entropy

MDPI

Article

Expected Logarithm of Central Quadratic Form and Its Use in KL-Divergence of Some Distributions

Pourya Habib Zadeh [1] and Reshad Hosseini [1,2,*]

[1] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, P.O. Box 14395-515, Tehran, Iran; p.habibzadeh@ut.ac.ir

[2] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran

* Correspondence: reshad.hosseini@ut.ac.ir; Tel.: +98-21-6111-9799

Academic Editor: Raúl Alcaraz Martínez

Received: 10 May 2016; Accepted: 21 July 2016; Published: 28 July 2016

Abstract: In this paper, we develop three different methods for computing the expected logarithm of central quadratic forms: a series method, an integral method and a fast (but inexact) set of methods. The approach used for deriving the integral method is novel and can be used for computing the expected logarithm of other random variables. Furthermore, we derive expressions for the *Kullback–Leibler* (KL) divergence of elliptical gamma distributions and angular central Gaussian distributions, which turn out to be functions dependent on the expected logarithm of a central quadratic form. Through several experimental studies, we compare the performance of these methods.

Keywords: expected logarithm; central quadratic form; Kullback–Leibler divergence; entropy; angular central Gaussian; elliptical gamma

1. Introduction

Expected logarithm of random variables usually appears in the expressions of important quantities like entropy and Kullback–Leibler (KL) divergence [1–3]. The second kind moment is an important statistics method used in estimation problems [4,5]. It also appears in an important class of inference algorithms called the variational Bayesian inference [6,7]. Furthermore, the geometric mean of a random variable which has been used in economics [8,9] is equal to the exponential of the expected logarithm of that random variable.

Central quadratic forms (CQFs) have many applications, and most of them stem from the fact that they are asymptotically equivalent to many statistics for testing null hypotheses. They are used for finding the number of components in mixtures of Gaussians [10], to test goodness-of-fit for some distributions [11] and as test statistics for dimensionality reduction in inverse regression [12].

In this paper, we develop three algorithms for computing the expected logarithm of CQFs. There is a need to develop special algorithms for it because CQFs do not have a closed-form probability density function, which makes the computation of their expected logarithms difficult. Although there is a vast literature on many different ways for calculating probability distributions of CQFs (see [13–16]), we have not found any work on calculating their expected logarithms. It is worth noting that one of our three algorithms is based upon works for computing the probability density function of CQFs using a series of gamma random variables [13,14]. We also derive expressions for the KL-divergence of two distributions that are subclasses of generalized elliptical distributions. These are *zero-mean elliptical gamma* (ZEG) distribution and *angular central Gaussian* (ACG) distribution. The only term in their KL-divergences that can not be computed in terms of

elementary functions is an expected logarithm of a CQF, which can be computed by one of our developed algorithms.

The KL-divergence or the relative entropy was first introduced in [17] as a generalization of Shannon's definition of information [18]. This divergence has been used extensively by statisticians and engineers. Many popular divergence classes like f-divergence and alpha-divergence have been introduced as generalizations to this divergence [19]. This divergence has several invariance properties like scale invariance that makes it an interesting dissimilarity measure in statistical inference problems [20]. KL-divergence is also used as a criterion for model selection [21], hypothesis testing [22], and merging in mixture models [23,24]. Additionally, it can be used as a measure of dissimilarity in classification problems, for example, text classification [25], speech recognition [26], and texture classification [27,28].

The wide applicability of KL-divergence as a useful dissimilarity measure persuade us to derive the KL-divergence for two important distributions. One of them is ZEG [29] that has a rich modeling power and allows heavy and light tail and different peak behaviors [30,31]. The other is ACG [32] which is a distribution on the unit sphere that has been used in many applications [33–36]. This distribution has many nice features; for example, its maximum likelihood estimator is asymptotically the most robust estimator of the scatter matrix of an elliptical distribution in the sense of minimizing the maximum asymptotic variance [37].

1.1. Contributions

To summarize, the key contributions of our paper are the following:

- Introducing three methods for computing the expected logarithm of a CQF.
- Proposing a procedure for computing the expected logarithm of an arbitrary positive random variable.
- Deriving expressions for the entropy and the KL-divergence of ZEG and ACG distributions (the form of KL-divergence between ZEG distributions appeared in [38] but without its derivations).

The methods for computing the expected logarithm of a CQF differ in running-time and accuracy. Two of these, namely integral and series methods, are exact. The third method is a fast but inexact set of methods. The integral method is a direct application of our proposed procedure for computing the expected logarithm of positive random variables. We propose two fast methods that are based on approximating the CQF with a gamma random variable. We show that these fast methods give upper and lower bounds to the true expected logarithm. This leads us to develop another fast method based on a convex combination of the other two fast methods. Whenever the weights of the CQF are eigenvalues of a matrix as in the case of KL-divergences, the fast methods can be very efficient because they do not need eigenvalue computation.

1.2. Outline

The remainder of this paper is organized as follows. Section 2 proposes three different methods for computing the expected logarithm of a CQF. Furthermore, a theorem is stated at the beginning of this section that has a pivotal role in the first two methods. Then, we derive expressions for the KL-divergence and entropy of ZEG and ACG distributions in Section 3. Afterwards, in Section 4, multiple experiments are conducted to examine the performance of three methods for computing the aforementioned expected logarithm in terms of accuracy and computational time. Finally, Section 5 presents our conclusions. To improve the readability of the manuscript, the proofs of some theorems are presented in appendices.

2. Calculating the Expected Logarithm of a Central Quadratic Form

Suppose N_i is the i-th random variable in the series of d independent standard normal random variables, i.e., normal random variables with zero-means and unit variances. Then the central (Gaussian) quadratic form is the following random variable:

$$U = \sum_{i=1}^{d} \lambda_i N_i^2, \tag{1}$$

where λ_is are non-negative real numbers. Note that N_i^2s are chi-square random variables with degree of freedoms equal to one; therefore, this random variable is also called the weighted sum of chi-square random variables. To the best of our knowledge, the expected logarithm of the random variable U does not have a closed-form expression using elementary mathematical functions. For its calculation, we propose three different approaches, namely an integral method, a series method and a set of fast methods. Each of them has its specific properties and does well in certain situations.

In the following theorem, a relation between the expected logarithm of two positive random variables distributed according to arbitrary densities and the Laplace transform of these two densities is given. This theorem is used in the integral method and the fast method. Note that the assumptions of the following theorem are unrestrictive. Therefore, it can also be used for computing the expected logarithm of other positive random variables.

Theorem 1. *Let X and Y be two positive random variables, F and G be their cumulative distribution functions, and f and g be their probability density functions. Furthermore, suppose that $\mathcal{F}$ and $\mathcal{G}$ are the Laplace transform of f and g, respectively. If $\lim_{x\to\infty} \log(x)(G(x) - F(x)) = 0$, $\lim_{x\to 0^+} \log(x)(G(x) - F(x)) = 0$ and $\int_0^\infty \frac{\mathcal{G}(\sigma)-\mathcal{F}(\sigma)}{\sigma} d\sigma$ exists, then*

$$\mathbb{E}[\log X] - \mathbb{E}[\log Y] = \int_0^\infty \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma} d\sigma. \tag{2}$$

Proof. Using the definition of Laplace transform, we have

$$\mathcal{G}(s) - \mathcal{F}(s) = \int_0^\infty (g(x) - f(x)) \exp(-sx)\, dx. \tag{3}$$

Using the integration property of Laplace transform, we have

$$\frac{\mathcal{G}(s) - \mathcal{F}(s)}{s} = \int_0^\infty \left[\int_0^x (g(\tau) - f(\tau))\, d\tau\right] \exp(-sx)\, dx.$$

Using the frequency integration property of Laplace transform, and the formulas $F(x) = \int_0^x f(\tau)d\tau$, and $G(x) = \int_0^x g(\tau)d\tau$ we further obtain

$$\int_s^\infty \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma}\, d\sigma = \int_0^\infty \frac{1}{x}(G(x) - F(x)) \exp(-sx)\, dx.$$

Letting s in the above equation go to zero and since $\int_0^\infty \frac{\mathcal{G}(\sigma)-\mathcal{F}(\sigma)}{\sigma}\, d\sigma$ exists, we further obtain

$$\int_0^\infty \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma}\, d\sigma = \int_0^\infty \frac{1}{x}(G(x) - F(x))\, dx.$$

Using the integration by parts formula having $\log(x)$ and $G(x) - F(x)$ as its parts, we have

$$\int_0^\infty \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma}\, d\sigma = \Big[\log(x)(G(x) - F(x))\Big]_0^\infty - \int_0^\infty \log(x)(g(x) - f(x))\, dx. \tag{4}$$

Since $\lim_{x\to\infty} \log(x)(G(x) - F(x)) = 0$, and $\lim_{x\to 0^+} \log(x)(G(x) - F(x)) = 0$, we obtain

$$\Big[\log(x)(G(x) - F(x))\Big]_0^\infty = 0. \tag{5}$$

Hence by using (5) in (4), we have

$$\int_0^\infty \log(x) f(x)\, dx - \int_0^\infty \log(x) g(x)\, dx = \int_0^\infty \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma} d\sigma. \tag{6}$$

From the definition of expectation, relation (2) is obtained. □

2.1. Integral Method

In this part, we will use Theorem 1 for computing the expected logarithm of a CQF. To this end, we choose a random variable Y that has a closed-form formula for its expected logarithm and Laplace transform of its density. A possible candidate is the gamma random variable. The density of gamma random variable has the following Laplace transform:

$$(1 + \theta s)^{-k}, \tag{7}$$

where k and θ are its shape and scale parameters, respectively. Also, the expected logarithm of this random variable is $\Psi(k) + \log(\theta)$, where $\Psi(\cdot)$ is digamma function.

Using the convolution property of Laplace transform, it is easy to see that the density function of the CQF given in (1) has the following closed-form Laplace transform:

$$\prod_{i=1}^{d} (1 + 2\lambda_i s)^{-\frac{1}{2}}. \tag{8}$$

Lemmas 2 and 3 show that a CQF and a gamma random variable satisfy the conditions of Theorem 1. For proving Lemma 2, we need Lemma 1. First of all, let us express the following trivial proposition.

Proposition 1. *Let $X_1, \ldots, X_n$ be arbitrary real random variables. Suppose we have two many-to-one transformations $Y = h(X_1, \ldots, X_n)$ and $Z = g(X_1, \ldots, X_n)$. If the following inequality holds for any x_is in the support of random variables X_is:*

$$h(x_1, \ldots, x_n) \geq g(x_1, \ldots, x_n), \tag{9}$$

then we have the following inequality between the cumulative distribution functions of random variables Y and Z:

$$F_Y(v) \leq F_Z(v), \quad \text{for all } v \in \mathbb{R}. \tag{10}$$

Lemma 1. *Let F be the cumulative distribution function of a CQF, that is $\sum_{i=1}^{d} \lambda_i N_i^2$, where λ_is are positive real numbers and N_is are independent standard normal random variables. Also, let $G(x; k, \theta)$ be the cumulative distribution function of a gamma random variable with parameters k and θ, then the following inequalities hold:*

$$G\big(x; \tfrac{d}{2}, 2\lambda_{\max}\big) \leq F(x) \leq G\big(x; \tfrac{d}{2}, 2\lambda_{\min}\big), \quad \text{for all } x \in \mathbb{R}^+, \tag{11}$$

where $\lambda_{\max} = \max\{\lambda_i\}_{i=1}^d$, and $\lambda_{\min} = \min\{\lambda_i\}_{i=1}^d$.

Proof. This lemma is an immediate consequence of Proposition 1 and the following relation, knowing that $\lambda \sum_{i=1}^{d} N_i^2$ is a gamma random variable with the shape parameter $d/2$ and the scale parameter 2λ:

$$\lambda_{\min} \sum_{i=1}^{d} x_i^2 \leq \sum_{i=1}^{d} \lambda_i x_i^2 \leq \lambda_{\max} \sum_{i=1}^{d} x_i^2, \quad \text{for all } x_i \in \mathbb{R}. \tag{12}$$

□

Lemma 2. *Let G be the cumulative distribution function of an arbitrary gamma random variable and F be the cumulative distribution function of random variable $\sum_{i=1}^{d} \lambda_i N_i^2$, where λ_is are positive real numbers and N_is are independent standard normal random variables, then $\lim_{x\to\infty} \log(x)(G(x) - F(x)) = 0$, and $\lim_{x\to 0^+} \log(x)(G(x) - F(x)) = 0$.*

The proof of this lemma can be found in Appendix A.

Lemma 3. *Let $\mathcal{G}$ be the Laplace transform of probability density function of an arbitrary gamma random variable and $\mathcal{F}$ be the Laplace transform of probability density function of $\sum_{i=1}^{d} \lambda_i N_i^2$, where λ_is are positive real numbers and N_is are independent standard normal random variables, then $\int_0^\infty \frac{\mathcal{G}(\sigma)-\mathcal{F}(\sigma)}{\sigma} d\sigma$ is convergent.*

The proof of this lemma can be found in Appendix B.

According to Lemmas 2 and 3, the conditions of Theorem 1 hold by choosing X to be a CQF given in (1), and Y to be an arbitrary gamma random variable. Therefore, we can use (2) for calculating the expected logarithm of a CQF, and it is given by

$$\mathbb{E}\left[\log\left(\sum_{i=1}^{d} \lambda_i N_i^2\right)\right] = \Psi(k) + \log(\theta) + \int_0^\infty \frac{(1+\theta\sigma)^{-k} - \prod_{i=1}^{d} (1+2\lambda_i\sigma)^{-\frac{1}{2}}}{\sigma} d\sigma. \tag{13}$$

The above equation holds for any choice of positive scalars k and θ. To the best of our knowledge, the above integral does not have a closed-form solution, so it must be computed numerically. This integral can be computed numerically using the variety of techniques available for one-dimensional integrals (see for example [39]).

2.2. Fast Methods

The integral method explained in the previous part can be computationally expensive for some applications. To this end, we derive three approximations that can be calculated analytically and, therefore, are much faster.

Using the first or higher order Taylor expansion around $\mathbb{E}[U]$ to approximate the expected logarithm of U has been already proposed in the literature [6,40]. However, we observed that lower order Taylor expansion does not give a very accurate approximation. Therefore, we use two different approximations, for which we can show that they provide a lower and an upper bound for the true expected logarithm. Finally, a convex combination of these two is used to get the final approximation.

Two different gamma distributions have been used in [15,41,42] to approximate a CQF. Since the expected logarithm of a gamma random variable has a closed-form solution, we use the expected logarithm of these gamma random variables to approximate the expected logarithm of a CQF. A further justification for this idea can be given based on (13) by choosing the shape and the scale parameters of gamma distribution such that the magnitude of the integral part in (13) becomes smaller.

Since the weights of CQF in the KL-divergence formulas in Section 3 are eigenvalues of a positive definite matrix $\mathbf{\Sigma}$, we express the approximations based on this matrix. This way of expressing the approximations also elucidates the fact that the eigenvalues do not need to be calculated, which shows further computational benefits of these approximations. The shape and scale parameters of the first

approximating gamma random variable are $d/2$ and $2\text{tr}(\mathbf{\Sigma})/d$, respectively. Therefore, for the *first fast approximation* we have

$$\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\lambda_i N_i^2\right)\right] \approx \Psi\left(\tfrac{d}{2}\right) + \log\left(\frac{2\text{tr}(\mathbf{\Sigma})}{d}\right). \tag{14}$$

The shape and scale parameters of the gamma random variable for the second approximation are $\text{tr}(\mathbf{\Sigma})^2/2\text{tr}(\mathbf{\Sigma}^2)$ and $2\text{tr}(\mathbf{\Sigma}^2)/\text{tr}(\mathbf{\Sigma})$, respectively. Then, we obtain the following formula for the *second fast approximation*:

$$\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\lambda_i N_i^2\right)\right] \approx \Psi\left(\frac{\text{tr}(\mathbf{\Sigma})^2}{2\text{tr}(\mathbf{\Sigma}^2)}\right) + \log\left(\frac{2\text{tr}(\mathbf{\Sigma}^2)}{\text{tr}(\mathbf{\Sigma})}\right). \tag{15}$$

The following theorem shows that these approximations are lower and upper bounds to the true expected logarithm.

Theorem 2. *If $U = \sum_{i=1}^{d}\lambda_i N_i^2$, where λ_is are eigenvalues of positive definite matrix $\mathbf{\Sigma}_{d\times d}$ and N_is are independent standard normal random variables, then*

$$\Psi\left(\frac{tr(\mathbf{\Sigma})^2}{2tr(\mathbf{\Sigma}^2)}\right) + \log\left(\frac{2tr(\mathbf{\Sigma}^2)}{tr(\mathbf{\Sigma})}\right) \leq \mathbb{E}[\log(U)] \leq \Psi\left(\tfrac{d}{2}\right) + \log\left(\frac{2tr(\mathbf{\Sigma})}{d}\right). \tag{16}$$

The proof of this theorem can be found in Appendix C.

From this theorem, we can conclude that there exist some convex combinations of the two previously mentioned approximations which perform equal or better than each of them, in the sense that they are closer to the true expected logarithm. Therefore, we define the *third fast approximation* to be

$$\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\lambda_i N_i^2\right)\right] \approx (1-l)\left[\Psi\left(\tfrac{d}{2}\right) + \log\left(\frac{2\text{tr}(\mathbf{\Sigma})}{d}\right)\right] + l\left[\Psi\left(\frac{\text{tr}(\mathbf{\Sigma})^2}{2\text{tr}(\mathbf{\Sigma}^2)}\right) + \log\left(\frac{2\text{tr}(\mathbf{\Sigma}^2)}{\text{tr}(\mathbf{\Sigma})}\right)\right]. \tag{17}$$

To determine parameter $l \in [0,1]$ in the above equation, we used the least squares fitting on thousands of positive definite matrices with different dimensions and unit trace sampled uniformly according to an algorithm given in [43]. We observed that the fitted value is roughly equal to $l = 0.7$ and dimensionality has a negligible effect on the best value of l. For the case of $d = 20$, the mean squared error for various values of l can be seen in Figure 1.

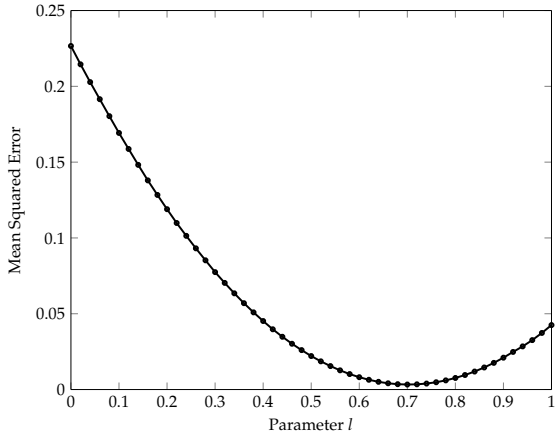

Figure 1. The mean squared error of the third fast method for approximating the expected logarithm of a CQF as a function of parameter l.

2.3. Series Method

One can represent the probability density function of a CQF given by (1) as an infinite weighted sum of gamma densities [13,14],

$$f_U(u) = \sum_{j=0}^{\infty} c_j g\left(u; \tfrac{d}{2} + j, 2\beta\right), \tag{18}$$

where $g(u; d/2 + j, 2\beta)$ is the probability density function of a gamma random variable with parameters $d/2 + j$ and 2β, and

$$c_k = \frac{1}{k} \sum_{r=0}^{k-1} v_{k-r} c_r, \tag{19}$$

$$v_k = \frac{1}{2} \sum_{j=1}^{d} \left(1 - \beta \lambda_j^{-1}\right)^k, \tag{20}$$

$$c_0 = \prod_{j=1}^{d} \sqrt{\beta \lambda_j^{-1}}. \tag{21}$$

This result can be used for deriving a series formula for the expected logarithm of U. Thus,

$$\mathbb{E}[\log U] = \sum_{j=0}^{\infty} c_j \left(\Psi(\tfrac{d}{2} + j) + \log(2\beta)\right). \tag{22}$$

Ruben [13] analyzed the effect of various βs on the behavior of the series expansion and proposed the following β as an appropriate one:

$$\beta = \frac{2\lambda_{\max}\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}. \tag{23}$$

By using this β, $\sum_{j=0}^{\infty} c_j = 1$ holds [13] and also knowing that the following relation holds for the digamma function:

$$\Psi(x+1) = \frac{1}{x} + \Psi(x),$$

then (22) can be simplified

$$\begin{aligned}
\mathbb{E}[\log U] &= \log(2\beta) + c_0 \Psi(\tfrac{d}{2}) + \sum_{j=1}^{\infty} c_j \left(\Psi(\tfrac{d}{2}) + \sum_{l=0}^{j-1} \frac{1}{\frac{d}{2} + l} \right) \\
&= \log(2\beta) + \Psi(\tfrac{d}{2}) + \sum_{j=1}^{\infty} c_j \sum_{l=0}^{j-1} \frac{1}{\frac{d}{2} + l} \\
&= \log(2\beta) + \Psi(\tfrac{d}{2}) + \sum_{i=0}^{\infty} \left(\frac{1}{\frac{d}{2} + i} \sum_{k=i+1}^{\infty} c_k \right) \\
&= \log(2\beta) + \Psi(\tfrac{d}{2}) + \sum_{i=0}^{\infty} \left(\frac{1}{\frac{d}{2} + i} \left(1 - \sum_{k=0}^{i} c_k\right) \right).
\end{aligned} \tag{24}$$

To approximate this formula, we cut the series coefficients, which means we only use a finite number of terms to evaluate the expectation:

$$\hat{\mathbb{E}}[\log U] = \log(2\beta) + \Psi(\tfrac{d}{2}) + \sum_{i=0}^{L-1} \left(\frac{1}{\frac{d}{2}+i} \left(1 - \sum_{k=0}^{i} c_k \right) \right). \tag{25}$$

For this approximation, it is possible to compute an error bound which is expressed by the following lemma.

Lemma 4. *The bound for error of the approximation* (25) *for* $L > d\epsilon/(2-2\epsilon)$ *is as follows:*

$$\mathbb{E}[\log U] - \hat{\mathbb{E}}[\log U] \;\leq\; c_0 \frac{\epsilon^{L+1}}{\left(1 - \frac{d/2+L}{L}\epsilon\right)^2} \frac{\Gamma(\frac{d}{2}+L)}{\Gamma(\frac{d}{2}+1)(L+1)!}, \tag{26}$$

where $\Gamma(\cdot)$ *is gamma function, and* $\epsilon = (\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min})$.

The proof of this lemma is in Appendix D.

By using this bound, it is possible to calculate the expected logarithm with a given accuracy by selecting an appropriate L. Note that the upper bound given by (26) is growing with respect to ϵ, and ϵ is also increasing with $\lambda_{\max}/\lambda_{\min}$. As we will see in the simulation studies, when the ratio $\lambda_{\max}/\lambda_{\min}$ as well as the dimensionality d are small, this method performs better than the integral method.

3. KL-Divergence of Two Generalized Elliptical Distributions

In this section, we derive expressions for the KL-divergence and the entropy of two subclasses of generalized elliptical distributions, namely ZEG and ACG distributions [44]. We first start by reviewing some related materials.

3.1. Some Background on the Elliptical Distributions

Suppose the d-dimensional random vector $\mathbf{X}$ is distributed according to a *zero-mean elliptical contoured* (ZEC) distribution with a positive definite scatter matrix $\mathbf{\Sigma}_{d\times d}$, that is $\mathbf{X} \sim \mathcal{ZEC}(\mathbf{\Sigma}, \varphi)$. The probability density function of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{\Sigma}, \varphi) = |\mathbf{\Sigma}|^{-\frac{1}{2}} \varphi\big(\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}\big), \tag{27}$$

for some density generator functions $\varphi\colon \mathbb{R}^+ \to \mathbb{R}$. We know that we can decompose the vector $\mathbf{X}$ into a uniform hyper-spherical component and a scaled-radial component so that, $\mathbf{X} = \mathbf{\Sigma}^{1/2} R\mathbf{U}$, where $\mathbf{U}$ is uniformly distributed over the unit sphere $\mathbb{S}^{d-1}$ and R is a univariate random variable given by $R = \|\mathbf{\Sigma}^{-1/2}\mathbf{X}\|_2$ [45]. Then, the random variable R has the density

$$f_R(r) = 2\pi^{\frac{d}{2}} \varphi(r^2) r^{d-1} / \Gamma(\tfrac{d}{2}). \tag{28}$$

Therefore, square radial component $\mathrm{Y} = R^2$ has the following density:

$$f_{\mathrm{Y}}(v) = \pi^{\frac{d}{2}} \varphi(v) v^{\frac{d}{2}-1} / \Gamma(\tfrac{d}{2}). \tag{29}$$

A ZEG is a ZEC whose square radial component is distributed according to a gamma distribution $\mathrm{Y} \sim \mathit{Gamma}(a, b)$. A gamma-distributed random variable has the density

$$f_{\mathrm{Y}}(v) = \Gamma(a)^{-1} b^{-a} v^{a-1} \exp\left(-v/b\right), \tag{30}$$

where a is a *shape* parameter and b is a *scale* parameter. So the probability density function of a d-dimensional random variable $\mathbf{X} \sim \mathcal{ZEG}(\mathbf{\Sigma}, a, b)$ is given by

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{\Sigma}, a, b) = \frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\Gamma(a)b^a|\mathbf{\Sigma}|^{\frac{1}{2}}}\left(\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}\right)^{a-\frac{d}{2}}\exp\left(-b^{-1}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}\right), \tag{31}$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{\Sigma} \succ 0$ is its scatter matrix, also $a, b > 0$ are certain scale and shape parameters [31].

When ZEC random variable is projected onto a unit sphere, the resulting random variable is called ACG and denoted by $\mathbf{X} \sim \mathcal{ACG}(\mathbf{\Sigma})$. This distribution unlike many other distributions on the unit sphere has a nice closed-form density given by

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{\Sigma}) = \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\left(\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}\right)^{-\frac{d}{2}}, \tag{32}$$

where $\mathbf{x} \in \mathbb{S}^{d-1}$ and $\mathbf{\Sigma} \succ 0$ is its scatter matrix.

3.2. KL-Divergence between ZEG Distributions

Suppose we have two probability distributions P and Q with probability density functions p and q, KL-divergence between these two distributions is defined by

$$\mathrm{KL}(P||Q) = \int \log(p(\mathbf{x}))p(\mathbf{x})d\mathbf{x} - \int \log(q(\mathbf{x}))p(\mathbf{x})d\mathbf{x}. \tag{33}$$

The negative of the first part, $\mathrm{H}(\mathbf{X}) = -\int \log(p(\mathbf{x}))p(\mathbf{x})d\mathbf{x}$, is the entropy and the second part, $\mathbb{E}[-\log(q(\mathbf{X}))] = -\int \log(q(\mathbf{x}))p(\mathbf{x})d\mathbf{x}$, is the averaged log-loss term, where $\mathbf{X}$ is a random variable distributed according to P.

Following lemma gives a general expression for the KL-divergence between two ZEC distributions. It is then used for deriving the KL-divergence between two ZEG distributions.

Lemma 5. *Suppose we have two probability distributions on random variable* $\mathbf{Y}$, $P_{\mathbf{Y}} = \mathcal{ZEC}(\mathbf{\Sigma}_1, \varphi)$ *and* $Q_{\mathbf{Y}} = \mathcal{ZEC}(\mathbf{\Sigma}_2, \varphi')$, *then the KL-divergence between these two distributions is given by the following expression:*

$$\begin{aligned} KL(P||Q) = &-\tfrac{1}{2}\log(|\mathbf{\Sigma}|) + \int_0^\infty \log\left(v^{1-\frac{d}{2}}f_{\mathrm{Y}}(v)\right)f_{\mathrm{Y}}(v)dv \\ &-\iint_0^\infty \tfrac{1}{r}f_{wd}(\tfrac{v}{r})f_{\mathrm{Y}}(r)\log\left(v^{1-\frac{d}{2}}f_{\mathrm{Y}'}(v)\right)dvdr, \end{aligned} \tag{34}$$

where f_{Y} *and* $f_{\mathrm{Y}'}$ *are the square radial components of distributions* P *and* Q, *respectively. Also,* f_{wd} *is the density of* $\sum_{i=1}^d \lambda_i N_i^2 / \sum_{i=1}^d N_i^2$, *where* N_i*s are independent standard normal random variables, and* $\lambda_1, \ldots, \lambda_d$ *are eigenvalues of matrix* $\mathbf{\Sigma} = \mathbf{\Sigma}_2^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1/2}$.

Proof. KL-divergence is known to be invariant against invertible transformations of random variable $\mathbf{Y}$ [46]. To simplify the derivations, we apply a linear transformation $\mathbf{X} = \mathbf{\Sigma}_2^{-1/2}\mathbf{Y}$ that makes the scatter matrix of the second distribution identity. By using this change of variable, the problem becomes that of KL-divergence computation between $P_{\mathbf{X}} = \mathcal{ZEC}(\mathbf{\Sigma}, \varphi)$ and $Q_{\mathbf{X}} = \mathcal{ZEC}(\mathbf{I}, \varphi')$, where $\mathbf{\Sigma} = \mathbf{\Sigma}_2^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1/2}$.

As expressed in (33), the KL-divergence is the subtraction of the entropy from the averaged log-loss. Firstly, let us derive the entropy of $\mathbf{X}$ having distribution $P_{\mathbf{X}}$, that is $\mathrm{H}(\mathbf{X}) = -\int \log(p(\mathbf{x}))p(\mathbf{x})d\mathbf{x}$.

Assume the change of integration variable $\mathbf{y} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$ and use (27), then we obtain the following expression for $\mathrm{H}(\mathbf{X})$:

$$\mathrm{H}(\mathbf{X}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}|) - \int \log\Big(\varphi(\mathbf{y}^\top\mathbf{y})\Big)\,\varphi(\mathbf{y}^\top\mathbf{y})d\mathbf{y}.$$

Let $r = \|\mathbf{y}\|_2$ and recall that the area of a sphere in dimension d with radius r equals $2r^{d-1}\pi^{d/2}/\Gamma(d/2)$, thus

$$\mathrm{H}(\mathbf{X}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}|) - \int_0^\infty \log\Big(\varphi(r^2)\Big)\,\frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}r^{d-1}\varphi(r^2)dr.$$

Using the change of variable $v = r^2$ and replacing φ by square radial density f_Y as expressed in (29), we obtain

$$\mathrm{H}(\mathbf{X}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}|) - \int_0^\infty \log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}}v^{1-\frac{d}{2}}f_\mathrm{Y}(v)\right)f_\mathrm{Y}(v)dv. \tag{35}$$

Now, we are deriving an expression for the averaged log-loss given by $\mathbb{E}[-\log(q(\mathbf{X}))] = -\mathbb{E}\big[\log\big(\varphi'(\mathbf{X}^\top\mathbf{X})\big)\big]$. The argument of function φ' is $\mathbf{X}^\top\mathbf{X}$; therefore, it is enough to compute the expectation of the function over the new random variable $Z = \|\mathbf{X}\|_2^2$:

$$\begin{aligned}\mathbb{E}[-\log(q(\mathbf{X}))] &= -\,\mathbb{E}_Z\big[\log\big(\varphi'(Z)\big)\big] \\ &= -\int_0^\infty \log\big(\varphi'(Z)\big)f_Z(v)dv,\end{aligned} \tag{36}$$

where f_Z is the density of $Z = \|\mathbf{X}\|_2^2$, wherein $\mathbf{X} \sim \mathcal{ZEC}(\mathbf{\Sigma}, \varphi)$. It is easy to see that the random variable Z can equally be written as $Z = \bar{\mathbf{X}}^\top\mathbf{\Sigma}\bar{\mathbf{X}}$ where $\bar{\mathbf{X}} \sim \mathcal{ZEC}(\mathbf{I}, \varphi)$. The density of Z with this representation has already been reported in [47]:

$$f_Z(z) = \int_0^\infty \frac{1}{r}f_{\mathrm{wd}}(z/r)f_\mathrm{Y}(r)dr, \tag{37}$$

where f_Y is the square radial density of $p_\mathbf{Y}$, and f_{wd} is the density of a linear combination of Dirichlet random variable components,

$$\Lambda = \sum_{j=1}^{s} l_j D_j, \tag{38}$$

where $\mathbf{D} = (D_1, \ldots, D_s)^\top$ is a Dirichlet random variable with parameters $(r_1/2, \ldots, r_s/2)$, and l_js are s distinct eigenvalues of the positive definite matrix $\mathbf{\Sigma}$ with respective multiplicities r_j, for $j = 1, \ldots, s$.

It is known that if random variables $C_1, \ldots, C_s$ are independent chi-square random variables having $r_1, \ldots, r_s$ degrees of freedom, and $C = \sum_{j=1}^s C_j$, then $(C_1/C, \ldots, C_s/C)^\top$ is a Dirichlet random variable with the parameters $(r_1/2, \ldots, r_s/2)$ [48]. Hence, the random variable Λ in (38) can be expressed as $\Lambda = \sum_{i=1}^s l_j C_j/C$. Equivalently, if $N_1, \ldots, N_d$ are independent standard normal random variables, then Λ can be written as

$$\Lambda = \frac{\sum_{i=1}^d \lambda_i N_i^2}{\sum_{i=1}^d N_i^2}. \tag{39}$$

Using (37) in (36) and replacing φ' by square radial density $f_{\mathrm{Y}'}$ as expressed in (29), we obtain the following expression for the averaged log-loss:

$$\mathbb{E}[-\log(q(\mathbf{X}))] = -\int_0^\infty\!\!\int \frac{1}{r}f_{\mathrm{wd}}(v/r)f_\mathrm{Y}(r)\log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}}v^{1-\frac{d}{2}}f_{\mathrm{Y}'}(v)\right)dvdr. \tag{40}$$

Subtracting (35) from (40), we obtain (34). □

Until now, we derived an expression for the KL-divergence between two ZEC distributions. We can further simplify the KL-divergence for the case of ZEG distributions to avoid computing double-integration, and the following theorem proves it.

Theorem 3. *Suppose we have two distributions $P_{\mathbf{Y}} = \mathcal{ZEG}(\mathbf{\Sigma}_1, a_p, b_p)$, and $Q_{\mathbf{Y}} = \mathcal{ZEG}(\mathbf{\Sigma}_2, a_q, b_q)$, then the entropy of random variable $\mathbf{Y}$ distributed according to $P_{\mathbf{Y}}$ and the KL-divergence between these two distributions are given by the following expressions:*

$$H(\mathbf{Y}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}_1|) - \log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\Gamma(a_p)b_p^{a_p}}\right) - (a_p - \tfrac{d}{2})\left(\Psi(a_p) + \log(b_p)\right) + a_p, \tag{41}$$

$$\begin{aligned} KL(P||Q) = &-\tfrac{1}{2}\log(|\mathbf{\Sigma}|) + \log\left(\frac{\Gamma(a_q)b_q^{a_q}}{\Gamma(a_p)b_p^{a_q}}\right) + (a_p - a_q)\Psi(a_p) - a_p \\ &+ \frac{a_p b_p}{d\, b_q} tr(\mathbf{\Sigma}) + (\tfrac{d}{2} - a_q)\left(\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\lambda_i N_i^2\right)\right] - \Psi(\tfrac{d}{2}) - \log(2)\right), \end{aligned} \tag{42}$$

where $\Psi(\cdot)$ is digamma function, and $tr(\cdot)$ is the trace of a matrix. Also N_is are independent standard normal random variables, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues of matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_2^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1/2}$.

Proof. Like the previous lemma, we apply the change of variable $\mathbf{X} = \mathbf{\Sigma}_2^{-1/2}\mathbf{Y}$ and compute the KL-divergence between the transformed distributions. The expression for entropy (35) in the case of ZEG distributions becomes

$$\mathrm{H}(\mathbf{X}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}|) - \int_0^\infty \log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\Gamma(a_p)b_p^{a_p}} v^{a_p - \frac{d}{2}} \exp\left(-\frac{v}{b_p}\right)\right) \frac{1}{\Gamma(a_p)b_p^{a_p}} v^{a_p - 1}\exp\left(-\frac{v}{b_p}\right) dv. \tag{43}$$

Next, recall the following gamma function identities [49]:

$$\frac{1}{\Gamma(a+1)b^{a+1}}\int_0^\infty r^a \exp\left(-\frac{r}{b}\right) dr = 1, \tag{44}$$

$$\frac{1}{\Gamma(a+1)b^{a+1}}\int_0^\infty \log(r) r^a \exp\left(-\frac{r}{b}\right) dr = \Psi(a+1) + \log(b). \tag{45}$$

Using (44) and (45), we can simplify (43) to obtain

$$\mathrm{H}(\mathbf{X}) = \tfrac{1}{2}\log(|\mathbf{\Sigma}|) - \log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\Gamma(a_p)b_p^{a_p}}\right) + (\tfrac{d}{2} - a_p)\left(\Psi(a_p) + \log(b_p)\right) + a_p.$$

Since $\mathbf{Y} = \mathbf{\Sigma}_2^{1/2}\mathbf{X}$, we can trivially derive the expression of $\mathrm{H}(\mathbf{Y})$ given in (41).

For deriving the averaged log-loss term, we obtain the following expression by putting the gamma square radial component (30) into (40):

$$\mathbb{E}[-\log(q(\mathbf{X}))] = -\int_0^\infty\!\!\int \frac{1}{r} f_{\mathrm{wd}}(v/r)\frac{r^{a_p-1}}{\Gamma(a_p)b_p^{a_p}}\exp\left(-\frac{r}{b_p}\right)\log\left(\frac{\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\Gamma(a_q)b_q^{a_q}} v^{a_q - \frac{d}{2}}\exp\left(-\frac{v}{b_q}\right)\right) dv dr.$$

We apply the change of variable $\mu = v/r$ and express the integrals in terms of new variables μ and r,

$$\begin{aligned}\mathbb{E}[-\log(q(\mathbf{X}))] = &-\log(\Gamma(\tfrac{d}{2})) + \log(\pi^{\frac{d}{2}}\Gamma(a_q)b_q^{a_q}) \\ &+ \frac{\frac{d}{2}-a_q}{\Gamma(a_p)b_p^{a_p}}\int_0^\infty f_{\mathrm{wd}}(\mu)\log(\mu)d\mu\int_0^\infty r^{a_p-1}\exp\left(-\frac{r}{b_p}\right)dr \\ &+ \frac{\frac{d}{2}-a_q}{\Gamma(a_p)b_p^{a_p}}\int_0^\infty f_{\mathrm{wd}}(\mu)d\mu\int_0^\infty \log(r)r^{a_p-1}\exp\left(-\frac{r}{b_p}\right)dr \\ &+ \frac{1}{b_q\Gamma(a_p)b_p^{a_p}}\int_0^\infty f_{\mathrm{wd}}(\mu)\mu d\mu\int_0^\infty r^{a_p}\exp\left(-\frac{r}{b_p}\right)dr.\end{aligned}$$

Using the equalities (44) and (45), we obtain

$$\begin{aligned}\mathbb{E}[-\log(q(\mathbf{X}))] = &-\log(\Gamma(\tfrac{d}{2})) + \log(\pi^{\frac{d}{2}}\Gamma(a_q)b_q^{a_q}) + (\tfrac{d}{2}-a_q)\left(\Psi(a_p)+\log(b_p)\right) \\ &+ \frac{a_pb_p}{d\,b_q}\int_0^\infty f_{\mathrm{wd}}(\mu)\mu d\mu + (\tfrac{d}{2}-a_q)\int_0^\infty f_{\mathrm{wd}}(\mu)\log(\mu)d\mu,\end{aligned}$$

where similar to the previous lemma, f_{wd} is the density of the random variable $\Lambda = \sum_{i=1}^d \lambda_i N_i^2 / \sum_{i=1}^d N_i^2$, where N_is are independent standard normal random variables, and $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of matrix $\mathbf{\Sigma}$. Subtracting the entropy from the averaged log-loss and knowing that $\int_0^\infty f_{\mathrm{wd}}(\mu)\mu d\mu = \mathbb{E}[\Lambda]$ and $\int_0^\infty f_{\mathrm{wd}}(\mu)\log(\mu)d\mu = \mathbb{E}[\log(\Lambda)]$, we obtain

$$\mathrm{KL}(P||Q) = -\tfrac{1}{2}\log(|\mathbf{\Sigma}|) + \log\left(\frac{\Gamma(a_q)b_q^{a_q}}{\Gamma(a_p)b_p^{a_q}}\right) + (a_p-a_q)\Psi(a_p) - a_p + \frac{a_pb_p}{d\,b_q}\mathbb{E}[\Lambda] + (\tfrac{d}{2}-a_q)\mathbb{E}[\log(\Lambda)]. \tag{46}$$

The moments of Λ were computed in [47], but we are giving a simple derivation of the first moment below. It is known that the random variable $V_i = N_i^2/\sum_{j=1}^d N_j^2$ has the following beta distribution:

$$V_i \sim \mathrm{Beta}(\tfrac{1}{2}, \tfrac{d-1}{2}). \tag{47}$$

Since $\mathbb{E}[V_i] = 1/d$, then

$$\begin{aligned}\mathbb{E}[\Lambda] &= \mathbb{E}\left[\sum_{i=1}^d \lambda_i V_i\right] \\ &= \frac{\mathrm{tr}(\mathbf{\Sigma})}{d}.\end{aligned} \tag{48}$$

Expected logarithm of Λ can be expressed as a difference of two expectations:

$$\mathbb{E}[\log(\Lambda)] = \mathbb{E}\left[\log\left(\sum_{i=1}^d \lambda_i N_i^2\right)\right] - \mathbb{E}\left[\log\left(\sum_{i=1}^d N_i^2\right)\right]. \tag{49}$$

Using the fact that the expected logarithm of a chi-square random variable with d degrees of freedom is equal to $\Psi(d/2) + \log(2)$, $\mathbb{E}[\log(\Lambda)]$ can be computed by the following equation:

$$\mathbb{E}[\log(\Lambda)] = \mathbb{E}\left[\log\left(\sum_{i=1}^d \lambda_i N_i^2\right)\right] - \Psi(\tfrac{d}{2}) - \log(2). \tag{50}$$

With substitution (48) and (50) into (46), we get (42). □

3.3. KL-Divergence between ACG Distributions

The following theorem gives expressions for the KL-divergence between ACG distributions and the entropy of a single ACG distribution.

Theorem 4. *Suppose we have two probability distributions $G_{\mathbf{Y}} = \mathcal{ACG}(\mathbf{\Sigma}_1)$ and $J_{\mathbf{Y}} = \mathcal{ACG}(\mathbf{\Sigma}_2)$, then the entropy of random variable $\mathbf{Y}$ distributed according to $G_{\mathbf{Y}}$ and the KL-divergence between these two distributions are given by the following expressions:*

$$H(\mathbf{Y}) = \log\left(\frac{2\pi^{\frac{d}{2}}|\mathbf{\Sigma}_1|^{1/2}}{\Gamma(\frac{d}{2})}\right) - \tfrac{d}{2}\left(\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\sigma_i N_i^2\right)\right] - \Psi(\tfrac{d}{2}) - \log(2)\right), \tag{51}$$

$$KL(G||J) = -\tfrac{1}{2}\log(|\mathbf{\Sigma}|) + \tfrac{d}{2}\left(\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\lambda_i N_i^2\right)\right] - \Psi(\tfrac{d}{2}) - \log(2)\right), \tag{52}$$

where N_is are independent standard normal random variables, $\lambda_1, \ldots, \lambda_d$ are eigenvalues of matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_2^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1/2}$, and $\sigma_1, \ldots, \sigma_d$ are eigenvalues of matrix $\mathbf{\Sigma}_1$.

Proof. Due to the invariance property of KL-divergence under invertible change of variables, we use the change of variable $\mathbf{\Omega} = \mathbf{\Sigma}_1^{-1/2}\mathbf{Y}$. It is easy to verify that $\mathbf{\Omega}$ is distributed according to a zero-mean generalized elliptical distribution with identity covariance [44]. From the definition of KL-divergence given by (33), we have

$$\mathrm{KL}(G||J) = \mathbb{E}\left[\log\left(\frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}}\left(\mathbf{\Omega}^\top\mathbf{\Omega}\right)^{-\frac{d}{2}}\right)\right] - \mathbb{E}\left[\log\left(\frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}|\tilde{\mathbf{\Sigma}}|^{1/2}}\left(\mathbf{\Omega}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{\Omega}\right)^{-\frac{d}{2}}\right)\right],$$

where $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$. By some simplifications, it is immediate that

$$\mathrm{KL}(G||J) = \tfrac{1}{2}\log(|\tilde{\mathbf{\Sigma}}|) + \tfrac{d}{2}\,\mathbb{E}\left[\log\left(\frac{\mathbf{\Omega}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{\Omega}}{\mathbf{\Omega}^\top\mathbf{\Omega}}\right)\right]. \tag{53}$$

Since projecting any zero-mean generalized elliptical distribution (with identity covariance) on the unit sphere gives an ACG random variable (with identity covariance) [50], we can substitute $\mathbb{E}[\log(\mathbf{\Omega}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{\Omega}/\mathbf{\Omega}^\top\mathbf{\Omega})]$ with $\mathbb{E}[\log(\mathbf{X}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{X}/\mathbf{X}^\top\mathbf{X})]$, where the random vector $\mathbf{X}$ is distributed according to a multivariate normal distribution with identity covariance and zero mean. Because $\mathbf{X}^\top\tilde{\mathbf{\Sigma}}^{-1}\mathbf{X}$ is a CQF and $\mathbf{X}^\top\mathbf{X}$ is a chi-square random variable, we have

$$\mathrm{KL}(G||J) = \tfrac{1}{2}\log(|\tilde{\mathbf{\Sigma}}|) + \tfrac{d}{2}\left(\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\tilde{\lambda}_i^{-1}N_i^2\right)\right] - \Psi(\tfrac{d}{2}) - \log(2)\right), \tag{54}$$

where $\tilde{\lambda}_i$s are the eigenvalues of $\tilde{\mathbf{\Sigma}}$. Additionally, it is easy to verify that $|\mathbf{\Sigma}| = |\tilde{\mathbf{\Sigma}}|^{-1}$ and $\lambda_i = \tilde{\lambda}_i^{-1}$, therefore (52) holds.

Since one of the terms in the KL-divergence is equal to the minus entropy, we use our derived expression for the KL-divergence between ACG distributions to find a formula for the entropy of an ACG distribution. Define $S_{\mathbf{Y}} = \mathcal{ACG}(\mathbf{I})$, then the KL-divergence between $G_{\mathbf{Y}}$ and $S_{\mathbf{Y}}$ can be easily derived from the main definition (33):

$$\mathrm{KL}(G||S) = -\mathrm{H}(\mathbf{Y}) - \log\left(\frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}}\right), \tag{55}$$

where $\mathrm{H}(\mathbf{Y})$ is the entropy of the random variable $\mathbf{Y}$. Now, we compute the above KL-divergence using (52) which is

$$\mathrm{KL}(G||S) = -\tfrac{1}{2}\log(|\mathbf{\Sigma}_1|) + \tfrac{d}{2}\left(\mathbb{E}\left[\log\left(\sum_{i=1}^{d}\sigma_i N_i^2\right)\right] - \Psi\left(\tfrac{d}{2}\right) - \log(2)\right). \tag{56}$$

Equating the right-hand sides of (55) and (56) gives (51). □

The following corollary shows a relation between the KL-divergence of ACG distributions and the KL-divergence of ZEG distributions. It is an immediate consequence of Theorems 3 and 4.

Corollary 1. *The KL-divergence between two ACG distributions $\mathcal{ACG}(\mathbf{\Sigma}_1)$ and $\mathcal{ACG}(\mathbf{\Sigma}_2)$ is equal to the KL-divergence between two ZEG distributions $\mathcal{ZEG}(\mathbf{\Sigma}_1, a, b)$ and $\mathcal{ZEG}(\mathbf{\Sigma}_2, a, b)$, when $a \to 0$.*

4. Simulation Study

In Section 2, we proposed three different methods for computing the expected logarithm of the CQF given in (1). We assume the weights of CQFs that are used in the simulations are eigenvalues of some random positive definite matrices. These random matrices are generated uniformly from the space of positive definite matrices with unit trace according to the procedure proposed in [43]. In this section, we numerically investigate the running time and accuracy of these approaches. All methods were implemented in MATLAB (Version R2014a) (64-bit), and the simulations were run on a personal laptop with an Intel Core i5 (2.5Ghz) processor under the OS X Yosemite 10.10.3 operating system. Since the series method depends heavily on loops that are slow in MATLAB, we implemented this method in a MATLAB MEX-file. For the integral method, the integral is numerically evaluated using Gauss–Kronrod 7-15 rule [51,52]. The absolute error tolerance is given as an input parameter of the numerical integration. In the integral method, the value can be computed with any given accuracy by choosing the absolute error tolerance; therefore, we do not analyze the integral method in the sense of the calculation error.

Figure 2 investigates the effects of dimensionality on the average running time of different methods for computing the expected logarithm of the CQF explained in Section 2. For the integral method (upper-left plot), two curves for two different absolute error tolerances are shown. The integral formula (13) has the parameters k and θ that can be chosen freely, and we choose those given in (14). Different curves for the series method (upper-right plot) correspond to different values of L, which is the truncation length of the series. The curve of the fast method (lower plot) corresponds to the computation time of the third fast method explained in Section 2. One reason of lower computation time of the fast method is its lack of need for any eigenvalue computation. There is a curve in upper-right plot showing the computational time of eigenvalue computation.

The approximation error of all three fast methods for different dimensions can be seen in Figure 3. The plot on the right-hand side of this figure magnifies the curve for the mean error of the third fast method (the blue curve with dots). As it can be observed in Figure 3, changing the dimensionality has a negligible effect on the mean and the *standard deviation* (SD) of the absolute approximation error for the fast methods. Small mean error and SD of the third method indicate the distinct advantage of the third fast method over the other two methods. This method uses a convex combination of the values of the other two approximations as explained in Section 2.

Approximating the expected logarithm of the CQF using the fast methods induces an error on the KL-divergence between ACG distributions given by (52). Figure 4 shows the mean percentage of relative error and its standard deviation as a function of dimensionality. It can be observed that the relative error decreases as the dimensionality increases. The third fast method is clearly superior to the other two fast methods. The reason for such a small percentage of relative error is the observation that whenever the error is large, then the KL-divergence is large too. We are not showing the results for the dimensions less than ten because the error percentage is quite large in that regime.

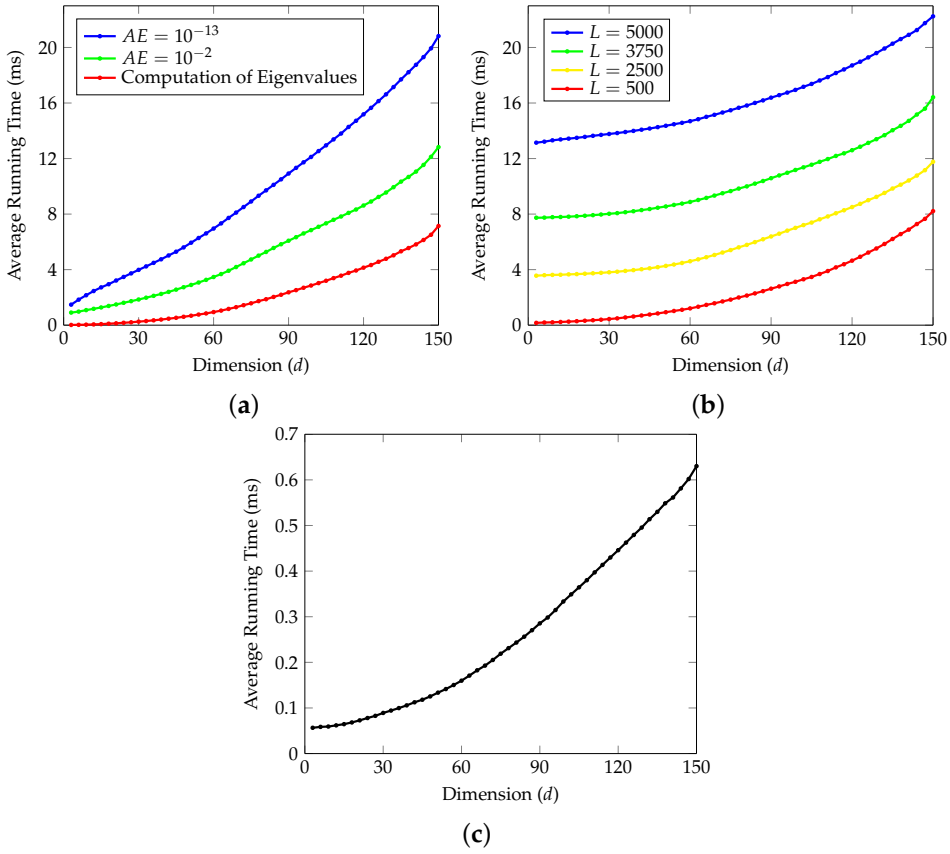

Figure 2. The average running time (in milliseconds) of the integral method (**a**), the series method (**b**) and the third fast method (**c**) in different dimensions for computing expected logarithm of the CQF. The red curve in the upper-left plot shows the computational time for computing the eigenvalues of random positive-definite matrices (using `eig` function in MATLAB) needed before applying the integral method or series method. Different curves for the upper-left plot correspond to the computational time of integral method for different absolute error tolerances including the time needed for computing the eigenvalues. Different curves for the series method correspond to the computational time for various values of truncation length of the series.

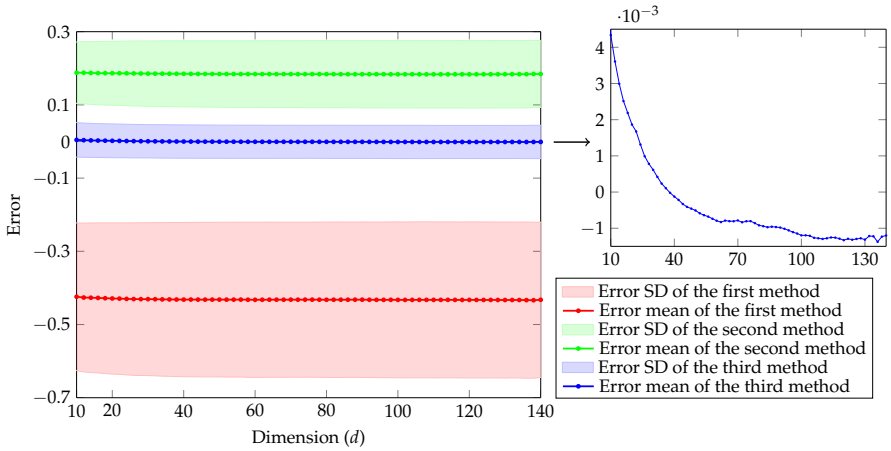

Figure 3. The absolute error for the approximation of the expected logarithm of the CQF by the fast methods explained in Section 2 for different dimensions. The third method uses a convex combination of the first two methods. The plot on the right shows the zoomed version of the error mean of the third method.

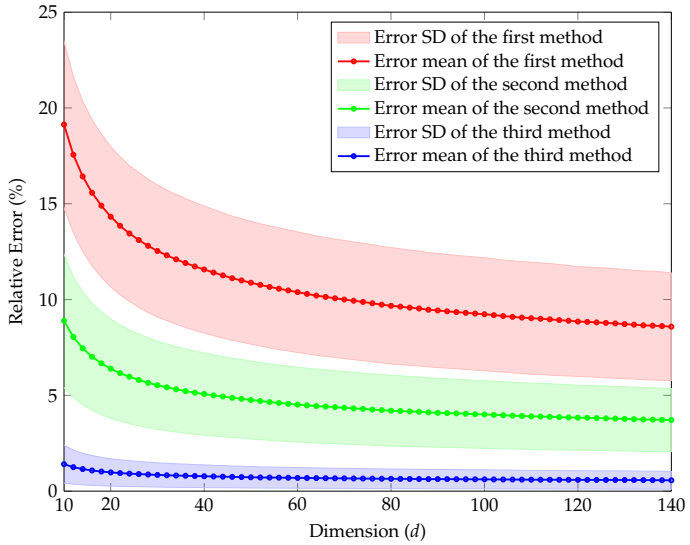

Figure 4. The relative error of the KL-divergence between ACG distributions in different dimensions when the fast methods are used for approximating the expected logarithm of the CQF.

The amount of error in the series method for a given truncation length L depends on the parameters d and ϵ, as it can be observed in the upper bound error formula (26). In Figure 5, this error together with its upper bound are shown for $\epsilon = 0.6$, $\epsilon = 0.8$ and $\epsilon = 0.95$ with $d = 2$. The curves are plotted for $L > d\epsilon/(2 - 2\epsilon)$, because the error upper bound is valid only for this region. It can be observed that by increasing ϵ, the slopes of the error and the error upper bound increase and the distance between these errors slightly increases too. This behavior indicates that the series method is more suitable for smaller ϵ values. For the case of $\epsilon = 0.6$, we can see that the real error remains constant after about $L = 60$, whereas the error upper bound still increases. It happens because the error is computed by the integral method, which has an error itself (in this case, the error is about 10^{-15}).

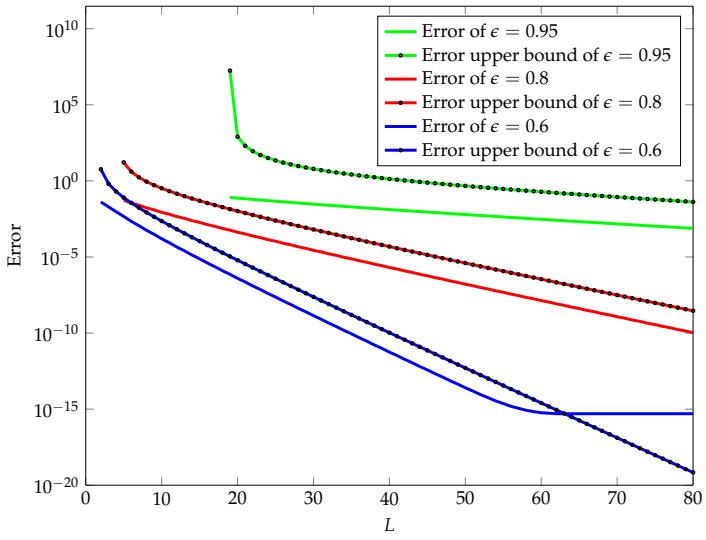

Figure 5. The approximation error of the series method for computing the expected logarithm of the CQF together with its upper bound. The result for three different values of ϵ is shown with different colors. In this figure, the number of weights (d) is equal to two.

Figure 6 shows how increasing dimension affects the performance of the series method. The parameter ϵ is set to 0.9 by choosing maximum and minimum weights in the CQF to be 1 and 1/19, respectively. The other weights of CQF are sampled uniformly between the maximum and minimum weights. It can be seen that the dimensionality has a negligible effect on the slope of the curves. This can be predicted from the formula of upper bound in (26), because the exponential term ϵ^L dominates other terms in the equation and the slopes of the curves are determined mainly by the parameter ϵ. In this figure, the standard deviations are due to the different distribution of the weights between the maximum and minimum weights. Figures 5 and 6 demonstrate that the error upper bound is a relatively tight bound for the actual error.

In Figure 7, we investigate the effect of ϵ and d on the averaged L to achieve an acceptable upper bound error (here 10^{-8}). We can see that as the amount of ϵ increases, the slopes of the curves increase and in the limit of $\epsilon \to 1$, it goes to infinity. This figure justifies our previous claim that when ϵ and the dimensionality are small, the series method is very efficient due to relatively small L needed to achieve an acceptable error.

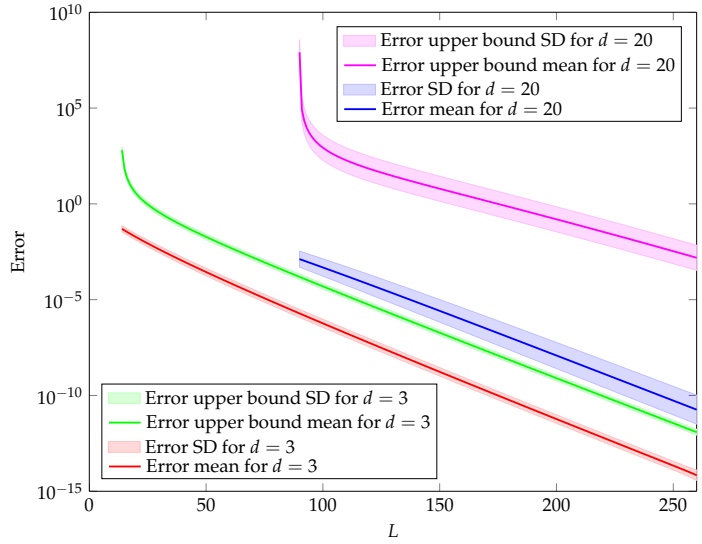

Figure 6. The relation between L and the error in the series method for $\epsilon = 0.9$.

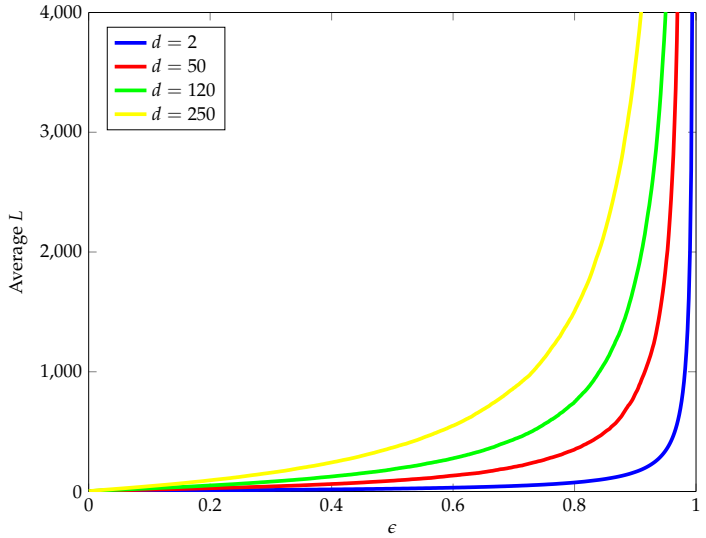

Figure 7. The average needed L computed by the error upper bound formula given by (26) for different values of d and ϵ.

5. Conclusions

In this paper, we developed three methods for calculating the expected logarithm of a central quadratic form. The integral method was a direct application of a more general result applicable for positive random variables. We then introduced three fast methods for approximating the expected logarithm. Finally, using an infinite series representation of central quadratic forms, we proposed a series method for computing the expected logarithm. By proving a bound for the approximation error, we investigated the performance of this method.

We also derived expressions for the entropy and the KL-divergence of zero-mean elliptical gamma and angular central Gaussian distributions. The expected logarithm of the central quadratic form appeared in the form of KL-divergences and entropy of the angular central Gaussian distribution.

By conducting multiple experiments, we observed that the three methods for computing the expected logarithm of a central quadratic form differ in running time and accuracy. The possible user can choose the most appropriate method based on his/her requirements.

The methodologies developed in this paper can be used in many applications. For example, one can use the result of Theorem 1 for computing the expected logarithm of other positive random variables like a non-central quadratic form. Another line of research would be to use the KL-divergence between angular central Gaussian distributions with the fast approximations in learning problems that have a divergence measure in their cost functions.

Acknowledgments: This research was in part supported by a grant from IPM (No. CS1395-4-42).

Author Contributions: The authors contributed equally to this work. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Lemma 2

From Lemma 1 and the form of gamma cumulative function, we have the following inequalities:

$$\frac{-\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\min}}\right)}{\Gamma\left(\frac{d}{2}\right)} \leq -F(x) \leq \frac{-\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\max}}\right)}{\Gamma\left(\frac{d}{2}\right)}, \tag{A1}$$

where $\lambda_{\max} = \max\{\lambda_i\}_{i=1}^d$, $\lambda_{\min} = \min\{\lambda_i\}_{i=1}^d$, and $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function. Adding $G(x)$ to all sides of the above inequality, we get

$$\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} - \frac{\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\min}}\right)}{\Gamma\left(\frac{d}{2}\right)} \leq G(x) - F(x) \leq \frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} - \frac{\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\max}}\right)}{\Gamma\left(\frac{d}{2}\right)}. \tag{A2}$$

Since $\log(x)$ is positive for $x > 1$, therefore by multiplying all sides of the above inequality by $\log(x)$, we obtain

$$\log(x)\left(\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} - \frac{\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\min}}\right)}{\Gamma\left(\frac{d}{2}\right)}\right) \leq \log(x)(G(x)-F(x)) \leq \log(x)\left(\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} - \frac{\gamma\left(\frac{d}{2},\frac{x}{2\lambda_{\max}}\right)}{\Gamma\left(\frac{d}{2}\right)}\right), \tag{A3}$$

which holds for all $x > 1$. For proving the first part of this lemma, namely

$$\lim_{x\to\infty} \log(x)(G(x)-F(x)) = 0, \tag{A4}$$

it is enough to show that

$$\lim_{x\to\infty} \log(x)\left(\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)} - \frac{\gamma\left(\hat{k},\frac{x}{\hat{\theta}}\right)}{\Gamma(\hat{k})}\right) = 0 \tag{A5}$$

holds for any positive choices of k, $\hat{k}$, θ, and $\hat{\theta}$ and then invoke squeeze theorem by taking the limits of all sides of (A3). From the definition of lower incomplete gamma function, the left-hand side (A5) can be rewritten as

$$\lim_{x\to\infty} \frac{\frac{1}{\Gamma(k)}\int_0^{\frac{x}{\theta}} t^{k-1}\exp(-t)\,dt - \frac{1}{\Gamma(\hat{k})}\int_0^{\frac{x}{\hat{\theta}}} t^{\hat{k}-1}\exp(-t)\,dt}{\log(x)^{-1}}.$$

Using L'Hôpital's rule, it can be seen that the above limit is equivalent to

$$\lim_{x\to\infty} x\log(x)^2\left(\frac{1}{\theta\Gamma(k)}\left(\frac{x}{\theta}\right)^{k-1}\exp\left(-\frac{x}{\theta}\right) - \frac{1}{\hat{\theta}\Gamma(\hat{k})}\left(\frac{x}{\hat{\theta}}\right)^{\hat{k}-1}\exp\left(-\frac{x}{\hat{\theta}}\right)\right). \tag{A6}$$

It is easy to see that (A6) is equal to zero and consequently (A5) and (A4) hold.

Now, we want to prove the second statement in the lemma, that is

$$\lim_{x\to 0^+} \log(x)(G(x)-F(x)) = 0. \tag{A7}$$

If we multiply all sides of (A2) by $\log(x)$, then for $0 < x < 1$ we have

$$\log(x)\left(\frac{\gamma(k,\frac{x}{\theta})}{\Gamma(k)} - \frac{\gamma(\frac{d}{2},\frac{x}{2\lambda_{\max}})}{\Gamma(\frac{d}{2})}\right) \le \log(x)(G(x)-F(x)) \le \log(x)\left(\frac{\gamma(k,\frac{x}{\theta})}{\Gamma(k)} - \frac{\gamma(\frac{d}{2},\frac{x}{2\lambda_{\min}})}{\Gamma(\frac{d}{2})}\right). \tag{A8}$$

Using the same strategy as above, we want to show that for any positive choices of k, $\hat{k}$, θ, and $\hat{\theta}$, the following limit holds:

$$\lim_{x\to 0^+} \log(x)\left(\frac{\gamma(k,\frac{x}{\theta})}{\Gamma(k)} - \frac{\gamma(\hat{k},\frac{x}{\hat{\theta}})}{\Gamma(\hat{k})}\right) = 0. \tag{A9}$$

Using L'Hôpital's rule, it can be seen that

$$\lim_{x\to 0^+} \log(x)\left(\frac{\gamma(k,\frac{x}{\theta})}{\Gamma(k)}\right) \overset{\text{H}}{=} \lim_{x\to 0^+} x\log(x)^2\left(\frac{1}{\theta\Gamma(k)}(\frac{x}{\theta})^{k-1}\exp(-\frac{x}{\theta})\right) = 0.$$

Therefore (A9) holds, and from (A8), we have

$$0 \le \lim_{x\to 0^+} \log(x)(G(x)-F(x)) \le 0. \tag{A10}$$

By squeeze theorem, we can conclude that (A7) holds.

Appendix B. Proof of Lemma 3

From the expression of $\mathcal{F}$ and $\mathcal{G}$, we have

$$\int_0^\infty \frac{\mathcal{G}(\sigma)-\mathcal{F}(\sigma)}{\sigma}d\sigma = \int_0^\infty \frac{\prod_{i=1}^d (1+2\lambda_i\sigma)^{\frac{1}{2}} - (1+\theta\sigma)^k}{\sigma(1+\theta\sigma)^k\prod_{i=1}^d (1+2\lambda_i\sigma)^{\frac{1}{2}}}d\sigma. \tag{B1}$$

In this proof, for the simplicity of notation, we define $\mathcal{L}(\sigma) = (\mathcal{G}(\sigma)-\mathcal{F}(\sigma))/\sigma$ and $\mathcal{V}(\sigma) = \prod_{i=1}^d (1+2\lambda_i\sigma)^{1/2} - (1+\theta\sigma)^k$.

We give separate proofs for the cases $d > 2k$, $d < 2k$ and $d = 2k$. For the first case $d > 2k$, we have

$$\begin{aligned}\lim_{\sigma\to\infty} \mathcal{V}(\sigma) &= \lim_{\sigma\to\infty} \left[\prod_{i=1}^{d}(1+2\lambda_i\sigma)^{\frac{1}{2}} - (1+\theta\sigma)^k\right] \\ &= +\infty.\end{aligned}$$

Consequently, it can be said that there exists a number $a > 0$ that for all $x \geq a$, the function $\mathcal{V}(\sigma)$ is positive.

Therefore, the integrand of $\int_a^\infty \mathcal{L}(\sigma)d\sigma$ is positive in its domain of integration. If we choose $1 < p < 1 + k$, then

$$\lim_{\sigma\to\infty} \frac{\mathcal{L}(\sigma)}{\frac{1}{\sigma^p}} = 0. \tag{B2}$$

Since $\int_a^\infty \frac{1}{\sigma^p}d\sigma$ is convergent and its integrand is positive in its domain, from the limit comparison test, it follows that the integral $\int_a^\infty \mathcal{L}(\sigma)d\sigma$ is convergent.

Now, we want to show that the integral $\int_0^a \mathcal{L}(\sigma)d\sigma$ is also convergent. Since $\mathcal{G}(\sigma) - \mathcal{F}(\sigma)$ is bounded for $\sigma \in \mathbb{R}^+$, and

$$\lim_{\sigma\to 0^+} \frac{\mathcal{G}(\sigma) - \mathcal{F}(\sigma)}{\sigma} = \sum_{i=1}^{d}\lambda_i - k\theta, \tag{B3}$$

then $\mathcal{L}(\sigma)$ is bounded and consequently $\int_0^a \mathcal{L}(\sigma)d\sigma$ is convergent.

Since $\int_0^a \mathcal{L}(\sigma)d\sigma$ and $\int_a^\infty \mathcal{L}(\sigma)d\sigma$ are convergent, the integral $\int_0^\infty \mathcal{L}(\sigma)d\sigma$ is also convergent.

For the case of $2k > d$, we have

$$\begin{aligned}\lim_{\sigma\to\infty} -\mathcal{V}(\sigma) &= \lim_{\sigma\to\infty} \left[(1+\theta\sigma)^k - \prod_{i=1}^{d}(1+2\lambda_i\sigma)^{\frac{1}{2}}\right] \\ &= +\infty.\end{aligned}$$

Therefore, there exists a number $a > 0$ that for all $x \geq a$, the function $-\mathcal{V}(\sigma)$ is positive. Therefore, the integrand of $\int_a^\infty -\mathcal{L}(\sigma)d\sigma$ is positive in its domain of integration. If we choose $1 < p < 1 + d/2$, then

$$\lim_{\sigma\to\infty} \frac{-\mathcal{L}(\sigma)}{\frac{1}{\sigma^p}} = 0. \tag{B4}$$

Knowing that $\int_a^\infty \frac{1}{\sigma^p}d\sigma$ is bounded, using limit comparison test, we can conclude that $\int_a^\infty -\mathcal{L}(\sigma)d\sigma$ is convergent. Now, with the same strategy as the previous case, we can show that the integral $\int_0^a -\mathcal{L}(\sigma)d\sigma$ is convergent and it is easy to see that $\int_0^\infty \mathcal{L}(\sigma)d\sigma$ is also convergent.

For $2k = d$, excluding the obvious case $\mathcal{G}(\sigma) = \mathcal{F}(\sigma)$, there exists a number $a > 0$ that for all $x \geq a$, the function $\mathcal{V}(\sigma)$ is either positive or negative. If it is positive, then we use the proof strategy for the case $d > 2k$. Otherwise, we exploit the strategy for the case $d < 2k$.

Appendix C. Proof of Theorem 2

We first give a proof for the inequality $\mathbb{E}[\log(U)] \leq \Psi(d/2) + \log(2\mathrm{tr}(\mathbf{\Sigma})/d)$. Using integral formula (13) with $k = d/2$ and $\theta = 2\mathrm{tr}(\mathbf{\Sigma})/d$ for $\mathbb{E}[\log(U)]$, we obtain

$$\int_0^\infty \frac{\left(1+\frac{2\mathrm{tr}(\mathbf{\Sigma})}{d}\sigma\right)^{-\frac{d}{2}} - \prod_{i=1}^{d}\left(1+2\lambda_i\sigma\right)^{-\frac{1}{2}}}{\sigma}d\sigma \leq 0. \tag{C1}$$

For proving this inequality, it is enough to show

$$\left(1+\frac{2\mathrm{tr}(\mathbf{\Sigma})}{d}\sigma\right)^{-\frac{d}{2}}-\prod_{i=1}^{d}(1+2\lambda_i\sigma)^{-\frac{1}{2}}\leq 0,\quad \text{for all } \sigma\in\mathbb{R}^+.$$

Assume $a_i=1+2\lambda_i\sigma$, then we can rewrite the above inequality as

$$\left(\prod_{i=1}^{d}a_i\right)^{\frac{1}{d}}\leq\frac{\sum_{i=1}^{d}a_i}{d},\quad \text{for all } a_i\in[1,+\infty), \tag{C2}$$

which is the well-known *arithmetic mean-geometric mean* inequality.

For proving the second inequality, $\mathbb{E}[\log(U)]\geq\Psi(\mathrm{tr}(\mathbf{\Sigma})^2/2\mathrm{tr}(\mathbf{\Sigma}^2))+\log(2\mathrm{tr}(\mathbf{\Sigma}^2)/\mathrm{tr}(\mathbf{\Sigma}))$, we use the integral formula in (13) with $k=\mathrm{tr}(\mathbf{\Sigma})^2/(2\mathrm{tr}(\mathbf{\Sigma}^2))$ and $\theta=2\mathrm{tr}(\mathbf{\Sigma}^2)/\mathrm{tr}(\mathbf{\Sigma})$ to obtain

$$\int_0^\infty\frac{\left(1+\frac{2\mathrm{tr}(\mathbf{\Sigma}^2)}{\mathrm{tr}(\mathbf{\Sigma})}\sigma\right)^{-\frac{\mathrm{tr}(\mathbf{\Sigma})^2}{2\mathrm{tr}(\mathbf{\Sigma}^2)}}-\prod_{i=1}^{d}\left(1+2\lambda_i\sigma\right)^{-\frac{1}{2}}}{\sigma}d\sigma\geq 0. \tag{C3}$$

For proving the above inequality, it is enough to show

$$\prod_{i=1}^{d}(1+2\lambda_i\sigma)\geq\left(1+\frac{2\mathrm{tr}(\mathbf{\Sigma}^2)}{\mathrm{tr}(\mathbf{\Sigma})}\sigma\right)^{\frac{\mathrm{tr}(\mathbf{\Sigma})^2}{\mathrm{tr}(\mathbf{\Sigma}^2)}},\quad \text{for all } \sigma\in\mathbb{R}^+. \tag{C4}$$

Assume $A(\sigma)=\prod_{i=1}^{d}(1+2\lambda_i\sigma)$, $B(\sigma)=(1+2\mathrm{tr}(\mathbf{\Sigma}^2)\sigma/\mathrm{tr}(\mathbf{\Sigma}))^{\mathrm{tr}(\mathbf{\Sigma})^2/\mathrm{tr}(\mathbf{\Sigma}^2)}$ and $P(\sigma)=A(\sigma)/B(\sigma)$, then (C4) can be rewrited as

$$P(\sigma)\geq 1,\quad \text{for all } \sigma\in\mathbb{R}^+. \tag{C5}$$

It is easy to see that $P(0)=1$. Therefore, for proving (C5), it is enough to show that the function $P(\sigma)$ is increasing for positive sigmas.

The derivative of the function $P(\sigma)$ is

$$P'(\sigma)=\frac{A(\sigma)'B(\sigma)-B(\sigma)'A(\sigma)}{B^2(\sigma)}.$$

We want to show $P'(\sigma)\geq 0$, for all positive sigmas, and it is equivalent to say

$$\frac{A'(\sigma)}{A(\sigma)}\geq\frac{B'(\sigma)}{B(\sigma)},\quad \text{for all } \sigma\in\mathbb{R}^+. \tag{C6}$$

Computing the left and right sides of the above inequality, we obtain

$$\sum_{i=1}^{d}\frac{2\lambda_i}{1+2\lambda_i\sigma}\geq\frac{2\mathrm{tr}(\mathbf{\Sigma})^2}{\mathrm{tr}(\mathbf{\Sigma})+2\mathrm{tr}(\mathbf{\Sigma})\sigma},\quad \text{for all } \sigma\in\mathbb{R}^+. \tag{C7}$$

We can rewrite the above inequality as

$$\left(\sum_{i=1}^{d}\frac{2\lambda_i}{1+2\lambda_i\sigma}\right)\left(\sum_{i=1}^{d}\lambda_i(1+2\lambda_i\sigma)\right)\geq\left(\sum_{i=1}^{d}\lambda_i\right)^2,\quad \text{for all } \sigma\in\mathbb{R}^+.$$

If we assume $x_i^2 = 2\lambda_i/(1+2\lambda_i\sigma)$ and $y_i^2 = \lambda_i(1+2\lambda_i\sigma)$, we can rewrite the above inequality as

$$\left(\sum_{i=1}^{d} x_i^2\right)\left(\sum_{i=1}^{d} y_i^2\right) \geq \left(\sum_{i=1}^{d} x_i y_i\right)^2, \quad \text{for all } \{x_i, y_i\} \in \mathbb{R}^+, \tag{C8}$$

which is the Cauchy-Schwarz inequality. So the function $P(\sigma)$ is increasing and consequently, the second inequality holds.

Appendix D. Proof of Lemma 4

As we can see in [14], the following bound exists for c_i:

$$|c_i| \leq c_0 \epsilon^i \frac{\Gamma(\frac{d}{2}+i)}{\Gamma(\frac{d}{2})i!}, \tag{D1}$$

where $\epsilon = \max_j |1-\beta\lambda_j^{-1}|$. If we put our chosen β given by (23) in ϵ formula, we obtain

$$\epsilon = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}. \tag{D2}$$

Consider $\hat{c}_i$ to be the corresponding bound for c_i,

$$\hat{c}_i = c_0 \epsilon^i \frac{\Gamma(\frac{d}{2}+i)}{\Gamma(\frac{d}{2})i!},$$

then we have

$$\frac{\hat{c}_{i+1}}{\hat{c}_i} = \frac{\frac{d}{2}+i}{i}\epsilon.$$

Since $(d/2+i)/i$ decreases with increasing i, we have

$$\frac{\hat{c}_{i+k}}{\hat{c}_i} \leq \Bigg(\underbrace{\frac{\frac{d}{2}+i}{i}\epsilon}_{\epsilon_i}\Bigg)^k. \tag{D3}$$

By summing up the both sides of above inequality and changing the variables, we can conclude

$$\begin{aligned} \sum_{k=i+1}^{\infty} \hat{c}_k &\leq \hat{c}_{i+1} \sum_{k=0}^{\infty} \epsilon_{i+1}^k \\ &\leq \hat{c}_{i+1} \sum_{k=0}^{\infty} \epsilon_i^k \\ &= \hat{c}_{i+1} \frac{1}{1-\epsilon_i}, \end{aligned} \tag{D4}$$

which is true if i is large enough such that $\epsilon_i < 1$. Since $L > d\epsilon/(2-2\epsilon)$, it can be observed that $\epsilon_i < 1$ for $i \geq L$, hence for the total approximation error, we obtain

$$\begin{aligned}
\mathbb{E}[\log U] - \hat{\mathbb{E}}[\log U] &= \sum_{i=L}^{\infty}\left(\frac{1}{\frac{d}{2}+i}\sum_{k=i+1}^{\infty}\hat{c}_k\right) \\
&\leq \sum_{i=L}^{\infty}\frac{1}{\frac{d}{2}+i}\hat{c}_{i+1}\frac{1}{1-\epsilon_i} \\
&\leq \frac{1}{(\frac{d}{2}+L)(1-\epsilon_L)}\sum_{i=L}^{\infty}\hat{c}_{i+1} \\
&\leq \frac{\hat{c}_{L+1}}{(\frac{d}{2}+L)(1-\epsilon_L)^2} \\
&\leq c_0\frac{\epsilon^{L+1}}{\left(1-\frac{\frac{d}{2}+L}{L}\epsilon\right)^2}\frac{\Gamma(\frac{d}{2}+L)}{\Gamma(\frac{d}{2}+1)(L+1)!}.
\end{aligned} \tag{D5}$$

References

1. Lapidoth, A.; Moser, S.M. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Trans. Inf. Theory* **2003**, *49*, 2426–2467.
2. Khodabin, M.; Ahmadabadi, A. Some properties of generalized gamma distribution. *Math. Sci.* **2010**, *4*, 9–28.
3. Eccardt, T.M. The use of the logarithm of the variate in the calculation of differential entropy among certain related statistical distributions. **2007**, arXiv:0705.4045.
4. Nicolas, J.M. Introduction to second kind statistics: Application of log-moments and log-cumulants to SAR image law analysis. *Trait. Signal* **2002**, *19*, 139–167.
5. Nicolas, J.M.; Tupin, F. Gamma mixture modeled with "second kind statistics": Application to SAR image processing. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Toronto, ON, Canada, 24–28 June 2002; Volume 4, pp. 2489–2491.
6. Teh, Y.W.; Newman, D.; Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 1353–1360.
7. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
8. Jean, W.H. The geometric mean and stochastic dominance. *J. Financ.* **1980**, *35*, 151–158.
9. Hakansson, N.H. Multi-period mean-variance analysis: Toward a general theory of portfolio choice. *J. Financ.* **1971**, *26*, 857–884.
10. Lo, Y.; Mendell, N.R.; Rubin, D.B. Testing the number of components in a normal mixture. *Biometrika* **2001**, *88*, 767–778.
11. Moore, D.S.; Spruill, M.C. Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann. Stat.* **1975**, *3*, 599–616.
12. Li, K.C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **1991**, *86*, 316–327.
13. Ruben, H. Probability content of regions under spherical normal distributions, IV: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *Ann. Math. Stat.* **1962**, *33*, 542–570.
14. Kotz, S.; Johnson, N.L.; Boyd, D.W. Series representations of distributions of quadratic forms in normal variables. I. Central case. *Ann. Math. Stat.* **1967**, *38*, 823–837.
15. Box, G.E. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Stat.* **1954**, *25*, 290–302.
16. Ha, H.T.; Provost, S.B. An accurate approximation to the distribution of a linear combination of non-central chi-square random variables. *REVSTAT Stat. J.* **2013**, *11*, 231–254.
17. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
18. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

19. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633.
20. Kanamori, T. Scale-invariant divergences for density functions. *Entropy* **2014**, *16*, 2611–2628.
21. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.
22. Pardo, L. *Statistical Inference Based on Divergence Measures*; CRC Press: London, UK, 2005.
23. Blekas, K.; Lagaris, I.E. Split–Merge Incremental LEarning (SMILE) of mixture models. In *Artificial Neural Networks–ICANN 2007*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4669, pp. 291–300.
24. Runnalls, A.R. Kullback–Leibler approach to Gaussian mixture reduction. *IEEE Trans. Aerosp. Electron. Syst.* **2007**, *43*, 989–999.
25. Dhillon, I.S.; Mallela, S.; Kumar, R. A divisive information theoretic feature clustering algorithm for text classification. *J. Mach. Learning Res.* **2003**, *3*, 1265–1287.
26. Imseng, D.; Bourlard, H.; Garner, P.N. Using Kullback–Leibler divergence and multilingual information to improve ASR for under-resourced languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4869–4872.
27. Do, M.N.; Vetterli, M. Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance. *IEEE Trans. Image Process.* **2002**, *11*, 146–158.
28. Mathiassen, J.R.; Skavhaug, A.; Bø, K. Texture Similarity Measure Using Kullback–Leibler Divergence Between Gamma Distributions. In *Computer Vision—ECCV 2002*; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2352; pp. 133–147.
29. Koutras, M. On the generalized noncentral chi-squared distribution induced by an elliptical gamma law. *Biometrika* **1986**, *73*, 528–532.
30. Fang, K.T.; Zhang, Y.T. *Generalized Multivariate Analysis*; Springer: Berlin/Heidelberg, Germany, 1990.
31. Hosseini, R.; Sra, S.; Theis, L.; Bethge, M. Inference and mixture modeling with the elliptical gamma distribution. *Comput. Stat. Data Anal.* **2016**, *101*, 29–43.
32. Watson, G.S. *Statistics on Spheres*; Wiley: New York, NY, USA, 1983.
33. Kent, J.T. The complex Bingham distribution and shape analysis. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 285–299.
34. Bethge, M.; Hosseini, R. Method and Device for Image Compression. U.S. Patent 8,750,603, 10 June 2014.
35. Zhang, T. Robust subspace recovery by Tyler's M-estimator. *Inf. Inference* **2015**, *5*, doi:10.1093/imaiai/iav012.
36. Franke, J.; Redenbach, C.; Zhang, N. On a mixture model for directional data on the sphere. *Scand. J. Stat.* **2015**, *43*, 139–155.
37. Tyler, D.E. A distribution-free M-estimator of multivariate scatter. *Ann. Stat.* **1987**, *15*, 234–251.
38. Sra, S.; Hosseini, R.; Theis, L.; Bethge, M. Data modeling with the elliptical gamma distribution. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 903–911.
39. Davis, P.J.; Rabinowitz, P. *Methods of Numerical Integration*; Dover: New York, NY, USA, 2007.
40. Benaroya, H.; Han, S.M.; Nagurka, M. *Probability Models in Engineering and Science*; CRC Press: Boca Raton, FL, USA, 2005; Volume 193.
41. Satterthwaite, F.E. Synthesis of variance. *Psychometrika* **1941**, *6*, 309–316.
42. Yuan, K.H.; Bentler, P.M. Two simple approximations to the distributions of quadratic forms. *Br. J. Math. Stat. Psychol.* **2010**, *63*, 273–291.
43. Mittelbach, Martin, B.M.; Jorswieck, E. Sampling uniformly from the set of positive definite matrices with trace constraint. *IEEE Trans. Signal Process.* **2012**, *60*, 2167–2179.
44. Frahm, G.; Jaekel, U. Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance. Available online: http://ssrn.com/abstract=1287683 (accessed on 26 July 2016).
45. Fang, K.T.; Kotz, S.; Ng, K.W. *Symmetric Multivariate and Related Distributions*; Chapman and Hall: London, UK, 1990.
46. Chen, B.; Zhu, Y.; Hu, J.; Principe, J.C. *System Parameter Identification: Information Criteria And Algorithms*; Newnes: Oxford, UK, 2013.
47. Provost, S.B.; Cheong, Y. The probability content of cones in isotropic random fields. *J. Multivar. Anal.* **1998**, *66*, 237–254.

48. Johnson, N.L.; Kotz, S. *Distributions in Statistics: Continuous Univariate Distributions*; Houghton Mifflin: Boston, MA, USA, 1970; Volume 1.
49. Polyanin, A.D.; Manzhirov, A.V. *Handbook of Integral Equations*; CRC Press: Boca Raton, FL, USA, 1998.
50. Soloveychik, I.; Wiesel, A. Tyler's estimator performance analysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015, pp. 5688–5692.
51. Piessens, R.; de Doncker-Kapenga, E.; Überhuber, C.W. *QUADPACK, A Subroutine Package for Automatic Integration*; Springer: Berlin/Heidelberg, Germany, 1983.
52. Shampine, L.F. Vectorized adaptive quadrature in Matlab. *J. Comput. Appl. Math.* **2008**, *211*, 131–140.