrac{2}{\sigma_i^3 \lambda_i^3}\left(1 + \dfrac{1}{\sigma_i^2 \lambda_i^2}\right)^{-\frac{3}{2}}$ where μ_{Z_i}, $\sigma^2_{Z_i}$ and $\lambda_{Z_i} \in \mathbb{R}$. Next we assume a random vector **Z** such that for $\boldsymbol{Z} \in \mathbb{R}^d$, we have mean $\mathbb{E}(\boldsymbol{Z}) = \mu_Z = [\mu_{Z_1}, \cdots, \mu_{Z_d}]$, Cov($\boldsymbol{Z}$) = Σ_Z

= $diag(\sigma^2_{Z_i}) = diag(\sigma_i^2 + \frac{1}{\lambda_i^2})$, and skew($\boldsymbol{Z}$) = $\lambda_Z = diag(\lambda_{Z_i}) = diag(\frac{2}{\sigma_i^3\lambda_i^3}\Big(1 + \frac{1}{\sigma_i^2\lambda_i^2}\Big)^{-\frac{3}{2}})$ where $\mu_Z \in \mathbb{R}^d$, Σ_Z and $\lambda_Z \in \mathbb{R}^{d\times d}$.

Affine transformation of a d independent 0-centered univariate EMG random variable Z. Given an affine transformation $\mathbb{L}$ that comprises of (a) the Euclidean transformations of rotation via matrix A and translation via vector b followed by (b) a shear transformation via matrix S where $A, S \in \mathbb{R}^{d\times d}$, both A and S are invertible and $b \in \mathbb{R}^d$. We define the shear matrix S as $S = S^* + \mathbb{I}_{d\times d}$ where S^* is skew-symmetric i.e. $S^*(ij) = \lambda_{Z_i}$, $S^*(ij) = S^*(-ji)$, $diag(S^*) = 0$ and $diag(S) = 1$. Applying $\mathbb{L}$ to $\boldsymbol{Z}$ such that $Z' = \mathbb{L}(\boldsymbol{Z})$ results in the image of $S(A\boldsymbol{Z} + b) := \mathbb{L}(\boldsymbol{Z}) = Z'$. We note here that $\mathbb{L}(\boldsymbol{Z})$ still linearly transforms $\boldsymbol{Z}$.

Moments of Z' The moments of this transformed random variable Z' are:

1. $\mathbb{E}(Z') = S(\mathbb{E}(A\mu_Z) + \mathbb{E}(b)) = S(A\mu_Z + b)$
2. $Cov(Z') = \mathbb{E}[(Z' - \bar{Z}')(Z' - \bar{Z}')^T]$
   $= \mathbb{E}[(SA(\boldsymbol{Z} + b) - S(A\mu_Z + b))(SA(\boldsymbol{Z} + b) - S(A\mu_Z + b))^T] = SA\mathbb{K}_{\boldsymbol{Z}\boldsymbol{Z}^T}A^TS^T$
   where $\mathbb{K}_{\boldsymbol{Z}\boldsymbol{Z}^T} = \mathbb{E}[(\boldsymbol{Z} - \mu_Z)(\boldsymbol{Z} - \mu_Z)^T] = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] - \mu_Z\mu_Z^T$
3. $Skew(Z') = \mathbb{E}[\Big(\frac{(Z' - \bar{Z}')}{cov(Z')}\Big)^3] = \mathbb{E}[\Big(\frac{SA(\boldsymbol{Z} - \mu_Z)}{SA\,\mathbb{K}_{\boldsymbol{Z}\boldsymbol{Z}^T}A^TS^T}\Big)^3]$

In our case, we will be considering $\mathbb{L}$ given by the rotation due to Σ_Z, the translation via μ_Z and the shearing via S. Through this construction, we bring in the covariance and mean of a multivariate Gaussian distribution and the rate parameter of a multivariate exponential distribution, respectively. From the properties of mean vectors and covariance matrices for a multivariate Gaussian distribution, we have $\mathbb{E}(\boldsymbol{Z}) = \mu_Z$, $Cov(\boldsymbol{Z}) = \Sigma_Z = diag(\sigma_i^2 + \frac{1}{\lambda_i^2}) = AA^T$ where $\mu_Z \in \mathbb{R}^d$ and $\Sigma_Z \in \mathbb{R}^{d\times d}$.

Therefore, the affine transformed $Z' = S(\Sigma_Z^{\frac{1}{2}}\boldsymbol{Z} + \mu_Z)$ and each z_i' can be written as: z_i' = (S($\Sigma_Z^{\frac{1}{2}}\boldsymbol{Z} + \mu_Z)$)$_i$ $= S\sqrt{\psi_i^T}U_i.\boldsymbol{Z} + S\mu_Z = V_i\sqrt{\psi_i^T}U_i.\boldsymbol{Z}.V_i^T + V_i\mu_Z V_i^T = \sum_j^d V_{ij}\sqrt{\psi_{ij}^T}U_{ij}.Z_j.V_{ij} + V_{ij}\mu_{Z_j}V_{ij} where \Sigma_Z = U\Psi U^T$, ψ_i is a row of Ψ, U_i is a row of U, λ_{Z_i} is a row of λ_Z, S is Cholesky decomposed to VV^T and V_i is a row of V.

PDF of Z' Suppose that $\boldsymbol{Z}$ is a random vector in the subset $T \subseteq \mathbb{R}^d$ and has continuous pdf f. Suppose that $Z' = r(\boldsymbol{Z})$ where r is a differentiable function from T to another subset $T^* \subseteq \mathbb{R}^d$, then we have the continuous pdf g for Z' as $g(Z') = f(\boldsymbol{Z})|\det(\frac{dZ}{dZ'})| = f(r^{-1}(Z'))|\det(\frac{d\bar{Z}}{dZ'})|$ where $\frac{dZ}{dZ'}$ is the Jacobian of the inverse of r. The inverse transformation is given by $\boldsymbol{Z} = \Sigma_Z^{-\frac{1}{2}}(S^{-1}Z' - \mu_Z)$, the Jacobian $\frac{dZ}{dZ'} = \Sigma_Z^{-\frac{1}{2}}S^{-1}$ and the determinant of the Jacobian of the inverse transformation is $\frac{1}{\sqrt{\det(\Sigma_Z)}\det S}$. Therefore

$$
\begin{aligned}
g(Z') &= f(r^{-1}(Z'))|\det(\frac{dZ}{dZ'})| \\
&\propto \frac{\det(\Lambda)}{\sqrt{\det(\Sigma_Z)}\det S}\exp(\frac{trace(\Sigma^{*T}\Sigma^*)}{2})\exp((-1)trace(\Sigma'))\frac{exp(\frac{(-1)trace(Z^{*T}Z^*)}{2})}{\det|Z^*|} \\
&:= mvEMG(Z'|\mu_Z, \Sigma_Z, \lambda_Z)
\end{aligned} \tag{1}
$$

where $z_i = \sum_j^d V_{ij}^{-1}(\sqrt{\psi_{ij}^T}U_{ij})^{-1}.Z_j'.V_{ij}^{-T} - (\sqrt{\psi_{ij}^T}U_{ij})^{-1}\mu_{Zj}$. We ensure that Z^* is non-singular.

4 Experiments

We compare the mvEMG with the multivariate skew-normal distribution (mvSN). The mvSN was introduced by *Azzalini and Capitanio (1999); Azzalini and DallaValle (1996)* and is a simple function of a multivariate Gaussian pdf and a univariate Gaussian cdf, given by: $f(\boldsymbol{x}) = 2\Phi(\boldsymbol{x}; \mathbf{0}, \Sigma)\boldsymbol{\Phi}(\alpha^T \boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{R}^d$ where Φ is the multivariate Gaussian density centered at $\mathbf{0}$, covariance matrix Σ and $\Phi(\cdot)$ is the CDF of the univariate spherical Gaussian, $\mathcal{N}(0, 1)$. α contains the shape for the Gaussian cdf. One convenient property of the mvSN is that the marginal distributions are scalar skew-normal distributions. The filled contour plots of the mvSN pdf are shown in Figure 2 (upper panel). Notice that when α are zeros (Figure 2 upper panel, first plot), the pdf reduces to that of a multivariate Gaussian. When α are positive, the pdfs attain right skewness. Figure 2 (lower panel) shows the filled contour plots for the mvEMG pdf generated using Equation 1. The scale parameters λ_Z constitute the off-diagonal elements of the shear matrix S that consists the skew-symmetric matrix as discussed in Section 3.2. We observe that the heights of the pdf are shorter than that of mvSN indicating that the mvEMG distributions, with the same shape and covariance parameters as that of the mvSN, are broader distributions with fatter tails.

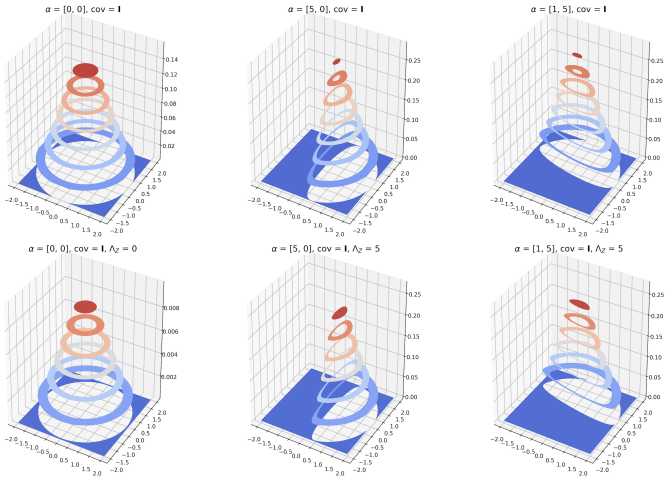

Figure 2: Filled contour plots of pdfs: (**Upper panel**): The multivariate skew normal (mvSN) pdf for varying shape parameters α and the spherical covariance matrix Σ. (**Lower panel**): The mvEMG pdf for the same varying shape parameters α (as above), the spherical covariance matrix Σ and varying scale parameters λ_Z that constitute the off-diagonal elements of the shear matrix S.

5 Conclusion

In this paper, we present mvEMG, the multivariate generalisation to a univariate exponentially-modified Gaussian (EMG) distribution, to handle multivariate skewed data giving access to mean, covariance and skewness of the distribution. We make use of an affine transformation involving rotation, translation and shearing to accommodate for the three moments in mvEMG. We present the statistical properties for mvEMG. The experiments on synthetic data show that the mvEMG is potentially better suited to model fat-tailed distributions than the mvSN distribution.

Self-critique: Even though the theory is in place and simple simulations show the usability of mvEMG, the work is stuck for two major reasons. We are unable to find (a) an efficient sampling strategy for probabilistic inference which is needed for downstream analysis such as cluster identification or building graphical models, and (b) a suitable real-world dataset to depict usefulness of mvEMG where skewness plays a crucial role in identifying underlying data patterns and that lead to further insights into the data. We hope that through this workshop, our work sparks interest and discussion for reviving this work.

References

[1] Zeisel, Amit & Muñoz-Manchado, Ana B. & Codeluppi, Simone & Lönnerberg, Peter & La Manno, Gioele & Juréus, Anna & Marques, Sueli & Munguba, Hermany & He, Liqun & Betsholtz, Christer & Rolny, Charlotte & Castelo-Branco, Gonçalo & Hjerling-Leffler, Jens & Linnarsson, Sten (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**(6226):1138-1142.

[2] Esary, James D & Marshall, Albert (1974) Multivariate distributions with exponential minimums. *The Annals of Statistics*.

[3] Naish, P. J. & Hartwell, S. (1998) Exponentially modified gaussian functions—a good model for chromatographic peaks in isocratic hplc? *Chromatographia*, **26**(1):285–296.

[4] Kong, H. & Ye, F. & Lu, X. & Guo, L. & Tian, J. & Xu, G. (2005) Deconvolution of overlapped peaks based on the expo-nentially modified gaussian model in comprehensive two-dimensional gas chromatography. *Journal Of Chromatography A*, **1086**(1-2):160–164.

[5] Golubev, A. (2010) Exponentially modified gaussian (emg) rel-evance to distributions related to cell proliferation and differentiation. *Journal of theoretical biology*, **262**(2):257–266

[6] Tyson, D. R. & Garbett, S. P. & Frick, P. L. & d Quaranta, V.Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. Nature methods,**9**(9):923, 2012

[7] Irizarry, R. A. & Hobbs, B. & Collin, F. & Beazer-Barclay, Y. D. & Antonellis, K. J. & Scherf, U. & Speed, T. P. (2003) Exploration, normalization, and summaries of high densityoligonucleotide array probe level data. *Biostatistics*, **4**(2):249–264.

[8] Azzalini, A. & Capitanio, C. (1999) Statistical applications of the multivariate skew normal distribution.*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(3):7579–602

[9] Azzalini, A. & DallaValle, A. (1996) The multivariate skew-normal distribution. *Biometrica*, 83:715–726.