Abstract

This paper studies mixture modeling using the Elliptical Gamma distribution (EGD)—a distribution that has parametrized tail and peak behaviors and offers richer modeling power than the multivariate Gaussian. First, we study maximum likelihood (ML) parameter estimation for a single EGD, a task that involves nontrivial conic optimization problems. We solve these problems by developing globally convergent fixed-point methods for them. Next, we consider fitting mixtures of EGDs, for which we first derive a closed-form expression for the KL-divergence between two EGDs and then use it in a “split-and-merge” expectation maximization algorithm. We demonstrate the ability of our proposed mixture modelling in modelling natural image patches.

1 Introduction

Statistical models in modern data analysis often involve non-Gaussian distributions: whether to capture manifold structure (Banerjee et al., 2005; Pennec, 2006; Chikuse, 2003), to model sparsity (Kotz et al., 2001; Seeger and Nickisch, 2011), to express heavy or light tailed behavior (Ollila et al., 2011; Kotz et al., 2001), to study independence (Lee et al., 1999; Hyvärinen, 1999), or to model a variety of other situations. The focus of this paper is on unsupervised learning with a particular non-Gaussian distribution, namely, the Elliptical Gamma (EG) distribution (Koutras, 1986). The EG density (the mean-zero case) with covariance matrix $\Sigma \succ 0$ is given by

$$p_{eg}(x; \Sigma, a, b) := \frac{\Gamma(q/2)}{\pi^{q/2}\Gamma(a)b^{a}} |\Sigma|^{1/2} (x^\top \Sigma^{-1} x)^{a-q/2} \exp\left(-b^{-1} x^\top \Sigma^{-1} x\right), \quad x \in \mathbb{R}^q,$$  \hspace{1cm} (1.1)

where $\Gamma$ is the usual Gamma function, and $a, b > 0$ are density shaping parameters (Fang and Zhang, 1990). Observe how (1.1) generalizes the Gaussian density, which corresponds to $a = q/2$; it reshapes the Gaussian density by including an additional factor $(x^\top \Sigma^{-1} x)^{a-q/2}$ that encodes different tail and peak behavior—Figure 3 provides an illustration. For the case $a < q/2$, we are able to prove that the distribution can be written as scale mixture of Gaussian (see Appendix A).

1.1 Motivation

Our motivation for studying EGDs stems from their broad applicability and rich modeling power; a mixture of zero-mean EGDs can approximate any symmetric distribution (Fang and Zhang, 1990).
Figure 1: EG density on $\mathbb{R}^2$ with shape parameter $a = 1/3, 1, \text{ and } 3 \text{ (from left to right). All displayed densities have equal covariances; the density corresponding to } a = q/2 = 1 \text{ (middle) is a Gaussian density.}$

particular, EGDs form a subclass of *Elliptically Contoured Distributions* (ECDs), which have been successfully used for multivariate density estimation (Ollila et al., 2011), Bayesian statistical data modeling (Arellano-Valle et al., 2006), signal denoising (Tan and Jiao, 2007), financial data modeling (Bingham and Kiesel, 2002), and pattern recognition (Theiler et al., 2010). Likewise, mixtures of ECDs have been used successfully in many applications like robust statistical modeling (Lange et al., 1989), denoising (Rabbani and Vafadust, 2008), signal processing, among others—the survey (Ollila et al., 2011) cites several more applications and references.

We consider the following two aspects of EGDs in this paper: (i) algorithms for efficient maximum likelihood (ML) parameter estimation; and (ii) theory and algorithms for calculating KL-divergence between EGDs.

A further motivation is the potential of opening of research directions such as robust recovery of multiple subspaces—see e.g., (Lerman et al., 2011). This topic has applications in various fields like unsupervised learning, computer vision and biomedical engineering—see (Soltanolkotabi et al., 2012) and references therein.

We hope that this paper encourages others to also apply EGDs for modeling data.

### 1.2 Contributions

ML estimation for EGDs is challenging, either due to nonconvexity or due to the positive-definiteness constraint $\Sigma \succeq 0$ and nonlinearity; no closed-form estimates exist (unlike the Gaussian case). We develop algorithms and theory that makes it numerically easy to find ML estimates. It is hoped that our algorithms, which we have implemented in a user friendly toolbox\(^1\), make EGDs an appealing alternative to other density models. As a proof-of-concept, we show a brief application to mixture modeling with EGDs within the context of modeling natural images.

More precisely, this paper makes the following contributions:

- It develops two new non-Euclidean fixed point algorithms for ML parameter estimation with EGDs; these algorithms handle the convex ($a \geq q/2$) yet numerically challenging case, as well as the non-convex ($a < q/2$) case, which, despite its nonconvexity is still efficiently solved to global optimality.
- Empirical results demonstrating the substantial computational gains obtained by our fixed-point algorithms, compared against state-of-the-art manifold optimization techniques.
- A derivation of the KL-divergence between two EGDs.

\(^1\)Please see: [http://suvrit.de/work/soft/meg.html](http://suvrit.de/work/soft/meg.html)
2 Background

We begin by recalling basics of ECDs (of which EGDs are a special subclass). A q-dimensional random vector $X$ is distributed according to an ECD with a mean parameter $\mu \in \mathbb{R}^q$ and “scatter” matrix $\Sigma \in \mathbb{R}^{q \times q}$ (i.e. $X \sim \text{ECD}(\mu, \Sigma, f)$) if its characteristic function is of the form $\Phi_X(t) = \exp(i t^\top \mu f(t^\top \Sigma t)$, for some function $f: \mathbb{R}_+ \to \mathbb{R}$. If it exists, the density of an ECD has the form

$$p_X(x) = |\Sigma|^{-1/2} f((x - \mu)^\top \Sigma^{-1} (x - \mu)).$$

For simplicity, we consider mean-zero ECDs, that is, $\mu = 0$, so that

$$p_X(x) = |\Sigma|^{-1/2} f(x^\top \Sigma^{-1} x).$$

Under this assumption, one can factor $X$ into a uniform hyperspherical component and a scaled-radial component, so that $X = \Sigma^{1/2} RU$ with $U$ uniformly distributed over the unit sphere $S^{n-1}$ and $R$ a univariate random variable given by $R = \|\Sigma^{-1/2} X\|_2$ (Fang et al., 1990). The random variable $R$ has the p.d.f.

$$p_R(r) := (2\pi)^{q/2} r^{q-1}/\Gamma(q/2).$$

An EGD is an ECD whose squared radial component $Y = R^2$ is Gamma distributed, according to the following density

$$p_Y(v) = v^{a-1} \Gamma(a)^{-1} b^{-a} \exp(-v/b),$$

where $a$ is a shape parameter and $b$ is a scale parameter.

Using (2.2) as the radial distribution, we obtain the density shaping function $f$ for (2.1); therewith, we obtain the EGD density shown in (1.1). If $\Sigma$ equals the covariance matrix of the distribution, i.e., $\Sigma = \mathbb{E}\{XX^\top\}$, then $b = q/a$ (see (Fang et al., 1990, Equation 2.16)).

Calculating maximum likelihood estimates of the parameters of an ECD is generally not analytically possible, though in special cases such as multivariate t-distributions, a recursive algorithm has been proposed to calculate the parameters (Lange et al., 1989). For a review of ML estimation of ECDs see Ollila et al. (2011) and references therein, as well as some more recent works (Sra and Hosseini, 2013a; Zhang et al., 2013).
However, for EGDs, we can derive efficient maximum likelihood parameter estimation procedures; these are presented in the next section.

3 Maximum likelihood parameter estimation

Let \( \{x_1, \ldots, x_n\} \) be i.i.d. samples from an EGD. Then, their log-likelihood is

\[
\ell(a, b, \Sigma; x_1, \ldots, x_n) := n \log(C) - \frac{q}{2} \log|\Sigma| + (a - \frac{q}{2}) \sum_{i=1}^{n} \log(x_i^\top \Sigma^{-1} x_i) - \frac{1}{b} \sum_{i=1}^{n} x_i^\top \Sigma^{-1} x_i. \tag{3.1}
\]

Below we consider maximum likelihood (ML) estimation of \( \Sigma \) when \( a, b \) are held fixed.

ML estimation splits naturally into two cases depending on \( a \):

(i) \textbf{Concave.} Here \( a \geq q/2 \), whereby \( \ell \) is seen to be (strictly) concave function in \( \Sigma^{-1} \) (not in \( \Sigma \)).

(ii) \textbf{Non-concave.} Here \( a < q/2 \), whereby the second term in (3.1) is no longer concave, hence \( \ell \) is not concave.

Clearly, if \( \ell \) is strictly concave and if it attains its maximum, then this must be unique. More remarkably, even when \( \ell \) is not concave, it turns out that \( \ell \) possesses enough geometric structure that uniqueness still holds (see end of this section).

To maximize (3.1), since the constraint \( \Sigma \succ 0 \) is an open set, a necessary optimality condition is \( \nabla_{\Sigma} \ell_n = 0 \) — since (3.1) has a unique global maximum, if this equation has a unique positive definite solution, then barring a pathological situation, this solution is the desired ML estimate. Differentiating (3.1), we thus obtain

\[
-\frac{q}{2} \Sigma^{-1} - (a - \frac{q}{2}) \sum_{i=1}^{n} \frac{\Sigma^{-1} x_i x_i^\top \Sigma^{-1}}{x_i^\top \Sigma^{-1} x_i} + \frac{1}{b} \sum_{i=1}^{n} \Sigma^{-1} x_i x_i^\top \Sigma^{-1} = 0. \tag{3.2}
\]

Adding \( \frac{q}{2} \Sigma^{-1} \) to and rescaling (3.2) by \( \sqrt{2/n} \Sigma^{1/2} \) results in the equation

\[
M(\Sigma, c, d) := c \sum_{i=1}^{n} \frac{\Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2}}{x_i^\top \Sigma^{-1/2} x_i} + d \sum_{i=1}^{n} \Sigma^{-1/2} x_i x_i^\top \Sigma^{-1/2} = I, \tag{3.3}
\]

where we define the constants

\[
c := -\frac{2(a - q/2)}{n}, \quad d := \frac{2}{bn}. \tag{3.4}
\]

We are now state our uniqueness theorem in the following that upon existence the solution to (3.3) is unique.

\textbf{Theorem 1.} If the data set \( \{x_i\}_{i=1}^{n} \) spans \( \mathbb{R}^q \) and \( \Sigma_1, \Sigma_2 \) are positive definite matrices for which \( M(\Sigma_1, c, d) = M(\Sigma_2, c, d) \) and \( c > 0 \), then this implies \( \Sigma_1 = \Sigma_2 \).

\textbf{Proof.} See Appendix B; another proof follows from (Kent and Tyler, 1991, Thm. 2.2)). \hfill \square

Two different fixed point algorithm can be obtained by rewriting (3.3). To make fixed point algorithms easier to be proved, we rewrite (3.3) by first introducing the matrix \( B \) and transformed vectors \( y_i \)

\[
B = d \sum_{i=1}^{n} x_i x_i^\top, \quad y_i = B^{-1/2} x_i. \tag{3.5}
\]

Then, defining \( \Gamma = B^{-1/2} \Sigma B^{-1/2} \) and noting that the square root of \( \Gamma \) has the form \( B^{-1/2} \Sigma^{1/2} Q^\top \) for some orthogonal matrix \( Q \), we turn (3.3) into the equation

\[
c \sum_{i=1}^{n} \frac{\Gamma^{-1/2} y_i y_i^\top \Gamma^{-1/2}}{y_i^\top \Gamma^{-1} y_i} + \Gamma^{-1} = I. \tag{3.6}
\]
From a solution $\Gamma^*$ to (3.6), we recover $\Sigma^* = B^{1/2} \Gamma^* B^{1/2}$, the solution to (3.3).

Our solution to (3.6) splits naturally into two cases: (i) the concave case ($a \geq q/2$, equivalently $c < 0$); and (ii) the nonconcave case ($a < q/2$, equivalently $c > 0$).

### 3.1 The concave case: $c < 0$

Rearrange (3.6) and consider the following “positivity-preserving” iteration

$$
\Gamma_{p+1} = \left( -c \sum_{i=1}^{n} \frac{\Gamma_p^{-1/2} y_i y_i^\top \Gamma_p^{-1/2}}{y_i y_i^\top} + I \right)^{-1}, \quad p = 0, 1, \ldots,
$$

(3.7)

which by construction ensures that if $\Gamma_p \succ 0$, then $\Gamma_{p+1} \succ 0$. Clearly, any limit point of (3.7) is also positive definite and satisfies (3.6), which is sufficient for global optimality, as the log-likelihood is concave. The limit point of the iteration (3.7) is the fixed point of the following map:

$$
\mathcal{G} \ni S \mapsto I + c'S^{1/2} Y DS Y^\top S^{1/2},
$$

(3.8)

where matrix $Y$ has $y_i$ as its $i$th column, $S = \Gamma^{-1}$, $DS = \text{Diag}(1/y_i^\top S y_i)$, and $c' = -c$.

We prove convergence of (3.7) by showing the map $\mathcal{G}$ to be a fixed-point map. This claim is proved by Theorem 4, which is the main result of this section. Crucial to the statement of Theorem 4 is the existence of a compact set $D$ on which the map $\mathcal{G}$ is a fixed-point. The existence of such a compact set is shown by Lemma 2.

**Lemma 2.** Let $\mathcal{G}$ be given by (3.8); let $D := [I, \mu I]$, where $\mu > (1 + c'n)$, then $\mathcal{G}(D) \subset D$.

**Proof.** Let $D = [I, \mu I]$, where $\mu > 1$ is some scalar (to be determined). We show that $\mathcal{G}(D) \subset D$, that is, if $S \in D$, then $\mathcal{G}(S) \in \text{int}(D)$.

To see that, first check that if $S = I$, then $\mathcal{G}(S) = I + c'Y D_I Y^\top \prec \mu I$ for suitably large $\mu$. Moreover, $\mathcal{G}(I) \succ I$ (since $c' > 0$ and not all $Y = 0$; anyways, if $Y = 0$, then $S = I$ is the only fixed point). If $S = \mu I$, then $\mathcal{G}(S) = I + c'Y D_I Y^\top$ (the $\mu$ cancels out because $\mathcal{G}(\alpha S) = \mathcal{G}(S)$ for all $\alpha > 0$). Thus, again in this case $I \prec \mathcal{G}(S) \prec \mu I$. It remains to show that for any $I \prec S \prec \mu I$, it holds that $I \prec \mathcal{G}(S) \prec \mu I$.

Notice that $S^{1/2} y_i y_i^\top S^{1/2} = \frac{z_i z_i^\top}{z_i^\top z_i} \prec I$. Thus, in particular $S^{1/2} Y DS Y^\top S^{1/2} \preceq n I$. So that we have

$$
I \prec \mathcal{G}(S) \preceq I + c'n I = (1 + c'n) I.
$$

(3.9)

So if $\mu > (1 + c'n)$, we are guaranteed that $\mathcal{G}(S) \prec \mu I$. \hfill \Box

Lemma 2 proves crucial because it shows that $\mathcal{G}$ maps a compact convex set $D$ to itself, and since $\mathcal{G}$ is continuous on $D$, it allows us to invoke Brouwer’s fixed-point theorem (Granas and Dugundji, 2003) to conclude existence of a fixed-point in $D$. Theorem 4 establishes the much harder result that this fixed point is unique and can be computed simply by iterating $\mathcal{G}$. We will need the following classic theorem on fixed-point maps for proving Theorem 4.

**Theorem 3.** Let $\mathcal{G}$ be a continuous map that maps a nonempty set $D$ to itself. If the iterated map $\mathcal{G}^k$ has a unique fixed point for every integer $k \geq 1$, then beginning with $S_0 \in D$ the Picard iteration $S_{k+1} = \mathcal{G}(S_k)$ converges to this unique fixed point.

**Proof.** Classic result in fixed-point theory; see e.g., (Goebel and Kirk, 1990). \hfill \Box

Using Theorem 3, we are able to prove Theorem 4 given below.

**Theorem 4.** Let $S_0 \in D$ (for $D$ defined by Lemma 2) be chosen arbitrarily. Then, the iteration $S_{k+1} = \mathcal{G}(S_k)$ converges to a unique fixed-point $S^*$.
Using (3.10), we get:

\[ Q = \begin{pmatrix} \sqrt{\alpha} & \ldots & \sqrt{\alpha} \\ \vdots & \ddots & \vdots \\ \sqrt{\alpha} & \ldots & \sqrt{\alpha} \end{pmatrix} \]

with an orthogonal matrix \( Q \).

Let \( N \) converge, and that in addition \( \alpha > 0 \), we obtain the iteration

\[ \Gamma_{p+1} = \alpha_p \Gamma_p^{1/2} N_p \Gamma_p^{1/2}, \quad p = 0, 1, \ldots, \quad (3.10) \]

where

\[ N_p = \alpha_p \sum_{i=1}^{n} \frac{\Gamma_p^{-1/2} y_i y_i^\top \Gamma_p^{-1/2}}{\gamma_i} + \alpha_p \Gamma_p^{-1} \quad (3.11) \]

with \( \alpha_p > 0 \) is a free scalar parameter. We show that under a specific choice of \( \alpha_p \), iterations (3.11) converge, and that in addition \( \alpha_p \rightarrow \alpha^* = 1 \). Thus, \( \lim_{p \rightarrow \infty} \Gamma_p = \Gamma^* \) satisfies (3.11), and \( \Gamma^* \) is therefore the desired ML solution.

Our proof relies on Lemma 5 which shows that one can find \( \alpha_p \) values that lead to the increase of the smallest eigenvalue of \( N_p \) and decrease of the largest eigenvalue of \( N_p \).

**Lemma 5.** Let \( \lambda_{1,p} > \alpha_p^{-1} \) and \( \lambda_{q,p} < \alpha_p^{-1} \) represent the largest and smallest eigenvalues of \( N_p \), respectively. If the data set \( \{x_i^\prime\}_{i=1}^n \) spans \( \mathbb{R}^q \) then \( \lambda_{1,p+1} \leq \lambda_{1,p} \) and \( \lambda_{q,p+1} \leq \lambda_{q,p+1} \).

**Proof.** By definition,

\[ I = N_p^{-1/2} N_p N_p^{-1/2} = \frac{c}{\alpha_p} \sum_{i=1}^{n} \frac{N_p^{-1/2} \Gamma_p^{-1/2} y_i y_i^\top \Gamma_p^{-1/2} N_p^{-1/2}}{\gamma_i} + N_p^{-1/2} \Gamma_p^{-1/2} N_p^{-1/2} \]

We multiply both numerator and denominator of the first term by the factor \( y_i^\top \Gamma_p^{-1} y_i \) and in the numerator we replace \( y_i^\top \Gamma_p^{-1} y_i \) by \( \frac{1}{\alpha_p} y_i^\top \Gamma_p^{-1} N_p^{-1} \Gamma_p^{-1} y_i \). In addition, we multiply from right and left with an orthogonal matrix \( Q_p \). This yields

\[ I = \frac{c}{\alpha_p} \sum_{i=1}^{n} \frac{Q_p N_p^{-1/2} \Gamma_p^{-1/2} y_i y_i^\top \Gamma_p^{-1/2} N_p^{-1/2} Q_p^\top y_i^\top \Gamma_p^{-1} N_p^{-1} \Gamma_p^{-1} y_i}{y_i^\top \Gamma_p^{-1} y_i} + Q_p N_p^{-1/2} \Gamma_p^{-1/2} N_p^{-1/2} \]

Since the square root of the symmetric positive definite matrix \( \Gamma_p^{1/2} N_p \Gamma_p^{1/2} \) can be written as \( \Gamma_p^{1/2} N_p^{1/2} Q_p^\top \) and using (3.10), we get:

\[ Q_p N_p^{1/2} \Gamma_p^{1/2} = \sqrt{\alpha_p} \Gamma_p^{-1/2}. \quad (3.14) \]
If we substitute (3.14) into (3.13), we get:

\[ I = c \sum_{i=1}^{n} \frac{\Gamma_{p+1}^{-\frac{1}{2}} y_i^\top y_i \Gamma_{p+1}^{-\frac{1}{2}} y_i^\top y_i - \frac{1}{2} N_{p} \Gamma_{p+1}^{-\frac{1}{2}} y_i}{y_i^\top \Gamma_{p+1}^{-1} y_i} + \alpha_p \Gamma_{p+1}^{-1}. \] (3.15)

By the extremal properties of the largest and smallest eigenvalues, we know:

\[
\lambda_{1,p}^{-1} \leq \frac{y_i^\top \Gamma_{p+1}^{-\frac{1}{2}} N_{p} \Gamma_{p+1}^{-\frac{1}{2}} y_i}{y_i^\top \Gamma_{p+1}^{-1} y_i} \leq \lambda_{q,p}^{-1}.
\] (3.16)

Applying inequality (3.16) to (3.15), we obtain following two inequalities:

\[
\lambda_{q,p}^{-1} \left[ \sum_{i=1}^{n} \frac{\Gamma_{p+1}^{-\frac{1}{2}} y_i^\top y_i \Gamma_{p+1}^{-\frac{1}{2}} y_i}{y_i^\top \Gamma_{p+1}^{-1} y_i} \right] + \alpha_p \Gamma_{p+1}^{-1} \geq I,
\] (3.17)

and

\[
I \geq \lambda_{q,p}^{-1} \left[ \sum_{i=1}^{n} \frac{\Gamma_{p+1}^{-\frac{1}{2}} y_i^\top y_i \Gamma_{p+1}^{-\frac{1}{2}} y_i}{y_i^\top \Gamma_{p+1}^{-1} y_i} \right] + \alpha_p \Gamma_{p+1}^{-1}.
\] (3.18)

From rearranging the equality in (3.11) we obtain:

\[
c \sum_{i=1}^{n} \frac{\Gamma_{p+1}^{-\frac{1}{2}} y_i^\top y_i \Gamma_{p+1}^{-\frac{1}{2}} y_i}{y_i^\top \Gamma_{p+1}^{-1} y_i} = N_{p+1} - \Gamma_{p+1}^{-1}.
\] (3.19)

Applying (3.19) to (3.17), we obtain the following inequality:

\[
\lambda_{q,p}^{-1} [N_{p+1} - \Gamma_{p+1}^{-1}] + \alpha_p \Gamma_{p+1}^{-1} \geq I,
\] (3.20)

If we rearrange (3.20), we obtain:

\[
N_{p+1} \geq \lambda_{q,p} I + \Gamma_{p+1}^{-1} (1 - \alpha_p \lambda_{q,p}).
\] (3.21)

Writing the singular value decomposition of \( N_{p+1} \) as \( U_{p+1} \Lambda_{p+1} U_{p+1}^\top \), and multiplying (3.21) from the left by \( U_{p+1}^\top \) and from the right by \( U_{p+1} \), we obtain:

\[
\Lambda_{p+1} \geq \lambda_{q,p} I + U_{p+1}^\top \Gamma_{p+1}^{-1} U_{p+1} (1 - \alpha_p \lambda_{q,p}).
\] (3.22)

Let \( \lambda_{q,p} \leq \alpha_p^{-1} \), then if the data points span \( \mathbb{R}^q \), the matrix

\[
\Xi = U_{p+1}^\top \Gamma_{p+1}^{-1} U_{p+1} (1 - \alpha_p \lambda_{q,p})
\] (3.23)

is a positive semi-definite matrix. Therefore all diagonal elements of \( \Xi \) in (3.23) are larger or equal to zero, and therefore all diagonal elements of \( \Lambda_{p+1} \) are larger or equal to \( \lambda_{q,p} \). Therefore if \( \lambda_{q,p} \leq \alpha_p^{-1} \), then \( \lambda_{q,p} \geq \lambda_{q,p} \) holds true.

Applying the same procedure to the other inequality, (3.18), we obtain:

\[
\Lambda_{p+1} \leq \lambda_{1,p} I - U_{p+1}^\top \Gamma_{p+1}^{-1} U_{p+1} (\alpha_p \lambda_{1,p} - 1).
\] (3.24)

Let \( \lambda_{1,p} \geq \alpha_p^{-1} \), then if the data points span \( \mathbb{R}^q \), the matrix

\[
\Xi' = U_{p+1}^\top \Gamma_{p+1}^{-1} U_{p+1} (\alpha_p \lambda_{1,p} - 1),
\] (3.25)

is a positive semi-definite matrix. Therefore, all diagonal elements of \( \Xi' \) in Equation 3.25 are larger or equal to zero, and therefore all diagonal elements of \( \Lambda_{p+1} \) are smaller or equal to \( \lambda_{1,p} I \). Therefore, if \( \lambda_{1,p} \geq \alpha_p^{-1} \) then \( \lambda_{1,p+1} \leq \lambda_{1,p} \) holds true.
The key result of this section is Theorem 6, which shows that there is a sequence \( \{\alpha_p\} \to 1 \), for which (3.11) converges.

**Theorem 6.** Let \( \lambda_{1,p} \geq 1 \) and \( \lambda_{q,p} \leq 1 \) represent the largest and smallest eigenvalues of \( N_p \), respectively. If the data set \( \{x_i\}_{i=1}^n \) spans \( \mathbb{R}^q \) then one can find an \( \alpha_p \) such that \( 1 \leq \lambda_{1,p+1} \leq \lambda_{1,p} \) and \( \lambda_{q,p} \leq \lambda_{q,p+1} \leq 1 \). This implies that iteration (3.11) converges. Moreover, \( \alpha_p \to 1 \), and one possible choice for \( \alpha_p \) is given by:

- \( \alpha_p = 1 \) if the largest eigenvalue \( \lambda_q^1 \) and the smallest eigenvalue \( \lambda_q^p \) of

\[
N' = \Gamma_p^{-1/2} \Gamma_p'^{-1} \Gamma_p^{-1/2} = c \sum_{i=1}^n \frac{\Gamma_p^{-1/2} y_i y_i^\top \Gamma_p'^{-1/2}}{y_i^\top \Gamma_p'^{-1} y_i} + \Gamma_p^{-1}
\]  

(3.26)

are larger and smaller than one, respectively.

- \( \alpha_p = \lambda_q^{-1} \) if \( \lambda_q^1 \leq 1 \) and \( \lambda_q^p \leq 1 \).

- \( \alpha_p = \lambda_q^{-1} \) if \( \lambda_q^1 \geq 1 \) and \( \lambda_q^p \geq 1 \).

**Proof.** According to the previous lemma \( \lambda_q^1 \leq \lambda_{1,p} \) and \( \lambda_q^p \geq \lambda_{q,p} \). If we rewrite \( N_{p+1} \) in terms of \( \Gamma' \), we obtain:

\[
N_{p+1} = c \sum_{i=1}^n \frac{y_i y_i^\top}{y_i^\top \Gamma_p'^{-1} y_i} + \alpha_p^{-1} \Gamma_p'.
\]  

(3.27)

We now consider three different cases for the eigenvalues of \( N' \) and we express possible values of \( \alpha_p \).

First case \( \lambda_q^1 \geq 1 \) and \( \lambda_q^p \leq 1 \): In this case \( \alpha_p = 1 \) would be the solution.

Second case \( \lambda_q^1 \leq 1 \) and \( \lambda_q^p \leq 1 \): If \( \alpha_p \) decreased toward zero, then \( \lambda_{1,p+1} \) increases toward infinity. In addition \( \lambda_{1,p+1} \) is a continuous function of \( \alpha_p \), therefore if we increase \( \alpha_p \) it will be possible to find an \( \alpha_p \) such that \( \lambda_{1,p+1} = 1 \). At the same time, because of the previous lemma since \( \lambda_{q,p} \leq 1 \leq \alpha_p^{-1} \), therefore \( \lambda_{q,p+1} \geq \lambda_{q,p} \). Note that finding \( \alpha_p \) is equivalent to an eigenvalue problem, we want to find an \( \alpha_p \), such that the largest eigenvalue of \( N_{p+1} \) becomes one. One can show that \( \alpha_p^{-1} \) should be the smallest eigenvalue of the following matrix (see case I in the end of the proof)

\[
\Gamma' - c \sum_{i=1}^n \frac{y_i y_i^\top}{y_i^\top \Gamma_p'^{-1} y_i}.
\]  

(3.28)

Third case \( \lambda_q^1 \geq 1 \) and \( \lambda_q^p \geq 1 \): If \( \alpha_p \) increases toward infinity, then \( \lambda_{q,p+1} \to \lambda < 1 \). Similar to previous case \( \lambda_{q,p+1} \) is a continuous function of \( \alpha_p \), therefore if we decrease \( \alpha_p \), it will be possible to find an \( \alpha_p \) such that \( \lambda_{q,p+1} = 1 \). At the same time since \( \lambda_{1,p} \geq 1 \geq \alpha_p^{-1} \), we have \( \lambda_{1,p+1} \leq \lambda_{1,p} \). One can show that in this case \( \alpha_p^{-1} \) should be the largest eigenvalue of the matrix given in (3.28) (see case II in the end of the proof).

Since the sequences \( \lambda_{1,p} \) and \( \lambda_{q,p} \) are both bounded and they are decreasing and increasing respectively, they are convergent sequences. From the explained procedure for finding \( \alpha_p \), it is easy to see that the convergent value of \( \lambda_{1,p} \) and \( \lambda_{q,p} \) are in the first case and therefore \( \alpha_p \) converges to one.

**Case I:**

We want to calculate \( \alpha_p^{-1} \) such that the largest eigenvalue of the following matrix becomes one given that eigenvalues of the matrix for \( \alpha_p = 1 \) are smaller than one

\[
N_{p+1} = c \sum_{i=1}^n \frac{\Gamma_p'^{-1/2} y_i y_i^\top \Gamma_p'^{-1/2}}{y_i^\top \Gamma_p'^{-1} y_i} + \alpha_p^{-1} \Gamma_p'.
\]  

(3.29)
Since both matrices $\Gamma^{-1}$ and $c \sum_{i=1}^{n} \frac{y_i y_i^\top \Gamma^{-1/2} y_i}{y_i^\top \Gamma^{-1} y_i}$ are positive definite, if we increase $\alpha_p^{-1}$ the largest eigenvalue increases. Therefore we need to find the smallest $\alpha_p^{-1}$ such that an eigenvalue of the matrix $N_{p+1}$ becomes one. We have the eigenvalue problem:

$$ N_{p+1} u = u. \quad (3.30) $$

Assume $u = \Gamma^{1/2} v$ and multiply (3.30) from left by the matrix $\Gamma^{1/2}$, we obtain:

$$ c \sum_{i=1}^{n} \frac{y_i y_i^\top}{y_i^\top \Gamma^{-1} y_i} v + \alpha_p^{-1} v = \Gamma' v. \quad (3.31) $$

If we rearrange the equation, we obtain:

$$ \Gamma' v - c \sum_{i=1}^{n} \frac{y_i y_i^\top}{y_i^\top \Gamma^{-1} y_i} v = \alpha_p^{-1} v. \quad (3.32) $$

Hence, $\alpha_p^{-1}$ is the smallest eigenvalue of the following matrix:

$$ \Gamma' - c \sum_{i=1}^{n} \frac{y_i y_i^\top}{y_i^\top \Gamma^{-1} y_i}. \quad (3.33) $$

**Case II:**

We want to calculate $\alpha_p^{-1}$ such that the smallest eigenvalue of the matrix in (3.31) becomes one given that eigenvalues of the matrix for $\alpha_p = 1$ are larger than one. Since both matrices $\Gamma'^{-1}$ and $c \sum_{i=1}^{n} \frac{y_i y_i^\top \Gamma'^{-1/2} y_i}{y_i^\top \Gamma'^{-1} y_i}$ are positive definite, if we decrease $\alpha_p^{-1}$, the smallest eigenvalue decreases. Therefore we need to find the largest $\alpha_p^{-1}$ such that an eigenvalue of the matrix $N_{p+1}$ becomes one. Here, we have the same eigenvalue problem, and therefore we obtain that $\alpha_p^{-1}$ is the largest eigenvalue of the matrix in (3.33).

One can invoke a result of (Kent and Tyler, 1991) or the theory of (Sra and Hosseini, 2013b) to obtain convergence proofs for an alternative iteration that computes $\Gamma$. However, the convergence results of (Kent and Tyler, 1991; Sra and Hosseini, 2013b) depend on the existence of an ML solution. Theorem 6 proves a stronger result because it does not depend on any existence requirement on the ML solution. This generality has some important consequences: if the ML solution exists, then inevitably iteration (3.11) converges to it. But when the ML solution does not exist (which is well possible), then the iterative algorithm still converges, though now the convergent solution is singular. This singular matrix possesses specific structure that can be then used for robust subspace recovery, generalizing subspace recovery approach of (Zhang, 2012). This topic is, however, beyond the present paper and will be considered in a future paper.

In addition to the stronger result of Theorem 6, it also gives a computationally efficient method for ML solution. It highly outperforms both the method mentioned in (Kent and Tyler, 1991; Sra and Hosseini, 2013b) and also manifold optimization techniques (see experimental results).

It worth mentioning that the above theorem suggests $\alpha_p$ values which are not necessarily optimal, though easy to calculate. In practice, we observed that if one chooses the parameter $\alpha_p$ such that the trace of the matrix $N_{p+1}$ becomes $q$, that is $\alpha_p = \text{tr}(\Gamma'^{-1})/(2a)$, then the convergence is faster. However, for this case our convergence proof does not apply.
4 KL Divergence between EGDs

In this section we first derive the KL-divergence between two arbitrary ECDs and then for two EGDs. Consider, we have two probability distribution \( P \) and \( Q \) with the densities \( p \) and \( q \), the KL-divergence between these two distributions is defined by:

\[
\text{KL}(P||Q) = \int \log \frac{p(x)}{q(x)} p(x) dx = \int \log p(x) p(x) dx - \int \log q(x) p(x) dx.
\] (4.1)

If \( P \) is the true distribution of the random variable \( X \), then the negative of the first part, \( H(X) = -\int \log p(x) p(x) dx \), is called entropy and the second part \( \mathbb{E}[-\log q(X)] = -\int \log q(x) p(x) dx \) is called the averaged log-loss of distribution \( Q \). The KL-divergence is invariant against invertible transformations of \( X \). Therefore, without loss of generality, we assume that the covariance matrix of the second distribution \( Q = \mathcal{ECD}(\Sigma_2, g) \) is the identity. This assumption can always be fulfilled by the change of variables \( y = \Sigma_2^{-1/2} x \).

For the calculations, we consider \( P \) and \( Q \) as two ECDs with the corresponding densities \( P = \mathcal{ECD}(\Sigma, f) \) and \( Q = \mathcal{ECD}(I, g) \), respectively. Firstly, we derive the entropy of \( X \) with distribution \( P \), namely \( H(X) \).

\[
H(X) = -\int \log p(x) p(x) dx.
\] (4.2)

Assume \( y = \Sigma^{-1/2} x \), then we obtain following expression for \( H(X) \)

\[
H(X) = \frac{1}{2} \log(|\Sigma|) - \int \log p(y) p(y) dy
\] (4.3)

\[
= -\frac{1}{2} \log(|\Sigma|) + \int \log (f(y^\top y)) f(y^\top y) dy.
\]

Consider \( r = ||y||_2 \) and using the knowledge that the surface of a sphere in dimension \( q \) with radius \( r \) is equal to \( 2r^{q-1}\pi^\frac{q}{2}/\Gamma(q/2) \), we get out:

\[
H(X) = \frac{1}{2} \log(|\Sigma|) - \int \log (f(r^2)) \frac{2\pi^\frac{q}{2}}{\Gamma(q/2)} r^{q-1} f(r^2) dr.
\] (4.4)

Then using

\[
p_R(r) = \frac{2\pi^\frac{q}{2}}{\Gamma(\frac{q}{2})} r^{q-1},
\] (4.5)

we obtain:

\[
H(X) = \frac{1}{2} \log(|\Sigma|) - \int \log \left( \frac{\pi^\frac{q}{2}}{\Gamma(q/2)} r^{1-q} p_R(r) \right) p_R(r) dr.
\] (4.6)

Consider the random variable change \( \Upsilon = R^2 \), then we obtain:

\[
H(X) = \frac{1}{2} \log(|\Sigma|) - \int \log \left( \frac{\Gamma(q/2)}{\pi^\frac{q}{2}} \nu^{1-q/2} p_T(\nu) \right) p_T(\nu) d\nu.
\] (4.7)

For the case of EGD, with parameter \( a_p \) and \( b_p \), the neg-entropy is:

\[
H(X) = -\frac{1}{2} \log(|\Sigma|) - \int \log \left( \frac{\pi^\frac{q}{2}}{\Gamma(a_p) b_p^a} \nu^{a_p-1} \exp (-\nu/b_p) \right) \frac{1}{\Gamma(a_p) b_p^a} \nu^{a_p-1} \exp (-\nu/b_p) d\nu.
\] (4.8)
We know that the following two equalities hold true:

\[
\frac{1}{\Gamma(a + 1)b^{a+1}} \int_{0}^{\infty} r^a \exp(-r/b) dr = 1, \quad (4.9)
\]

\[
\frac{1}{\Gamma(a + 1)b^{a+1}} \int_{0}^{\infty} \log(r) r^a \exp(-r/b) dr = \Psi(a + 1) + \log(b), \quad (4.10)
\]

where \(\Psi\) is digamma function. Using relations (4.9) and (4.10), if we simplify (4.8), we obtain:

\[
H(X) = +\frac{1}{2} \log(|\Sigma|) - \log \left( \frac{\Gamma(q/2)}{\pi^{q/2} \Gamma(a_p) b_p^{a_p}} \right) - (a_p - q/2) (\Psi(a_p) + \log(b_p)) + a_p. \quad (4.11)
\]

**General Case: P and Q are arbitrary ECDs**

The expression for the averaged log-loss is as below:

\[
E[-\log q(X)] = - \int \log q(x)p(x)dx = - \int \log (q(x^T x)) f(x^T \Sigma^{-1} x)dx. \quad (4.12)
\]

Since \(q\) is only dependent on the radial component, we just need to calculate the projection of \(p\) on the radial component and then to calculate the following expression:

\[
E[-\log q(X)] = - \int \log \left( \frac{\Gamma(q/2)}{\pi^{q/2}} v^{1-q/2} q_T(v) \right) p_Z(v)dv, \quad (4.13)
\]

where \(q_T\) is the square radial density of \(Q\), and \(p_Z\) is the p.d.f of \(Z = \|X\|^2\) where \(X \sim \mathcal{ECD}(\Sigma, f)\). It is easy to see that the random variable \(Z\) is equal to \(Z = Y^T \Sigma Y\) where \(Y \sim \mathcal{ECD}(I, f)\). The form of the p.d.f has already been reported by Provost and Cheong (1998).

\[
p_Z(z) = \int_{0}^{\infty} \frac{1}{r} \mathcal{K}(z/r)p_T(r)dr, \quad (4.14)
\]

where \(p_T\) is the square of the radial distribution of \(Y\), and \(\mathcal{K}\) is a density that is related to the component sum of a Dirichlet random variable (see appendix C). We replace \(p_Z\) in (4.13) by the expression in (4.14), and we obtain the following expression for the averaged log-loss:

\[
E[-\log q(X)] = - \int_{0}^{\infty} \frac{1}{r} \mathcal{K}(v/r)p_T(r) \log \left( \frac{\Gamma(q/2)}{\pi^{q/2}} v^{1-q/2} q_T(v) \right) dvdr. \quad (4.15)
\]

Thus the KL-divergence between two arbitrary ECDs can be evaluated by the following expression:

\[
KL(P||Q) = -\frac{1}{2} \log(|\Sigma|) + \int_{0}^{\infty} \log \left( v^{1-q/2} p_T(v) \right) p_T(v)dv - \int_{0}^{\infty} \frac{1}{r} \mathcal{K}(v/r)p_T(r) \log \left( v^{1-q/2} q_T(v) \right) dvdr. \quad (4.16)
\]
Special Case: P and Q are EGDs

For the case that \( P \) and \( Q \) are EGDs with parameters \( a_p, b_p \) and \( a_q \) and \( b_q \) respectively, we obtain the following expression for the averaged log-loss:

\[
\mathbb{E}[-\log q(X)] = -\int_0^\infty \int_0^1 \frac{1}{r} \mathcal{K}(v/r) \frac{1}{\Gamma(a_p)b_p^a} r^{a_p-1} \exp(-r/b_p) \log \left( \frac{\Gamma(q/2)}{\pi^{q/2} \Gamma(a_q)b_q^a} r^{a_q-q/2} \exp(-v/b_q) \right) dvdr.
\]

(4.17)

We apply the change of variables \( \mu = v/r \) and \( r = r \), therefore:

\[
\mathbb{E}[-\log q(X)] = \log(\pi^{q/2} \Gamma(a_q) b_q^a) - \log(\Gamma(q/2))
\]

(4.18)

\[
- \frac{a_q - q/2}{\Gamma(a_p)b_p^a} \int_0^\infty \mathcal{K}(\mu) \log(\mu) d\mu \int_0^\infty r^{a_p-1} \exp(-r/b_p) dr
\]

\[
- \frac{a_q - q/2}{\Gamma(a_p)b_p^a} \int_0^\infty \mathcal{K}(\mu) d\mu \int_0^\infty \log(r) r^{a_p-1} \exp(-r/b_p) dr
\]

\[
+ \frac{1}{b_q \Gamma(a_q)b_q^a} \int_0^\infty \mathcal{K}(\mu) \mu d\mu \int_0^\infty r^{a_q} \exp(-r/b_q) dr.
\]

The numerical methods for computing \( \mathcal{A} = \int_0^\infty \mathcal{K}(\mu) \log(\mu) d\mu \) and closed-form formula for \( \int_0^\infty \mathcal{K}(\mu) d\mu \) are given in appendix C. Using the equalities (4.9) and (4.10), we obtain:

\[
\mathbb{E}[-\log q(X)] = \log(\pi^{q/2} \Gamma(a_q) b_q^a) - \log(\Gamma(q/2)) - (a_q - q/2) (\Psi(a_p) + \log(b_p))
\]

(4.19)

\[
+ \frac{a_p b_p}{q b_q} \operatorname{tr}(\Sigma) - (a_q - q/2) \mathcal{A}.
\]

Summing up the equations for neg-entropy and averaged log-loss, we obtain:

\[
KL(P||Q) = -\frac{1}{2} \log(|\Sigma|) + \log \left( \frac{\Gamma(a_q)b_q^a}{\Gamma(a_p)b_p^a} \right) + (a_p - a_q) \Psi(a_p) - a_p + \frac{a_p b_p}{q b_q} \operatorname{tr}(\Sigma) - (a_q - q/2) \mathcal{A}
\]

(4.20)

The left plot in Fig. 2 shows the KL-Divergence between two EGDs as dimension \( q \) increases. The observed distribution is an EGD with arbitrary scale parameter while the model is a Gaussian with the same covariance. This plot reveals that if the scale parameter is very small or if it is very large, the KL-Divergence becomes very large, growing to infinity in the limit. This shows that the goodness-of-fit can be substantially improved for the EGD model relative to the Gaussian model.

If we have two EGDs with the same shape parameter \( a \) such that the covariance matrix, say \( \Sigma_1 \), of one of the distributions is treated as a rotation of the other \( \Sigma_2 \). Then, their KL-Divergence is

\[
KL(P||Q) = a (\operatorname{tr}(\Sigma_1 \Sigma_2^{-1})/q \right) - (a - q/2) \mathcal{A}
\]

The upper plot in Fig. 2 shows the contour plot of the term \( \operatorname{tr}(\Sigma_1 \Sigma_2^{-1})/q \) and the lower plot shows the contour plot of the term \( \mathcal{A} \). The x-axis is the rotation degree and the y-axis is the ratio of the largest eigenvalue to the smallest one. For small \( a \) the two terms get similar signs and since the behavior of two terms look similar, KL-divergence changes more by changing rotation degree and condition number of the covariance matrix. For large \( a \), those terms will get opposite signs and cancel each other.

5 Mixture modeling with EGDs

After developing algorithms and theory for ML estimation with EGDs, we are ready to discuss mixture modeling. In Section 5.2, we present a “split-and-merge” expectation maximization (EM) algorithm for
estimating parameters of an EGD mixture model. This algorithm uses the KL-Divergence computation presented in Section 4 to help decide whether two mixture components should be merged into one.

5.1 EM Algorithm for mixture of EGDs

A K-component mixture of Elliptical Gamma distributions (MEG) has the distribution

$$p(x) = \sum_{k=1}^{K} p_k p_{eg}(x; \Sigma_k, a_k, b_k),$$

(5.1)

where $\sum_k p_k = 1$ ($p_k \geq 0$). We use a block coordinate descent algorithm for computing an ML solution. First, we fix $a_k$ and $b_k$ and apply one step of EM to obtain $\Sigma_k$ ($1 \leq k \leq K$) using our proposed fixed-point algorithms. Next, we fix $\Sigma_k$ and update $a_k$, $b_k$. Here, the following variable change $v^k = x^T \Sigma_k x$ proves helpful, because with it the density (5.1) turns into

$$p(x) = \sum_{k=1}^{K} p_k p_{ga}(v^k; a_k, b_k),$$

(5.2)

where $p_{ga}$ is the Gamma density (2.2).

Each step of EM algorithm consists of two steps Expectation step (E-step) and Maximization step (M-step). These steps for the first stage is written as follows:

- **E-step:** Based on the current estimate of weight parameters for each data point $x_i$ and each mixture $k$:

$$t_i^k = \frac{p_k p_{eg}(x_i; \Sigma_k, a_k, b_k)}{\sum_{i=1}^{K} p_i p_{eg}(x_i; \Sigma_i, a_i, b_i)} = \frac{p_k p_{ga}(v_i^k; a_k, b_k)}{\sum_{i=1}^{K} p_i p_{ga}(v_i^k; a_i, b_i)}$$
M-step: This step updates the parameters of the mixture model by maximizing weighted log-likelihood sub-problems:

\[ \ell_k(\Sigma_k, a_k, b_k; x_1, \ldots, x_n) = \sum_{i=1}^{n} t_i^k \log p_{eg}(x_i; \Sigma_k, a_k, b_k) \]

parameters \( p_k \) are updated by the following equation:

\[ p_k = n^{-1} \sum_{i=1}^{n} t_i^k \]

ML-estimation methods for an EGD explained in section two can be easily modified to the case of parameter estimation of weighted log-likelihood of an EGD.

Similar to the first stage, one step of the EM algorithm for the second stage is also consists of two steps that we apply sequentially till convergence. The E-step and updating \( p_k \) parameters in M-step is similar to that of the first stage. For updating \( a_k \) and \( b_k \) parameters in M-step, the following objective function is maximized:

\[ \ell_k(a_k, b_k; v_1^k, \ldots, v_n^k) = \sum_{i=1}^{n} t_i^k \log p_{ga}(v_i^k | a_k, b_k) \]

The maximum weighted log-likelihood estimates of the parameters can be calculated efficiently using Generalized Newton method (Minka, 2002). Modifying the method explained in (Minka, 2002) to account for weights, we will obtain the following fix-point algorithm:

\[ \frac{1}{a_{k_{new}}} = \frac{1}{a_k} + \frac{1}{a_k} \left( \frac{a_k \Psi(a_k) - \Psi(\bar{a}_k)}{a_k^2 (\bar{1} - \Psi'(a_k))} \right) \]  \hspace{1cm} (5.3)

Where \( \bar{z} \) is weighted mean over \( z \) \((\sum_i t_i^k z_i^k / \sum_i t_i^k)\) and \( \Psi \) is the digamma function. The other parameter is calculated simply using the following equation:

\[ b_k = \frac{\bar{v}_k}{a_k}. \]  \hspace{1cm} (5.4)

To counter the problem of poor local optima in EM, we derive below a more refined “split-and-merge” EM procedure (Ueda et al., 2000a) for the EG mixture model.

### 5.2 Split-and-Merge EM for mixture of EGDs

Ueda et al. (2000a) identified false division of the number of mixture components in different parts of the data cloud as a major cause of local minima that plague EM. They proposed a remedy for countering local minima that is quite effective in practice. The idea is to iteratively find candidates to merge and candidates to split while maximizing the log-likelihood. This process is continued until further splitting or merging fail to improve the model fit. (Blekas and Lagaris, 2007) added more criteria for splitting and merging to the ones given in (Ueda et al., 2000a,b) and also modified the original split-and-merge algorithm. One criterion explained in (Blekas and Lagaris, 2007) is to find two components with minimum symmetric KL-Divergence difference. We observed that this criterion can correctly specify the components, merging which leads to the highest improvement in likelihood. However, all three different criteria for splitting given in (Blekas and Lagaris, 2007) have problems in identifying the correct component to split. In practice, these methods often select the component with the highest entropy, though clearly this component is not necessarily the best candidate for splitting.
Therefore, we propose a variant of the split-and-merge algorithm that not only solves the problem of finding the component that needs to be split, but does so computationally efficiently—pseudo-code is provided as Algorithm 1. Therein, \( \theta_k \) represents parameters of component \( k \), i.e., \( \{ \theta_k \} = \{ p_k, \ a_k, \ b_k, \ \Sigma_k \} \).

The first stage of the algorithm splits components until no further improvement is possible, or when the number of components reaches an upper limit \( K_{\text{max}} \). The algorithm chooses components, which in their previous split led to the highest improvement in log-likelihood (\( \Delta \ell_i \)). If after a split the improvement is less than a threshold denoted by \( t \), the split is not accepted and the component is not further split.

The second stage of the algorithm finds components with the minimum KL-Divergence and merges them. This process is continued until the number of components reaches \( K \). At the last stage, the algorithm performs one step of optimization over all components. The proposed algorithm is efficient: optimization is performed only over the parameters of the split or merged components; computational complexity of the splitting step is equal to optimizing the mixture model with two components; complexity of merging step is equal to ML estimation for just one component.

Here we discuss some points regarding split-and-merge algorithm described above.

- For each step of splitting stage, initialization over split mixture is done simply by adding a small amount of noise to the parameters before splitting.

- For initialization in the merging step, we can simply use one of the mixtures that are merged.

- Parameter \( K_{\text{max}} \) is chosen based on the maximum computational time. The typical value of \( K_{\text{max}} \approx 2K \) works well in practice.

- If there is no validation set that is checked during optimization to avoid overfitting. Then, the amount of threshold \( t \) can be chosen by

\[
t = d \sum_{k=1}^{K+1} \frac{n_{k_{\text{new}}}}{n_{k_{\text{new}}} - d - 1} + d \sum_{k=1}^{K} \frac{n_{k_{\text{old}}}}{n_{k_{\text{old}}} - d - 1},
\]
where $d = q(q + 1)/2 + 1$ is the number of parameters for each component and $n_{k_{new}}$, $n_{k_{old}}$ are the number of data in component $k$ after and before split, respectively. The quantity $dn/(n - d - 1)$ is corrected Akaike Information Criterion (AICc) (Hurvich and Tsai, 1991) that approximately measures expected cross-validation bias between training and test sets.

- If early-stopping is used to avoid overfitting, then threshold $t$ can be chosen to be smaller number (a fraction of $d$ like $d/10$ works fine in practice).

The results of applying proposed algorithm to a mixture of one-dimensional Gaussian is shown in Fig. 3. Proposed algorithm successfully recovers the distribution. Fig. 4 shows the result of an alternative algorithm explained in (Blekas and Lagaris, 2007) to the same problem. Due to the problem with finding good candidate for splitting, this approach could not recover the distribution.

6 Experiments and Application

6.1 ML-estimation using fixed-point iteration

In our first set of experiments we report results on the convergence speed of our fixed-point iterations, namely (3.7) and (3.11). Fig. 5 shows the convergence speed of our fixed-point algorithms compared to three manifold optimization techniques, namely limited-memory BFGS (lbfgs), trust-region and conjugate gradient and also Kent-Tyler method explained in (Kent and Tyler, 1991). We also tested other optimization techniques like interior-point methods (Nesterov and Nemirovski, 1994) but did not include them in the plot because they were vastly slower than manifold optimization techniques. We sampled
10,000 points from an EGD with a random covariance matrix, and initialized the iterations with a random covariance. The left plot in Fig. 5 shows the result for the case $a = 1$ (non-concave case) and right plot is for the case $a = 50$ (concave case).

6.2 Application: Natural Image Modeling

We use MEG to model statistical distribution of natural image patches. The data used for fitting the model is patches sampled from random locations in a natural image dataset. We extracted image patches of two different sizes $8 \times 8$ and $16 \times 16$ from random locations in the van Hateren dataset (van Hateren and van der Schaaf, 1998). This dataset contains 4167 images; we excluded images that had problems, e.g., were noisy, blurred, etc. We extracted 200,000 training image patches, and 10 sets of 100,000 test image patches from remaining 3632 images. We preprocess image patches by log-transforming pixel intensities.

We evaluate the performance of different models using the Multi-Information Rate (MI-Rate) criterion. MI-Rate (in bits/pixel) has the intuitive flavor that it approximately shows the number of bits per pixel that one saves if the patch-level model distribution is used compared to the case that all pixels are modeled independently. Formally, it is defined as

$$\text{MI-Rate} \approx \frac{H(X_0) + \frac{1}{n-1} \ell(\theta; x_1, \ldots, x_n)}{\log 2},$$

where $H(X_0)$ is the entropy of one pixel. The relation becomes exact if $n \to \infty$ (Hosseini et al., 2010).

Table 1 summarizes the performance of different procedures in terms of MI-Rate. In all models the DC component is modeled independently using mixture of Gaussians with 10 components. Two different sizes are included in order to observe how the MI-Rate estimate of different models are changed if the patch size is increased. Among different methods, MEG shows the best performance, yielding the highest MI-Rate per pixel.

In the table, Gauss denotes the simple Gaussian model; the MI-Rate captured by this model is called the amount of second-order information present in the data. The number of layers in hierarchical ICA (HICA) is 8 for $8 \times 8$ patches and 4 for $16 \times 16$ patches (Hosseini and Bethge, 2009). The number of mixtures for MoG and MEG is 8 (Zoran and Weiss, 2012). Note that both MoG and HICA are universal approximators, therefore theoretically they may reach the performance of MEG but with more parameters. In practice, however, parsimonious models are usually preferred. $L_p$-spherical model is a density model proposed in Sinz et al. (2009). RG+ICA corresponds to radial Gaussianisation followed by one layer ICA (Hosseini and Bethge, 2009). DBN corresponds to Deep Belief Networks and GRBM corresponds to Gaussian Restricted Boltzmann Machine. The MI-Rate of DBN and DRBM were evaluated by the method explained in (Theis et al., 2011).

We emphasize that the differences in MI-Rate shown in Table 1 are significant, because closer to the upper limit of the MI-rate any improvement means capturing a lot of perceptually relevant regularities.
Figure 5: Comparison of the proposed fixed-point algorithms against manifold optimization techniques for EG distributions with dimension equal to 64 (left plot) $a = 1$ (right plot) $a = 50$.

| Model   | $8 \times 8$ | $16 \times 16$ |
|---------|---------------|----------------|
| Gauss   | 2.50          | 2.60           |
| GRBM    | 2.69          | 2.74           |
| DBN     | 2.73          | 2.79           |
| ICA     | 2.73          | 2.83           |
| EG      | 2.83          | 2.90           |
| HICA    | 2.84          | 2.91           |
| $L_p$-spherical | 2.85   | 2.95           |
| RG + ICA| 2.87          | 3.00           |
| MoG     | 2.89          | 2.98           |
| MEG     | **2.93**      | **3.02**       |

Table 1: MI-Rate (bits/pixel; higher is better) for different models and two different patch sizes.

Figure 6: (blue plot) MI-Rate for MEG with different number of components (dotted red plot) MI-Rate for Gaussian.

of the underlying distribution, a claim grounded in the recent psychophysical results in (Gerhard et al., 2013).

Finally, Fig. 6 visualizes the effect of number of mixture components on the performance. As a baseline the Gaussian MI-Rate is plotted as a dotted line in the figure.

7 Discussion and future work

We studied a powerful class of symmetric distributions, namely, Elliptical Gamma distributions. We presented theory outlining existence and uniqueness of maximum likelihood estimators for EGDs and developed simple and computationally effective algorithms computing these. As an application of our theory, we studied mixture models based on EGDs for modeling natural image statistics, for which we showed state-of-the-art performance.

Several avenues of further research remain open. The most important direction is to study robust subspace recovery and its applications (Soltanolkotabi et al., 2012). Other potential directions involve
developing mathematical tools to study stochastic processes based on EGDs, as well as to investigate other applications where non-Gaussian data can benefit from EGDs or their mixture models. We hope that the basic theory and practical application outlined in this paper encourage other researchers to also study non-Gaussian modeling with EGDs or families richer than them.

References

Jo. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. ORSA J. Computing, 7(1):36–43, 1995.

R. B. Arellano-Valle, G. del Pino, and P. Iglesias. Bayesian inference in spherical linear models: robustness and conjugate analysis. J. Multivariate Analysis, 97(1):179–197, 2006.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. JMLR, 6:1345–1382, Sep 2005.

Nicholas H Bingham and Rüdiger Kiesel. Semi-parametric modelling in finance: theoretical foundations. Quantitative Finance, 2(4):241–250, 2002.

K. Blekas and I. E. Lagaris. Split–merge incremental learning (SMILE) of mixture models. In Artificial Neural Networks—ICANN 2007, pages 291–300. Springer Berlin Heidelberg, 2007.

Yasuko Chikuse. Statistics on special manifolds, volume 174. Springer, 2003.

K.-T. Fang and Y.-T. Zhang. Generalized multivariate analysis. Springer, 1990.

K.-T. Fang, S. Kotz, and K. W. Ng. Symmetric multivariate and related distributions, volume 1. Chapman and Hall London, 1990.

Holly E Gerhard, Felix A Wichmann, and Matthias Bethge. How sensitive is the human visual system to the local statistics of natural images? PLoS computational biology, 9(1):e1002873, 2013.

K. Goebel and W. A. Kirk. Topics in Metric Fixed Point Theory. Cambridge University Press, 1990.

A. Granas and J. Dugundji. Fixed Point Theory. Springer, 2003.

R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, Newyork, 1985.

R. Hosseini and M. Bethge. Hierarchical models of natural images. In Frontiers in Computational Neuroscience, 2009.

R. Hosseini, F. Sinz, and M. Bethge. Lower bounds on the redundancy of natural images. Vision research, 50(22):2213–2222, 2010.

Clifford M Hurvich and Chih-Ling Tsai. Bias of the corrected aic criterion for underfitted regression and time series models. Biometrika, 78(3):499–509, 1991.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on, 10(3):626–634, 1999.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. Biometrika, 48(3/4):419–426, December 1961.

J. T. Kent and D. E. Tyler. Redescending M-estimates of multivariate location and scatter. The Annals of Statistics, 19(4):2102–2119, December 1991.

S. Kotz, N. L. Johnson, and D. W. Boyd. Series representations of distributions of quadratic forms in normal variables. i. central case. The Annals of Mathematical Statistics, 38(3):823–837, June 1967.

Samuel Kotz, Tomasz Kozubowski, and Krzysztof Podgorski. The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomies, Engineering, and Finance. Springer, 2001.
M. Koutras. On the generalized noncentral Chi-Squared distribution induced by an elliptical gamma law. *Biometrika*, 73(2):528–532, August 1986.

K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modeling using the t distribution. *J. the American Statistical Association*, 84(408):881–896, December 1989.

Te-Won Lee, Michael S Lewicki, and Terrence J Sejnowski. Unsupervised classification with non-Gaussian mixture models using ICA. In *Advances in neural information processing systems*, pages 508–514, 1999.

Gilad Lerman, Teng Zhang, et al. Robust recovery of multiple subspaces by geometric lp minimization. *The Annals of Statistics*, 39(5):2686–2715, 2011.

T. P. Minka. Estimating a gamma distribution. http://research.microsoft.com/en-us/um/people/minka/papers/minka-gamma.pdf, 2002.

Yu. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

E. Ollila, D.E. Tyler, V. Koivunen, and H. V. Poor. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, 2011.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *J. Mathematical Imaging and Vision*, 25(1):127–154, 2006.

S. B. Provost and Y.-H. Cheong. The probability content of cones in isotropic random fields. *J. Multivariate Analysis*, 66:237–254, August 1998.

H. Rabbani and M. Vafadust. Image/video denoising based on a mixture of Laplace distributions with local parameters in multidimensional complex wavelet domain. *Signal Processing*, 88(1):158–173, 2008.

H. Ruben. Probability content of regions under spherical normal distributions, IV: the distribution of homogeneous and Non-Homogeneous quadratic functions of normal variables. *The Annals of Mathematical Statistics*, 33(2):542–570, June 1962.

Matthias W. Seeger and Hannes Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011.

F. Sinz, E. Simoncelli, and M. Bethge. Hierarchical modeling of local image features through lp-nested symmetric distributions. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2009.

Mahdi Soltanolkotabi, Emmanuel J Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–238, 2012.

S. Sra and R. Hosseini. Conic geometric optimisation on the manifold of positive definite matrices. *arXiv preprint*, 2013a.

S. Sra and R. Hosseini. Geometric optimisation on positive definite matrices for elliptically contoured distributions. In *Advances in Neural Information Processing Systems*, pages 2562–2570, 2013b.

S. Tan and L. Jiao. Multivariate statistical models for image denoising in the wavelet domain. *International J. Computer Vision*, 75(2):209–230, 2007.

J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy. Elliptically contoured distributions for anomalous change detection in hyperspectral imagery. *IEEE Geosci. & Remote Sens. Lett.*, 7(2):271–275, 2010.

L. Theis, S. Gerwinn, F. Sinz, and M. Bethge. In ALL likelihood, deep belief is not enough. *J. Machine Learning Research*, 12:3071–3096, 2011.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Split and merge EM algorithm for improving gaussian mixture density estimates. *J. VLSI Signal Processing*, 26(1):133–140, 2000a.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Smem algorithm for mixture models. *Neural computation*, 12(9):2109–2128, 2000b.
A Showing EG can be expressed as a GSM

From properties of Laplace transform, we know that the inverse Laplace transform of the following function

\[ p^{-v} e^{-cp}, \quad v > 0 \]  

is equal to

\[ w(t) = \begin{cases} 0, & 0 < t < c \\ \frac{(t-c)^{v-1}}{\Gamma(v)}, & t > c \end{cases} \]  

We can therefore use the definition of Laplace transform:

\[ p^{-v} e^{-cp} = \int_0^\infty w(t) e^{-tp} dt \]  

Now assume \( v = q/2 - a \) and \( c = 2b^{-1} \) and \( p = \frac{1}{2} x^\top \Sigma^{-1} x \) in definition of EG in (1.1), then we obtain:

\[
p_{eg}(x) = C |\Sigma|^{-1/2} \int_{2b^{-1}}^\infty w(t) e^{-tx^\top \Sigma^{-1} x} dt = \int_{2b^{-1}}^\infty \frac{\Gamma(q/2)}{\Gamma(a) b^a} (t - 2b^{-1})^{q/2 - a - 1} (2\pi)^{-q/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} t x^\top \Sigma^{-1} x\right) dt \]

\[
= \int_{2b^{-1}}^\infty \frac{\Gamma(q/2)}{\Gamma(a) (q/2 - a)} \frac{(t - 2b^{-1})^{q/2 - a - 1} - 2a}{(t)^{q/2} b^a} p_n(x; \alpha, t^{-1} \Sigma) dt \]  

Without loss of generality assume \( b = 2 \) and use the change of variable \( u = 1/t \), we obtain:

\[
p_{eg}(x) = \int_0^1 \frac{\Gamma(q/2)}{\Gamma(q/2 - a) \Gamma(a)} (1 - u)^{q/2 - a - 1} u^{a-1} p_n(x; \alpha, u \Sigma) du \]

Interestingly, the first term is equal to beta distribution with parameters \( q/2 - \alpha \) and \( \alpha \).

\[
p_{eg}(x) = \int_0^1 p_\beta(u; q/2 - \alpha, \alpha) p_n(x; \alpha, u \Sigma) du \]
B  Uniqueness of the log-likelihood critical point

Proof. By the assumption \( M(\Sigma_1, c, d) = M(\Sigma_2, c, d) \), we have:

\[
\begin{align*}
&c \sum_{i=1}^{n} \frac{\Sigma_1^{-1/2} x_i \Sigma_1^{-1/2}}{x_i \Sigma_1^{-1/2}} + d \sum_{i=1}^{n} \Sigma_1^{-1/2} x_i x_i^\top \Sigma_1^{-1/2} \\
&= c \sum_{i=1}^{n} \frac{\Sigma_2^{-1/2} x_i \Sigma_2^{-1/2}}{x_i \Sigma_2^{-1/2}} + d \sum_{i=1}^{n} \Sigma_2^{-1/2} x_i x_i^\top \Sigma_2^{-1/2}
\end{align*}
\]

(B.1)

Substituting \( z_i = \Sigma_2^{-1/2} x_i \) and \( S = \Sigma_1^{-1/2} \Sigma_2^{1/2} \) in the equation, we obtain:

\[
\begin{align*}
&c \sum_{i=1}^{n} \frac{S_z_i z_i^\top S}{S^\top S z_i} + d \sum_{i=1}^{n} \Sigma z_i z_i^\top \Sigma^\top = c \sum_{i=1}^{n} \frac{z_i z_i^\top}{z_i} + d \sum_{i=1}^{n} z_i z_i^\top,
\end{align*}
\]

(B.2)

Let \( u \) be a right eigenvector for \( S \) corresponding to the eigenvalue \( \Gamma \), then multiplying (B.2) from left by \( u^\top \) and from right by \( u \) and using the fact that the following equality holds for the eigenprojection:

\[
u^\top S z = \Gamma u^\top z,
\]

we obtain:

\[
c\Gamma^2 \sum_{i=1}^{n} \frac{(u^\top z_i)^2}{z_i^\top S z_i} + d\Gamma^2 \sum_{i=1}^{n} (u^\top z_i)^2 = c \sum_{i=1}^{n} \frac{(u^\top z_i)^2}{z_i} + d \sum_{i=1}^{n} (u^\top z_i)^2.
\]

(B.3)

Using the fact that the product of two positive definite matrices has positive eigenvalues (Horn and Johnson, 1985, p.465), following two inequalities can be derived by straightforward computations:

\[
\lambda_1^2 z_i^\top z_i \leq z_i^\top S^\top S z_i \leq \lambda_q^2 z_i^\top z_i,
\]

(B.4)

where \( \lambda_1 \) and \( \lambda_q \) are the largest and the smallest eigenvalues respectively. It is clear that if \( \lambda_1 > 1 \) or \( \lambda_q < 1 \) then inequalities in (B.5) contradicts the equality in (B.4). Therefore all eigenvalues of \( S \) need to be equal to one which implies \( S = I \) or \( \Sigma_1 = \Sigma_2 \) and the proof is complete.

C  Calculating the parameter \( \mathcal{A} \) in KL-divergence

The distribution of \( K(z) \) in (4.14) is a weighted sum of a Dirichlet random variable.

\[
Z \sim \sum_{j=1}^{s} l_j d_j, l_s < Z < l_1,
\]

(C.1)

where \( D = (d_1,...,d_s) \) is distributed according to Dirichlet with parameters \( r_1/2,..,r_s/2 \) and \( l_j \)'s are the eigenvalues of \( \Sigma \) with respective multiplicities \( r_j \), \( j = 1,...,s \).

A random variable \( D = (d_1,...,d_l)^\top \) is said to have a Dirichlet distribution with parameters \( (a_1, ..., a_l) \) if its p.d.f. is given by:

\[
p_D(d) = \begin{cases} 
\frac{\Gamma(\sum_{i=1}^{l} a_i)}{\prod_{i=1}^{l} \Gamma(a_i)} d_1^{a_1-1} \cdots d_l^{a_l-1}, & d_i = 1 - \sum_{i=1}^{l-1} d_i, \\
0, & 0 \leq d_i \leq 1, a_i > 0, i = 1, ..., l, \\
\text{elsewhere.} & 
\end{cases}
\]

(C.2)
It can be shown that if \( Y_1, \ldots, Y_s \) are independent chi-square variables having \( r_1, \ldots, r_s \) degrees of freedom, and \( Y = \sum_{j=1}^{s} Y_j \), then \((Y_1/Y, \ldots, Y_s/Y)^{\top}\) is Dirichlet distribution with parameters \((r_1/2, \ldots, r_s/2)\).

The random variable \( Z \) in (C.1) can be expressed as \( Z = \sum_{i=1}^{q} l_i Y_i / Y \). Equivalently, if \( N_1, \ldots, N_q \) are normally distributed, then \( z \) can be written as \( \sum_{i=1}^{q} \lambda_i N_i^2 / \sum_{i=1}^{q} N_i^2 \), where \( \lambda_i \)'s are eigenvalues of \( \Sigma \).

The moments of \( z \) in (C.1) were computed in (Provost and Cheong, 1998). For the first moment, we have:

\[
E[Z] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ \frac{q}{\sum_{i=1}^{q} \lambda_i x_i^2} \right\} \prod_{i=1}^{q} \frac{1}{\sqrt{2\pi}} \exp(-x_i^2/2) \, dx_1 \cdots dx_q. \tag{C.3}
\]

Knowing that

\[
\left( \sum_{i=1}^{q} x_i^2 \right)^{-1} = \int_{0}^{\infty} \exp \left\{ -t \left( \sum_{i=1}^{q} x_i^2 \right) \right\} dt \tag{C.4}
\]

and replace that in (C.3), we obtain:

\[
E[Z] = \sum_{i=1}^{q} \int_{0}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \lambda_i x_i^2 \prod_{i=1}^{q} \frac{1}{\sqrt{2\pi}} \exp(-x_i^2(2t + 1)/2) \, dx_1 \cdots dx_q dt. \tag{C.5}
\]

Simplifying the integral, we get out:

\[
E[Z] = \sum_{i=1}^{q} \int_{0}^{\infty} \lambda_i (2t + 1)^{-q/2 - 1} dt. \tag{C.6}
\]

Therefore, we obtain:

\[
E[Z] = \frac{1}{q} \sum_{i=1}^{q} \lambda_i = \frac{1}{q} \text{tr}(\Sigma). \tag{C.7}
\]

We could not find an analytic solution for \( E[\log Z] \). Since generating samples for random variable \( z \) is easy, one can use Monte-Carlo integration for computing the expression. Consider \( \{z_i\}_{i=1}^{n} \) are our random samples, then \( E[\log Z] \) can be approximated using:

\[
E[\log Z] \approx \sum_{i=1}^{n} \log z_i. \tag{C.8}
\]

Another approach for computing the expected logarithm of \( z \) is using the Laplace transform. Expected logarithm can be written as subtraction of two expectations:

\[
E[\log Z] = E[\log \sum_{i=1}^{q} \lambda_i N_i^2] - E[\log \sum_{i=1}^{q} N_i^2]. \tag{C.9}
\]

Therefore the problem is reduced into calculating the expected logarithm of weighted chi-square distribution. The Laplace transform of the density function of random variable \( U = \sum_{i=1}^{q} \lambda_i N_i^2 \) is given by:

\[
\prod_{i=1}^{q} \left( 1 + 2\lambda_i s \right)^{-\frac{q}{2}}. \tag{C.10}
\]
Knowing that the Laplace transform of natural logarithm has the following form:

\[ \Psi(1) - \log(s) \frac{1}{s}, \]  

(C.11)

and using the properties of Laplace transform, the expected logarithm could be evaluated using the following integral in the complex domain:

\[ \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{\Psi(1) - \log(s)}{s} \prod_{i=1}^{q} \left(1 - 2\lambda_i s\right)^{-\frac{1}{2}} ds. \]  

(C.12)

where \( \sigma \) is a positive constant smaller than \( 1/\lambda_q \). This integral can be computed numerically (See for example (Abate and Whitt, 1995)). It is worth mentioning that a similar procedure has been used in the literature for evaluating the cumulative distribution function of \( u \) (Imhof, 1961).

When the ratio \( \lambda_1/\lambda_q \) as well as the dimensionality \( q \) are small, a more efficient algorithm for calculating aforementioned expected logarithm can be derived. One can approximate the probability density of \( u \) as an infinite weighted sum of gamma distribution functions (Ruben, 1962; Kotz et al., 1967):

\[ P_U(u) = \sum_{j=0}^{\infty} c_j p_{\text{ga}}(u; \frac{q}{2} + j, 2\beta), \]  

(C.13)

where

\[ c_k = \frac{1}{k} \sum_{r=0}^{k-1} d_{k-r} c_r, \]  

(C.14)

\[ d_k = \frac{1}{2} \sum_{j=1}^{q} (1 - \beta \lambda_j^{-1})^k, \]  

(C.15)

\[ c_0 = \prod_{j=1}^{q} \sqrt{\beta \lambda_j^{-1}}. \]  

(C.16)

Then for the expected logarithm, we obtain:

\[ \mathbb{E}[\log U] = \sum_{j=0}^{\infty} c_j \left( \Psi\left(\frac{q}{2} + k\right) + \log(2\beta) \right) \]  

(C.17)

Ruden explained the effect of different \( \beta \)s on the behavior of the series expansion and gives the following \( \beta \) as a close to optimal one in calculating the cumulative distribution function:

\[ \beta = \frac{2\lambda_1 \lambda_q}{\lambda_1 + \lambda_q}. \]  

(C.18)

Using this \( \beta \), the sum of the weights is equal to one:

\[ \sum_{j=0}^{\infty} c_j = 1, \]  

(C.19)
and also knowing that the following relationship holds for the digamma function:

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

(C.20)

then the (C.17) can be simplified:

\[
\mathbb{E}[\log U] = c_0(\Psi\left(\frac{q}{2}\right) + \sum_{j=1}^{\infty} c_j(\Psi\left(\frac{q}{2}\right) + \sum_{l=1}^{j} 1/l) + \log(2\beta) \\
= \Psi\left(\frac{q}{2}\right) + \log(2\beta) + \sum_{j=1}^{\infty} \sum_{l=1}^{j} \frac{1}{l} c_j \\
= \Psi\left(\frac{q}{2}\right) + \log(2\beta) + \sum_{l=1}^{\infty} \left(\frac{1}{l} \sum_{k=0}^{l-1} c_k\right) \\
= \Psi\left(\frac{q}{2}\right) + \log(2\beta) + \sum_{l=1}^{\infty} \left(\frac{1}{l} (1 - \sum_{k=0}^{l-1} c_k)\right). \tag{C.21}
\]

Therefore we get the following expression for the expected logarithm of \( z \):

\[
\mathbb{E}[\log Z] = \mathbb{E}[\log \sum_{i=1}^{q} \lambda_i N_i^2] - \mathbb{E}[\log \sum_{i=1}^{q} N_i^2] \\
= \log(\beta) + \sum_{l=1}^{\infty} \left(\frac{1}{l} (1 - \sum_{k=0}^{l-1} c_k)\right). \tag{C.22}
\]

To approximate it, we cut the series coefficients and only finite terms are used to evaluate the expectation.

\[
\hat{\mathbb{E}}[\log Z] = \log(\beta) + \sum_{l=1}^{L} \left(\frac{1}{l} (1 - \sum_{k=0}^{l-1} c_k)\right). \tag{C.23}
\]

It is possible to compute the bound for the approximation error. The following bound exists for \( c_k \) (Kotz et al., 1967):

\[
|c_k| \leq c_0 \epsilon^k \frac{\Gamma(q/2 + k)}{\Gamma(q/2)k!}, \tag{C.24}
\]

where \( \epsilon = \max_j |1 - \beta \lambda_j| \).

Consider \( \hat{c}_k \) to be the corresponding bound for \( c_k \),

\[
\hat{c}_k = c_0 \epsilon^k \frac{\Gamma(q/2 + k)}{\Gamma(q/2)k!}, \tag{C.25}
\]

then we have:

\[
\frac{\hat{c}_{k+1}}{\hat{c}_k} = \frac{q/2 + k}{k} \epsilon. \tag{C.26}
\]

Since \((q/2 + k)/k\) decreases with increasing \( k \) we have:

\[
\frac{\hat{c}_{k+1}}{\hat{c}_k} \leq \left(\frac{q/2 + k}{k} \epsilon\right)^i. \tag{C.27}
\]
First we calculate the following bound:

\[
\sum_{k=l}^{\infty} \hat{c}_k \leq \hat{c}_l \sum_{k=0}^{\infty} \hat{c}_l^k = \hat{c}_l \frac{1}{1 - \epsilon_l}.
\] (C.28)

Which is true only if \(l\) is large enough such that \(\epsilon_l < 1\). For the total approximation error, we obtain:

\[
\mathbb{E}[\log Z] - \tilde{\mathbb{E}}[\log Z] = \sum_{l=L}^{\infty} \left( \frac{1}{l} \sum_{k=l}^{\infty} c_k \right)
\leq \sum_{l=L}^{\infty} \frac{1}{l} \hat{c}_l \frac{1}{1 - \epsilon_l}
\leq \frac{\hat{c}_L}{L(1 - \epsilon_L)^2}
\leq c_0 \frac{\epsilon^L}{L \left(1 - \frac{q/2 + L}{L} \epsilon\right)^2} \frac{\Gamma(q/2 + L)}{\Gamma(q/2)L!}.
\] (C.29)

Given the bound we are able to calculate the expected logarithm with a given accuracy.