Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ADAPTIVE ESTIMATION IN ELLIPTICAL DISTRIBUTIONS WITH EXTENSIONS TO HIGH DIMENSIONS

By Sharmodeep Bhattacharyya[*,‡] and Peter J. Bickel[†,‡]

University of California, Berkeley[‡]

*Abstract*The goal of this paper is to propose efficient and adaptive regularized estimators for the nonparametric component, mean and covariance matrix in both high and fixed dimensional situations. Although, semiparametric estimation of elliptical distribution has also been discussed in [8], we wish to expand the model in two ways. First, study adaptive estimation methods with a novel scheme of estimating the nonparametric component and second, we perform regularized estimation of Euclidean parameters of the elliptical distribution such that high dimensional inference of the Euclidean parameters under certain additional structural assumption can be carried out. Some methods have already been developed. But we extend the work in [5] [6] [10] [18] [19]. The estimate of elliptic densities can also be used to approximately estimate certain sub-class of log-concave densities by using results from convex geometry. The problem of estimation of mixture of elliptical distributions is also important in clustering, as the level sets produce disjoint elliptical components, which can be viewed as model of clusters of specific shape high dimensional space. The regularized estimation of mixture of elliptical distributions will also lead to an algorithm for finding elliptical clusters in high dimensional space under highly relaxed tail conditions.

1. Introduction.

We consider the estimation of semi parametric family of elliptic distributions for the purpose of data description and classification (regression and clustering). The class of elliptically contoured or elliptical distributions provide a very natural generalization of the class of Gaussian distribution. An elliptical density has elliptic contours like a Gaussian distribution, but can have either heavier or lighter tails than the Gaussian density. The class of elliptical distributions is also very attractive for statistical inference as it has the same location-scale Euclidean parameters as in Gaussian distribution with an additional univariate function parameter.

[*] Supported in part by NSF grant

[†] Supported in part by NSF grant

AMS 2000 subject classifications: Primary 62G05, 62G35, 62F35; secondary 62H12, 62J07

Keywords and phrases: Elliptical density, Monotone density estimation, Robustness, High-dimensional inference, Covariance estimation, High-dimensional Regression

There has been extensive work done on estimation of Euclidean parameters for elliptical distributions. Adaptive and efficient estimation of the Euclidean parameters were addressed by Bickel et.al. (1993) [8], Bickel (1982) [4] and Anderson et al (1986) [1]

It may be argued that semi parametric family is too restrictive and one instead should focus on the more general family of shape-constrained densities and their mixtures. This are has been heavily studied theoretically in different contexts since the seminal work of Grenander [23] on nonparametric maximum likelihood estimation of monotone univariate density. In particular the natural generalization of Gaussian and elliptical families the log-concave densities and their generalizations have received much attention (algorithmic [52] [46], theoretical [16] [14] and extensions [34] [2]). However, for all these problems, estimation of densities in large dimensions become a computationally challenging problem. The algorithm proposed by Cule et.al. (2010)[14] Koenker and Mizera [34] works for estimation of multivariate log-concave densities but is too slow to work in large dimensions. So, the application of such models to clustering and classification are very limited. That is why, we consider a smaller class of multivariate density functions, which can be estimated with relative ease for large dimensions.

Semiparametric estimation of elliptic densities for fixed dimension were first addressed in Stute and Werner (1991) [48] by using kernel density estimators for estimating the function parameter. Cui and He (1995) [13] addressed a similar problem. Liebscher (2005) [38] used transformed data and kernel density estimators for estimating the unknown function parameter. Battey and Linton (2012) [3] used finite mixture sieves estimate for the function parameter by using the scale mixture of normal representation of consistent elliptic density. ML-Estimation of only the Euclidean parameters of the elliptical distributions were considered in the work of Tyler [50] and Kent and Tyler [32]. These works were extended to get shrinkage estimates of covariance matrix of the elliptical distributions by Chen, Wiesel and Hero (2011) [12] and Wiesel (2012) [53] with the shrinking being towards identity, diagonal or given positive-semi-definite matrix. In all of these works the theoretical properties of the estimators were also addressed. We focus on maximum likelihood estimation of the elliptic densities using penalized likelihood functions. The estimation consists of nonparametric estimation of a univariate function as well as parametric estimation of the location and scale parameters.

Recently, there has been a general focus on statistical inference in high-dimensional problems, with examples of high-dimensional data coming from biology especially genetics and neuroscience, imaging, finance, atmospheric science, astronomy and so on. In most of these cases, the number of dimensions of data is nearly of the same order or greater than the number of data points. So, the appropriate asymptotic framework is as both $n \to \infty$ and $p \to \infty$, where n is the number of data points and p is the dimension of data points.

Little is possible without some restriction on parameters. In case of regression restrictions are put on regression parameters such as sparsity and size and on the design matrix such as incoherence and related conditions. A variety of regularization methods have been studied such as [40] [42] [55] [7] [11] and effective algorithms proposed such as [17] [41]. A good reference book on all different forms of regularization and algorithms is Bühlmann and Van de Geer (2011) [9]. Another problem that has been considered is covariance or precision matrix estimation where again you need regularization if you has to have consistency as $p, n \to \infty$. Again there has been considerable theoretical work [5] [6] [10] [21] [45] [36] focussing on Gaussian and sub-Gaussian distributions, except some like [37], which are distribution-free. However, there has been little focus on tail behavior, except some on sample covariance matrix behavior [47] [15], although there has been earlier work in the robustness literature (Chapter 5 of [25], [29]). Here we consider elliptical distributions, which can have arbitrary tail behavior. For such distributions, we have attempted to estimate sparse mean and covariance matrices using penalized likelihood loss function. Thus, we have generalized the class of regularized covariance estimators, so that estimation of sparse covariance and precision matrix becomes possible under arbitrary tail behavior of the underlying distribution in high-dimensions. We have tried to provide a framework of semiparametric inference of elliptical distributions for Euclidean parameters as well as mixtures.

1.1. *Contributions and Outline of the Paper.*

So, in this paper we have done the following.

1. We develop estimation procedure for the density generator function of the elliptical distribution in a log-linear spline form in Section 3 and derive respective error bounds.
2. We use the estimate of the density generator function of elliptical distribution to adaptively estimate Euclidean parameters of elliptical distribution in Section 4. We show how using appropriate regularization we can obtain, under conditions similar to those of [7], [45] and [36], consistent estimates of Euclidean parameters for both fixed dimensional case and when $p, n \to \infty$ in Section 5.3.
3. Develop feasible algorithms for all these methods in Section 5 and illus-

trate our method by simulation and one real data example in Section 7 and Section 8.

4. Extend the results to three special cases - (a) Estimation of Covariance and Precision matrix (b) Regression with Elliptical errors and (c) Clustering via mixtures of elliptical distribution in Section 6.

We give the main definitions and results in Section 2.

2. Elliptical Distributions and Main Results.

The formal definition of *elliptically symmetric* or *elliptical* distributions is given in the following way in [20] -

Definition 1. *Let X be a p-dimensional random vector. X is said to be 'elliptically distributed' (or simply 'elliptical') if and only if there exist a vector $\mu \in \mathbb{R}^p$, a positive semidefinite matrix $\Omega \equiv \Sigma^{-1} \in \mathbb{R}^{p\times p}$, and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $t \mapsto \phi_{X-\mu}(t)$ of $X - \mu$ corresponds to $t \mapsto \phi(t^T\Sigma t)$, $t \in \mathbb{R}^p$.*

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu, \Omega)$. Then the density function $f(\cdot; \mu, \Omega)$ is of the form

$$f(x; \mu, \Omega) = |\Omega|^{1/2} g_p\left((x-\mu)^T\Omega(x-\mu)\right) \tag{2.1}$$

where $\theta = (\mu, \Omega) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and covariance parameters respectively with $\mu \in \mathbb{R}^p$ and $\Omega \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x)dx = 1$$

$\mathbb{R}^+ = [0, \infty)$. g_p is also called the *density generator* of the elliptical distribution in $\mathbb{R}^p$.

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\mu_0||_0 = s_1$ and $||\Omega^-||_0 = s_2$, where, $||\cdot||_0$ calculates the number of nonzero entries in the vector or the matrix in vectorized form. Small values of s_1 and s_2 indicates *sparsity*. We first consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

Now, if we define $Y = (X-\mu)^T\Sigma^{-1}(X-\mu)$, then, by transformation of variables, Y has the density

$$f_Y(y) = c_p y^{p/2-1} g_p(y) \quad y \in \mathbb{R}^+ \tag{2.2}$$

where, $c_p = \frac{\pi^{p/2}}{\Gamma(p/2)}$. So, we can now use estimate of f_Y from the data $Y_1, \ldots, Y_n$ to get an estimate of the non-parametric component g_p. From here onwards we shall drop the suffix and denote g_p by g.

We divide the family of density generators into two different classes - *monotone* and *non-monotone*. The following proposition gives the equivalence between *monotone* density generator and *unimodal* elliptical density.

PROPOSITION 2. *The density generator g_p is monotonically non-increasing if and only if the elliptical density f with density generator g_p is unimodal. Also the mode of density f is at μ.*

PROOF. *If:* If g_p is not monotonically increasing, then, there exists a mode of g_p which is not at zero. Let that mode be at ν. By symmetry of f now f has a mode at all points at in an ellipse around μ whose points are $(\nu - \mu)^T \Sigma^{-1} (\nu - \mu)$. So, f does not remain unimodal anymore. So, if f is unimodal, then, g_p has to be monotonically non-increasing.
Only If: This part is obvious. □

So, we divide the class of elliptical densities into two - *unimodal* and *multimodal.* Unimodal elliptical densities have monotone density generator, where as, multimodal elliptical density has non-monotone density generator. Examples of unimodal elliptical density include normal, t, logistic distributions, and examples of multimodal elliptical density include a subclass of Kotz type and multivariate Bessel type densities. See Table 3.1 (pp. 69) of [20] for more examples.

Another desirable property of the class of elliptical distributions is *consistency.*

DEFINITION 3. *An elliptical distribution with density $f_p(\cdot; \mathbf{0}_p, \mathbf{I}_p)$ and density generator g_p is said to possess* ***consistency*** *property if and only if*

$$\int_{-\infty}^{\infty} g_{p+1}\left(\sum_{i=1}^{p+1} x_i^2\right) dx_{p+1} = g_p\left(\sum_{i=1}^{p} x_i^2\right) \tag{2.3}$$

This consistency property of elliptical distributions is quite desirable and natural, since it ensures that marginal distribution of the elliptical distributions also follow the elliptical distribution with same density generator. This property becomes indispensable if we go for high-dimensional situation, since, in high-dimensions, we have to depend on the projection of the random variable in low-dimensions and if in the low-dimensions, we have a different density generator, we can not devise any adaptive estimator. Equivalent conditions for the consistency property is given in Theorem 1 of [31]. We just mention an excerpt of that Theorem in form of Lemma below

LEMMA 4 ([31]). *Let g_p be the density generator for a p-variate random variable, X_p following elliptical distribution with mean and covariance matrix parameters $(\mathbf{0}_p, \mathbf{I}_p)$. Then, X_p follows consistent elliptical distribution if and only if $X_p \overset{d}{=} Z_p/\sqrt{\xi}$, where, Z_p is a p-variate normal random variable with parameters $(\mathbf{0}_p, \mathbf{I}_p)$ and $\xi > 0$ is some random variable unrelated with p and independent of Z_p.*

Examples of elliptical distributions with consistency property include multivariate Gaussian distributions, multivariate t-distributions, multivariate stable laws and such, where as, examples of elliptical distribution without consistency property include multivariate logistic distributions. For more discussion and insight on the issue see [31].

We shall first try to estimate density generator g_p with monotonicity constraint in Section 3.1. Unimodal elliptical density is more commonly seen in practice and is easier to handle. We shall only estimate Euclidean parameters for consistent elliptical distributions in high dimensions.

2.1. *Some Notations.* For any vector $x \in \mathbb{R}^p$, we define,

$$\begin{aligned}
||x||_2 &= \sqrt{\sum_{i=1}^n x_i^2}, \\
||x||_1 &= \sum_{i=1}^n |x_i|, \\
||x||_\infty &= \max\{x_1, \ldots, x_n\}, \\
||x||_0 &= \sum i = 1^n \mathbf{1}(x_i \neq 0)
\end{aligned}$$

For any matrix $M = [m_{ij}]$, we write $|M|$ for the determinant of M, $tr(M)$ for the trace of M, and $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ for the largest and smallest eigenvalues of M, respectively. We write $M^+ = \text{diag}(M)$ for a diagonal matrix with the same diagonal as M and $M^- \equiv M - M^+$. We will use $||M||_F$ to denote the Frobenius matrix norm and $||M|| \equiv \lambda_{\max}(MM^T)$ to denote the operator or spectral norm (also known as matrix 2-norm). We will also write $|\cdot|_1$ for the l_1 norm of a vector or matrix vectorized $|M|_1 = \sum_{i,j} |m_{ij}|$. $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form.

For two numerical sequences a_n and b_n, $a_n = o(b_n)$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$ and $a_n = O(b_n)$ means there exists constant C such that $a_n \leq Cb_n$ for $n \geq N$. Also, for two random or numerical sequences X_n and Y_n, $X_n = o_P(Y_n)$ means that $\frac{X_n}{Y_n} \overset{P}{\to} 0$ and $X_n = O_P(Y_n)$ means that X_n is stochastically bounded by Y_n, that is, given $\varepsilon > 0$ there exists constant C and integer $N \geq 1$, such that $\mathbb{P}\left[\left|\frac{X_n}{Y_n}\right| \leq C\right] \leq \varepsilon$ for $n \geq N$.

2.2. *Main Results.* Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p\left((x-\mu_0)^T \Omega (x-\mu_0)\right) \tag{2.4}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance or precision matrix parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x) dx = 1$$

$\mathbb{R}^+ = [0, \infty)$.

We start with the following assumption -

(A1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] \leq J_1(t, n, p) \tag{2.5}$$

$$\mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] \leq J_2(t, n, p). \tag{2.6}$$

For fixed dimensions, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(n))$, where, $\omega(n) \to 0$ as $n \to \infty$ with given tail bounds. For high dimensions, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(p, n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(p, n))$, where, $\omega(p, n) \to 0$ as $p, n \to \infty$ with given tail bounds.

Now, let us consider the high-dimensional situation. So, in this case, the dimension of Euclidean parameters grows with n. In the high-dimensional situation we assume the additional structure of sparsity imposed on the Euclidean parameters $\theta_0 = (\mu_0, \Omega_0)$. Density generator g comes from a consistent family elliptical distributions and so by Lemma 4, $\Omega_0^{1/2}(X - \mu_0) \overset{d}{=}$

$Z_p/\sqrt{\xi}$. Let us consider the probability density function of X_j to be h $(j = 1, \ldots, p)$ and

$$H(u) \equiv \int_u^\infty h(v)dv \tag{2.7}$$

We consider the following assumptions -

(A2) Suppose that $||\Omega^-||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*.

(A3) $\lambda_{\min}(\Sigma_0) \geq \underline{k} > 0$, or equivalently $\lambda_{\max}(\Omega_0) \leq 1/\underline{k}$. $\lambda_{\max}(\Sigma_0) \leq \overline{k}$.

(A4) The function $H(t)$ defined in Eq. (2.7) and $J_1(t), J_2(t)$ defined in Eq. (2.5), satisfies the following conditions -

(a) there exists a function $\sigma_1(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $c, d > 0$, for $t = O(\sigma_1(p,n))$,

$$p(c\exp(-dt)J_1(t,n,p)J_2(t,n,p))^n \to 0 \text{ as } p, n \to \infty \tag{2.8}$$

(b) there exists a function $\sigma_2(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $d_1, d_2, d_3 > 0$, for $t = O(\sigma_2(p,n))$,

$$d_1 p^2 (H(t)J_1(t,n,p)J_2(t,n,p))^n (\exp(-nd_2 t))(d_3 \exp(-nd_3 t^2)) \to 0 \text{ as } p, n \to \infty \tag{2.9}$$

Let us consider that we have obtained estimators of the Euclidean parameters $\tilde{\mu}$ and $\tilde{\Omega}$ and the nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 5. Let us consider first the fixed dimensional situation. So, in this case, the dimension of Euclidean parameters does not grow with n

THEOREM 5. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (A1-A4),*

(a) $||\mu_0 - \tilde{\mu}||_2 = O_P(\sigma_1(n))$.
(b) $||\Omega_0 - \tilde{\Omega}||_F = O_p(\sigma_2(n))$
(c) $\hat{g}$ is an uniform consistent estimator of g with rate $O_p\left(\left(\frac{\log n}{n}\right)^{1/3}\right)$.

We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow. We have the estimates, $\tilde{\mu}$ and $\tilde{\Omega}$ as stated in Section 5.

THEOREM 6. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (A1)-(A4),*

(a) $||\mu_0 - \tilde{\mu}||_2 = O_P\left(\sqrt{p}\sigma_1(p,n)\right)$
(b) $||\Omega_0 - \tilde{\Omega}||_F = O_P\left(\sqrt{(p+s)}\sigma_2(p,n)\right)$ *for* $\nu = O\left(\sigma_2(p,n)\right)$.
(c) $\hat{g}$ *is an uniform consistent estimator of* g *with rate* $O_p\left(\left(\frac{\log n}{n}\right)^{1/3}\right)$.

We apply the semi parametric inference technique of estimating elliptical distributions to two parametric inference and one semi-parametric inference problems -

(a) In Section 6.1, we apply it for robust regularized covariance and precision matrix estimation in high-dimensions.
(b) In Section 6.2, we apply it for robust regularized regression in high-dimensions.
(c) In Section 6.3, we apply the semi parametric inference technique of estimating elliptical distributions to clustering by devising an inference scheme for mixtures of elliptical distributions.

3. Inference I: Estimation of Density Generator g_p.

We try to find a maximum likelihood estimate for the semiparametric elliptical distribution. The main idea is using non-parametric maximum likelihood estimate (NPMLE) to estimate density generator g_p and then use that NPMLE estimate of g_p, to get a likelihood estimate of the Euclidean parameters. Throughout this section, we shall consider that the Euclidean parameters $\theta = (\mu, \Omega)$ are given and the dimension of data p is fixed.

We shall propose non-parametric maximum likelihood estimates (NPMLE) of density generator g_p under the monotonicity assumption. This is the most common situation for elliptical distributions as monotone density generators gives rise to unimodal elliptical distributions according to Proposition 2. We shall principally focus on this case. We consider this case in Section 3.1.

3.1. *Maximum Likelihood Estimation of Monotone Density Generator.*

The likelihood for (θ, g) is

$$\mathcal{L}(\theta, g|X_1, \ldots, X_n) \quad = \quad \prod_{i=1}^{n} |\Sigma|^{-1/2} g\left((X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right)$$

The log-likelihood is

$$\ell(\theta, g|X_1, \dots, X_n) = -\frac{n}{2}\log|\Sigma| + \sum_{i=1}^{n} \log g\left((X_i - \mu)^T \Sigma^{-1}(X_i - \mu)\right)$$

Let us start with an ideal case when the Euclidean parameters μ and Σ are known. Then, the non-parametric likelihood of g in terms of data Y_i, $i = 1, \dots, n$, obtained by the transformation $Y_i = (X_i - \mu)^T \Sigma^{-1}(X_i - \mu)$ becomes

$$\mathcal{L}(g|\boldsymbol{Y}) = (c_p)^n \prod_{i=1}^{n} Y_i^{\frac{p}{2}-1} g\left(Y_i\right)$$

The log-likelihood ignoring the constants becomes

$$\ell(g|\boldsymbol{Y}) = \sum_{i=1}^{n} \left(\left(\frac{p}{2} - 1\right)\log(Y_i) + \log g\left(Y_i\right)\right)$$

So, the NPMLE $\hat{g}_n$ can be written as

$$\hat{g}_n = \arg\max_g \sum_{i=1}^{n} \left(\left(\frac{p}{2} - 1\right)\log(Y_i) + \log g\left(Y_i\right)\right) = \arg\max_g \sum_{i=1}^{n} \left(\log g\left(Y_i\right)\right)$$

Now, $g(y)$ is a monotonically non-increasing function, however, $f_Y(y)$ as defined in (2.2), which is the density of Y_i's, is not monotone. But, we can still formulate the problem as a generalized isotonic regression problem as done in Example 1.5.7 (pp. 38-39) in Robertson et.al. (1988) [44].

First note that the NPMLE $\hat{g}_n$ must be constant on intervals $(Y_{(i-1)}, Y_{(i)}]$, $i = 1, \dots, n$ $(Y_0 = 0)$, where $Y_{(i)}$ is the i^{th} order statistic of Y_i, and $\hat{g}_n$ must be zero on $(Y_{(n)}, \infty)$. It follows by observing that if $\hat{g}_n$ is not constant on $(Y_{(i-1)}, Y_{(i)}]$, then, we can always construct another estimator $\tilde{g}_n = (Y_{(i)} - Y_{(i-1)})^{-1} \int_{Y_{(i-1)}}^{Y_{(i)}} \hat{g}_n(t)dt$ constant on $(Y_{(i-1)}, Y_{(i)}]$ which gives larger likelihood than $\hat{g}_n$. So, the NPMLE $\hat{g}_n$ has to be piecewise constant and left-continuous.

Hence the problem of finding NPMLE boils down to the optimization problem on
$(g_1, \dots, g_n)$, where $g_i = g(Y_i)$ for $i = 1, \dots, n$. The optimization problem is defined as

$$\max_{(g_1,\dots,g_n)} \sum_{i=1}^{n} \log g_i \tag{3.1}$$

such that

$$\sum_{i=1}^{n} c_p \int_{Y_{(i-1)}}^{Y_{(i)}} y^{p/2-1} g_i dy = \frac{2c_p}{p} \sum_{i=1}^{n} g_i \left(Y_{(i)}^{p/2} - Y_{(i-1)}^{p/2} \right) = 1 \tag{3.2}$$

and

$$g_1 \geq g_2 \geq \cdots \geq g_n \tag{3.3}$$

The above defined optimization problem can be solved in the similar way as described in Example 1.5.7 of [44] (pp. 38-39) and by following Theorem 1.4.4 of [44] the solution can be written as

$$\hat{g}_i = \frac{p}{2c_p} \min_{s \leq i-1} \max_{t \geq i} \frac{F_n(Y_{(t)}) - F_n(Y_{(s)})}{Y_{(t)}^{p/2} - Y_{(s)}^{p/2}} \tag{3.4}$$

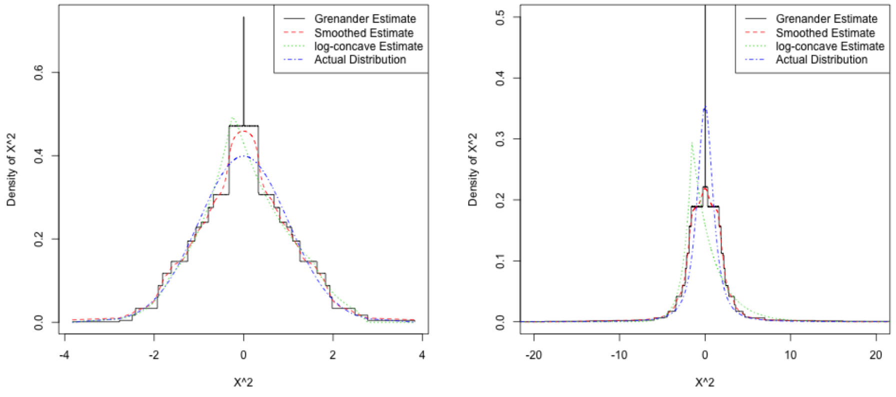
where, F_n is the empirical cumulative distribution function (CDF) of the data $(Y_1, \ldots, Y_n)$. The NPMLE $\hat{g}_n$ is given by

$$\hat{g}_n(y) \quad = \quad \begin{cases} \hat{g}_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

Note that, the NPMLE $\hat{g}_n$ is quite related to the Grenander estimator [23] of monotonically non-increasing densities. The Grenander estimator is a piece-wise constant or histogram type density estimate, where the constant values come from the left-derivative of the least concave majorant of the empirical CDF function. Similarly, NPMLE $\hat{g}_n$ is also a piece-wise constant or histogram type density generator estimate, where the constant values come from the left derivative of the least concave majorant of the empirical CDF plotted against the abscissa of $\frac{2c_p}{p} y^{p/2}$ instead of y. Figure 1 gives an example of the NPMLE $\hat{g}_n$ for a simulated small sample case.

The above description of the NPMLE $\hat{g}_n$ also provides us with a cautionary note while implementing this estimator. The transformation $y \mapsto \frac{2c_p}{p} y^{p/2}$ highly stretches the abscissa for large values of p, so for numerical implementation of the algorithm care should be taken so that machine precision problems does not hurt the computation of the estimator. However, in this discourse we shall not dwell on these numerical issues. Some thoughts on this issue is given in Section 7.

The asymptotic properties NPMLE $\hat{g}_n$ is provided in Lemma 7. The asymptotic properties are quite as expected of isotonic regression estimates. The proof borrows techniques from Groeneboom (1985) [24], Jonker and Van der Vaart (2002) [30] and Example 3.2.14 of citeMR1385671.



(a)

(b)

FIGURE 1. *(a) Estimated univariate normal density curve using several NPML techniques (b) Estimated univariate t with 2 degrees of freedom density curve using several NPML techniques.*

LEMMA 7. *Let g be the monotonically decreasing density generator of the elliptical distribution and the NPMLE of g is $\hat{g}_n$, whose definition is given in (3.5). Suppose that g is continuously differentiable on the interval $(0, \infty)$ with derivative g' bounded away from 0 (from above) and $-\infty$. Then,*

(a) For any $y > 0$, as $n \to \infty$,

(3.6)
$$n^{1/3}\left(\hat{g}_n(y) - g(y)\right) \xrightarrow{w} \left|4g(y)g'(y)\right|^{1/3} \, arg\max_{h} \left\{W(h) - \sqrt{c_p y^{p/2-1}} h^2\right\}.$$

where, W is the Wiener process on $(0, 1)$.

(b) For any $x_n \to \infty$, $\delta_n = O\left(n^{-1/3}(\log n)^{1/3}\right)$, $U > x_n\delta_n$ and $n \to \infty$

$$\mathbb{P}\left[\sup_{x_n\delta_n \leq y \leq U} \left(\frac{n}{\log n}\right)^{1/3} |\hat{g}_n(y) - g(y)| \geq x\right] \leq O\left(\frac{1}{\sqrt{x}}\right). \tag{3.7}$$

so, we have,

$$\sup_{x_n\delta_n \leq y \leq U} |\hat{g}_n(y) - g(y)| = O_P\left(\left(\frac{\log n}{n}\right)^{1/3}\right). \tag{3.8}$$

(c) For any $U > 0$*, if* $||\hat{g}_n - g||_1 = \int_0^U |\hat{g}_n(y) - g(y)|dy$*, then, as* $n \to \infty$

$$n^{1/3}\mathbb{E}||\hat{g}_n - g||_1 \to \int_0^U |4\mathbb{E}|V_y|g'(y)g(y)|dy \tag{3.9}$$

where, $V_y = \mathit{arg}\max_h \left\{W(h) - \sqrt{c_p y^{p/2-1}}h^2\right\}$.

Proof. (a) Let us define a stochastic process $\{\hat{s}_n(a) : a > 0\}$ by

$$\hat{s}_n(a) = \arg\max_s \left\{F_n(s) - \frac{2ac_p}{p}s^{p/2}\right\}$$

where the largest value is chosen when multiple maximizers exist. It is easy to see that $\hat{g}_n(t) \le a$ if and only if $\hat{s}_n(a) \le t$. It follows that,

$$\mathbb{P}\left(n^{1/3}\left|\frac{g(t)g'(t)}{2}\right|^{1/3}(\hat{g}_n(t) - g(t)) \le x\right) = \mathbb{P}\left(\hat{s}_n\left(g(t) + \delta_n\right) \le t\right)$$

where, $\delta_n = xn^{-1/3}\left|\frac{g(t)g'(t)}{2}\right|^{1/3}$. By definition,

$$\hat{s}_n(a + \delta_n) = \sup\left\{s \ge 0 : F_n(s) - \frac{2c_p}{p}(a + \delta_n)s \text{ is maximal}\right\}$$

Hence, we can write,

$$\hat{s}_n(a + \delta_n) = \sup\left\{s \ge 0 : \sqrt{n}\left(F_n(s) - F(s)\right) + \sqrt{n}\left(F(s) - \frac{2c_p}{p}(a + \delta_n)s\right) \text{ is maximal}\right\}$$

By Hungarian embedding theorem [35],

$$\sqrt{n}\left(F_n(t) - F(t)\right) = B_n(F(t)) + O_P\left(n^{-1/2}\log n\right)$$

where, $(B_n, n \in \mathbb{N})$ is a sequence of Brownian bridges, constructed on the same space as F_n and where, F is the CDF of $Y = (X - \mu)^T\Sigma^{-1}(X - \mu)$. So by (2.2),

$$F'(t) = f(t) = c_p t^{p/2-1}g(t)$$
$$f'(t) = c_p t^{p/2-1}g'(t) + \frac{c_p(p-2)}{2}t^{p/2-2}g(t)$$

So, the limiting distribution of $n^{1/3}(\hat{s}_n(a + \delta_n) - t)$ will be the same as limiting distribution of $n^{1/3}(s_n(a+\delta_n) - t)$, where, $s_n(b)$ is the location of the maximum of the process $\left\{B(F(s)) + \sqrt{n}(F(s) - \frac{2c_p}{p}bs^{p/2}), s \ge 0\right\}$ and B is a standard Brownian bridge on $[0, 1]$.

Now, location of the maximum of the process

$$\Big\{B(F(s)) + \sqrt{n}(F(s) - \frac{2c_p}{p}(a+\delta_n)s^{p/2}), s \geq 0\Big\}$$

behaves as $n \to \infty$ as the location of maximum of the process

$$\Big\{B(F(t)+f(t)(s-t)) + \sqrt{n}(F(t)+f(t)(s-t)+\frac{f'(t)}{2}(s-t)^2 - \frac{2c_p}{p}(a+\delta_n)s^{p/2}), s \geq 0\Big\}$$

Consider, $a = g(t)$, $c = -\frac{g'(t)}{2}$ and $h = \left(\frac{nc^2}{a}\right)^{1/3}(s-t)$, location of the maximum of above mentioned process behave as location of the maximum of following process as $n \to \infty$

$$\begin{aligned}\{B(F(t)+f(t)(s-t)) + \sqrt{n}(F(t)+f(t)(nc^2d/a)^{-1/3}h + \frac{f'(t)}{2}(nc^2/a)^{-2/3}h^2 \\ -\frac{2c_p}{p}(a+\delta_n)\left(t+(nc^2/a)^{-1/3}h\right)^{p/2}), h \in \mathbb{R}\}\end{aligned}$$

and as $n \to \infty$ it is equivalent to the location of the maximum of the process

$$\begin{aligned}\{B(F(t)+f(t)(s-t)) + \sqrt{n}(F(t)+f(t)(nc^2/a)^{-1/3}h + \frac{f'(t)}{2}(nc^2/a)^{-2/3}h^2 \\ -\frac{2c_p}{p}(a+\delta_n)\left((p/2)t^{p/2-1}(nc^2/a)^{-1/3}h + \binom{p/2}{2}t^{p/2-2}(nc^2/a)^{-2/3}h^2\right)), h \in \mathbb{R}\}\end{aligned}$$

and as $n \to \infty$ it is equivalent to the location of the maximum of the process

$$\begin{aligned}\{B(F(t)+f(t)(s-t)) - \sqrt{n}((c_p/2)t^{p/2-1}g'(t)(nc^2/a)^{-2/3}h^2 \\ +(c_p\delta_n t^{p/2-1}(nc^2/a)^{-1/3}h),\ h \in \mathbb{R}\}\end{aligned}$$

Since, a Brownian bridge behaves locally as a Brownian motion in $(0,1)$, the limiting distribution of

$$\begin{aligned}\{W(c_p t^{p/2-1}g(t)(nc^2/a)^{-1/3}h) - \sqrt{n}(c_p t^{p/2-1}(g'(t)/2)(nc^2/a)^{-2/3}h^2 \\ +(c_p\delta_n t^{p/2-1}(nc^2/a)^{-1/3}h),\ h \in \mathbb{R}\}\end{aligned}$$

where, W is the Wiener process on $(0,1)$. Now, by writing the values of a, c and $\delta_n = xn^{-1/3}\left(d(t)ac\right)^{1/3}$ and using Brownian scaling, we get that

$$\begin{aligned}\{\sqrt{c_p}t^{(p/2-1)/2}a^{2/3}(nc^2)^{-1/6}W(h) - c_p t^{(p/2-1)}a^{2/3}(nc^2)^{-1/6}h^2 \\ -c_p t^{(p/2-1)}a^{2/3}(nc^2)^{-1/6}xh,\ h \in \mathbb{R}\}\end{aligned}$$

The location of maximum of the above process is equivalent to the location of maximum of the process

$$\left\{W(h) - \sqrt{c_p t^{p/2-1}}\left(h^2 + xh\right),\ h \in \mathbb{R}\right\}$$

Let

$$V(a) \equiv \arg\max_h \left\{W(h) - \sqrt{c_p t^{p/2-1}}(h-a)^2, h \in \mathbb{R}\right\}$$

The, $\{V(a) - a : a \in \mathbb{R}\}$ is a stationary process and $\mathbb{P}(V(a) \leq t) = \mathbb{P}(V(0) \leq t - a)$. So, in summary, as $n \to \infty$

$$\begin{aligned}\mathbb{P}\left(n^{1/3}\left|\frac{g(t)g'(t)}{2}\right|^{-1/3}(\hat{g}_n(t) - g(t)) \leq x\right) &= \mathbb{P}\left(\hat{s}_n(a+\delta_n) - t \leq 0\right) \\ &\to \mathbb{P}\left(V(-x/2) \leq 0\right) = \mathbb{P}\left(2V(0) \leq x\right)\end{aligned}$$

and we prove for each $y > 0$ as $n \to \infty$,

$$n^{1/3}\left(\hat{g}_n(y) - g(y)\right) \xrightarrow{w} \left|4g(y)g'(y)\right|^{1/3} \arg\max_h \left\{W(h) - \sqrt{c_p t^{p/2-1}}h^2\right\}.$$

(b) Let us use the stochastic process $\hat{s}_n(a)$ again for this proof. Now,

$$\begin{aligned}\hat{s}_n(a) &= \arg\max_{s:s\geq 0}\left\{F_n(s) - \frac{2ac_p}{p}s^{p/2}\right\} \\ &= \arg\max_{h:h\geq -\delta_n^{-1}t}\left\{F_n(t+\delta_n h) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2}\right\} \\ &= \arg\max_{h:h\geq -\delta_n^{-1}t}\left\{G_n(t+\delta_n h) + \sqrt{n}\left(F(t+\delta_n h) - F(t) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2}\right)\right\}\end{aligned}$$

where, $G_n(s) = \sqrt{n}\,(F_n(s) - F(s))$ and F_n and F are as defined in part (a). Consider that $s \in (0, L)$. So, $h \in (\delta_n^{-1}t, \delta_n^{-1}(L-t))$. Let us take $a = g(t) + x\delta_n$ with $x > 0$ fixed. Now, by Taylor expansion,

$$\begin{aligned}&F(t+\delta_n h) - F(t) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2} \\ &= f(t)\delta_n h + \frac{f'(t+\xi\delta_n h)}{2}\delta_n^2 h^2 - \frac{2c_p}{p}(g(t)+x\delta_n)(t+\delta_n h)^{p/2} \\ &= c_p t^{p/2-1}\delta_n^2\left(\frac{g'(t)}{2}h^2 - xh\right) + r_n(h) \\ &\quad\begin{cases}\leq -c\delta_n^2 h^2 - \gamma_n x\delta_n^2 h \\ \geq -d\gamma_n\delta_n^2 h^2 - Cx\delta_n^2 h\end{cases}\end{aligned}$$

for certain $c, d > 0$ (since, $g(t)$ is monotonically decreasing) independent of δ_n, t, h. $\gamma_n > 0$ is a lower bound for t and $\gamma_n \to 0$ and $n \to \infty$. Now, if $(\hat{g}_n(t) - g(t)) > x\delta_n$, then, for any $h_0 \in (-t\delta_n^{-1}, 0)$,

$$\sup_{h>0}\left(G_n(t+\delta_n h) - \sqrt{n}\delta_n^2(ch^2 + \gamma_n x h)\right) \geq \left(G_n(t+\delta_n h_0) - \sqrt{n}\delta_n^2(d\gamma_n h_0^2 + Cxh_0)\right)$$

Choose, $h_0 \equiv -\frac{xC}{2d\gamma_n}$ and note that $ch^2 + \gamma_n xh \geq ch^2$ for $h \geq 0$. So, we can write,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in(\max(\delta_n x/(2dC\gamma_n),\gamma_n),U)} (\hat{g}_n(t) - g(t)) > x\delta_n\right)\\
&\quad\leq \mathbb{P}\left(\sup_{t\in(0,U)}\left(G_n(t+\delta_n h) - \sqrt{n}\delta_n^2 ch^2 - G_n(t+\delta_n h_0)\right) \geq \sqrt{n}C^2\delta_n^2\frac{x^2}{4d\gamma_n}\right)\\
&\quad\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in(0,U),j\leq h\leq j+1}(G_n(t+\delta_n h) - G_n(t+\delta_n h_0)) \geq \sqrt{n}\delta_n^2\left(cj^2 + \frac{x^2C^2}{4d\gamma_n}\right)\right)
\end{aligned}$$

We can define the class of functions

$$\mathcal{G}_{n,j} = \{\mathbf{1}(((t+\delta_n' h_0'), (t+\delta_n h)]) : t \in (0,U), j \leq h \leq j+1\}$$

where, $\delta_n' = \delta_n/\gamma_n$ and $h_0' = -Cx/(2d)$ and using Markov inequality we get that

$$\mathbb{P}\left(\sup_{t\in(\delta_n x/(2dC\gamma_n),L)}(\hat{g}_n(t) - g(t)) > x\delta_n\right) \leq \text{const}\sum_{j=0}^{\infty}\frac{\mathbb{E}||G_n||_{\mathcal{G}_{n,j}}}{\sqrt{n}\delta_n^2(j^2 + x^2/\gamma_n)}$$

Now, using bracketing integral entropy bounds from [51] and $\gamma_n = O((\log n)^{-1/3})$, we get that,

$$\begin{aligned}
\text{const}\sum_{j=0}^{\infty}\frac{\mathbb{E}||G_n||_{\mathcal{G}_{n,j}}}{\sqrt{n}\delta_n^2(j^2 + x^2/\gamma_n)} &\leq \sum_{j=0}^{\infty}\frac{1}{\sqrt{n}\delta_n^{3/2}(j+x)^{3/2}}\\
&= o\left(\frac{1}{\sqrt{x}}\right)
\end{aligned}$$

with the last equality coming by taking $\delta_n = O(n^{-1/3})$ and the RHS goes to zero for$x = x_n \to \infty$. Thus, we can combine all the arguments to get that, for any $\epsilon > 0$, there exists a $x_n > 0$ for sufficiently large n with $\delta_n = n^{-1/3}$ and $\gamma_n = (\log n)^{-1/3}$ such that

$$\mathbb{P}\left(\sup_{x_n\delta_n/\gamma_n \leq t \leq U}|\hat{g}_n(t) - g(t)| > x\delta_n\right) \leq O\left(\frac{1}{\sqrt{x}}\right) < \epsilon$$

(c) Follows from (a) and arguments of Groeneboom (1985) [24].

□

3.1.1. *Spline approximation of NPMLE $\hat{g}_n$.* In the previous section, we constructed an isotropic regression based NPMLE $\hat{g}_n$ for density generator g. One of the main problems with NPMLE $\hat{g}_n$ is that it is piece-wise constant and thus discontinuous. This is in general a problem with isotonic estimators. A number of works has been done to address this issue and obtain isotonic continuous or smooth estimators. Some of the approaches are - (a) kernel or spline smoothing of isotonic estimators [22], (b) finding isotonic estimator for smoothed empirical CDF, (c) kernel and spline fitting with additional isotonic constraints on the fitted models. For each of these approaches several different methods have been proposed with proper theoretical justification for most of them. But there still exist some unanswered questions in this domain. However, we shall not go into those questions in this discourse. We present approach (a) version here only.

We have already found the NPMLE $\hat{g}_n$ and proved some of its properties in Lemma 7. Now, consider the density generator

$$g(y) \equiv \exp(\phi(y)) \text{ and } (\phi_1, \ldots, \phi_n) \equiv (\phi(Y_1), \ldots, \phi(Y_n)) \tag{3.10}$$

and the NPMLE in the form

$$\exp(\hat{\phi}_n(y)) \equiv \hat{g}_n(y) \quad = \quad \begin{cases} \hat{g}_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\hat{\phi}_n(y) \quad \equiv \quad \begin{cases} \hat{\phi}_i \equiv log(\hat{g}_i), & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

Now, consider $\varphi(y)$ to be a twice continuously differentiable monotonically non-increasing function with bounded second derivative and

$$(\varphi(Y_1), \ldots, \varphi(Y_n)) \equiv (\hat{\phi}_1, \ldots, \hat{\phi}_n).$$

We consider the problem of estimating monotonically decreasing $\varphi(y)$ with the help of
$(\phi_1, \ldots, \phi_n)$. We have $((Y_1, \phi_1), \ldots, (Y_n, \phi_n))$ as the data and we want to solve regression problem

$$\hat{\phi}_i = \varphi(Y_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{3.12}$$

where, ϵ_i are mean zero random variables with variance σ^2 and exponentially decaying tails. Now, we solve the regression problem by finding the monotone continuous function $\psi(y)$ which minimizes the penalized least-squares loss function

$$L(\psi) = \sum_{i=1}^{n} \left(\hat{\phi}_i - \psi(Y_i)\right)^2 + \lambda \int_0^U (\psi'(t))^2 dt \tag{3.13}$$

Note that the true regression function is $\varphi(y)$. We choose the above smoothing spline loss function in order to get a natural linear spline estimate for the function $\varphi(y)$ and in turn $\phi(y)$ on design points $(Y_1, \ldots, Y_n)$ and thus get a log-linear spline estimate for the density generator $g_p(y)$ on design points $(Y_1, \ldots, Y_n)$.

The algorithm to solve the minimization problem (3.13) was given in [49]. Let us denote the resultant linear spline estimate by $\hat{\psi}_n(y)$. So, our estimate of $\phi(y)$ is a linear spline estimate $\hat{\psi}(y)$ of the form

$$\hat{\psi}_n(y) \quad \equiv \quad \begin{cases} a_i y + b_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

where, $(a_i, b_i))_{i=1}^n$ are estimated by solving the optimization problem in Eq (3.13).

Pal and Woodroofe (2007) [43] provided the asymptotic properties of the estimator $\hat{\phi}_n(y)$ in the Theorem 2 of their paper [43], which we restate in following lemma

Lemma 8.

$$\hat{\psi}(y) = \tau_\lambda(y) + \frac{\nu}{n}\sum_{i=1}^{n} \exp(-\nu(y - Y_i))\epsilon_i + O_P\left(n^{-2/3}\log n\right)\nu + \exp(-\nu y(U-y))O_P(\nu) \tag{3.15}$$

uniformly in λ and in $y \in (0, U)$ with $\nu = \lambda^{-1/2}$ and $\tau_\lambda(y) = \varphi(y) + \lambda\varphi''(y) + o(\lambda)$.

Now, we can use the above lemma and some extensions of it based on [43] to get concentration of $\hat{\psi}_n(y)$ for a special case.

Lemma 9. *Let $\lambda = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right)$ and we minimize loss function in Eq. (3.13) with the λ considered to get $\hat{\psi}(y)$. Suppose also that $\phi(y)$ defined in Eq. (3.10) is twice continuously differentiable with bounded second derivative and ϕ' bounded away from 0 (from above) and $-\infty$. Then as $n \to \infty$ for each Y_i, for some constants, $c_1, c_2 > 0$, we have,*

$$\mathbb{P}\left[\left|\hat{\psi}_n(Y_i) - \phi(Y_i)\right| \geq t\right] \leq c_1 \exp(-c_2 t). \tag{3.16}$$

PROOF. From Lemma 8 using $\lambda = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right)$ and substituting λ in Eq. (3.15), we have that, for $y \in (0, U)$,

$$\begin{aligned}
\hat{\psi}(y) &= \varphi(y) + o(n^{-2/3}) + \frac{\nu}{n}\sum_{i=1}^{n} \exp(-n^{1/3}(y - Y_i))\epsilon_i \\
&\quad + \nu O_P\left(n^{-1/3}\log n\right) + O_P\left(n^{1/3}\exp(-n^{2/3}y(U-y))\right) \\
\hat{\psi}(y) - \varphi(y) &= \frac{\nu}{n}\sum_{i=1}^{n} \exp(-n^{1/3}(y - Y_i))\epsilon_i \\
&\quad + \nu O_P\left(n^{-1/3}\log n\right) + O_P\left(n^{1/3}\exp(-n^{2/3}y(U-y))\right) + o(n^{-2/3})
\end{aligned}$$

Now, the first term of the RHS has sub-Gaussian concentration with rate $O_P(n^{-1/3})$ following Hoeffding's inequality, since ϵ_i are iid sub-Gaussian random variables. The second term is bounded by $\nu||\hat{\Phi} - \Phi||_\infty$ and $o(1/n)\nu$, where, $\hat{\Phi}(y) = \frac{1}{n}\sum_{i:Y_i \le y} \hat{\phi}(Y_i)$ and $\Phi(y) = \int_0^y \phi(y)$. Now, $||\hat{\Phi} - \Phi||_\infty$ has sub-Gaussian tails by Marshall's Lemma and Dvoretsky-Kiefer-Wolfowitz Theorem [33], we have that, $||\hat{\Phi} - \Phi||_\infty = O_P\left((\log n/n)^{2/3}\right)$ with sub-Gaussian concentration. So, the second term has sub-Gaussian concentration with rate $O_P(n^{-1/3})$. The third term is bounded by $\frac{\nu}{n}\exp(-\nu(U-y))\sum_{i=1}^{n}\epsilon_i + \nu\exp(-\nu y(U-y))$ and thus has sub-Gaussian concentration with rate $o_P(n^{-1/3})$.

Now, $(Y_1, \ldots, Y_n) \in (0, U)$. So, given $(Y_1, \ldots, Y_n) \in (0, U)$, we have, for some constants $c_1, c_2 > 0$,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left|\hat{\psi}_n(Y_i) - \hat{\phi}(Y_i)\right| \ge t\right] \le c_1\exp(-c_2 t^2). \tag{3.17}$$

From, Lemma 7(b), we have that,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left|\exp(\phi(Y_i)) - \exp(\hat{\phi}(Y_i))\right| \ge \exp(t)\right] = O\left(\frac{1}{\sqrt{\exp(t)}}\right).$$

So, we have,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left(\frac{\exp(\phi(Y_i))}{\exp(\hat{\phi}(Y_i))} - 1\right) \ge \exp(t)\right] = O\left(\frac{1}{\sqrt{\exp(t)}}\right),$$

which implies,

$$\mathbb{P}\left[\log n\left|\phi(Y_i) - \hat{\phi}(Y_i)\right| \ge C\log(\exp(t) \pm 1)\right] = O\left(\exp(-t/2)\right),$$

for some constant $C > 0$. So, for large $t > 0$, we have, for some constants $c_1 > 0$ and $c_2 > 0$,

$$\mathbb{P}\left[\log n \left|\phi(Y_i) - \hat{\phi}(Y_i)\right| \geq t\right] = c_1 \exp(-c_2 t),$$

Now, combining the above equation with Eq. (3.17), we get that, for each Y_i and large t,

$$\mathbb{P}\left[\left|\hat{\psi}(Y_i) - \phi(Y_i)\right| \geq t\right] \leq c_1 \exp(-c_2 t),$$

and thus the Lemma follows. □

4. Inference II: Estimation of Euclidean Parameters.

The estimation of Euclidean parameters is carried in an iterative fashion. We start with a consistent estimate of the Euclidean parameters and then by using the estimate of density generator in Section 3, we try to get better estimates of the Euclidean parameters. In this section, we shall try to devise the estimation procedure for improving the initial estimate of the Euclidean parameters.

4.1. *Initial Estimates of Euclidean Parameters.*

We have $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. We shall try to give different initial estimates of Euclidean parameters for fixed dimension and high-dimensional cases.

4.1.1. *Fixed dimensional case.*

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [25] is a good source. We can also use sample mean and covariance estimates, as they are also consistent estimates of mean and covariance parameters for the class of Euclidean distributions. We suggest using Stahel-Donoho robust estimator of multivariate location and scatter [39]. Stahel-Donoho estimators $(\hat{\mu}, \hat{\Sigma})$ of (μ_0, Σ_0) are also weighted mean and covariance matrix estimators, which are of the form

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \tag{4.1}$$

$$\hat{\Sigma} = \frac{\sum_{i=1}^n w_i (X_i - \hat{\mu})(X_i - \hat{\mu})^T}{\sum_{i=1}^n w_i} \tag{4.2}$$

where, the weight, w_i is a function on "oulyingness" of a data point X_i from the center $(i = 1, \ldots, n)$. See [25] for more details on weight function $\boldsymbol{w}$.

From Theorem 1 of [25], we get $\sqrt{n}$ consistency of the estimators $(\hat{\mu}, \hat{\Sigma})$. So, we have,

$$\sqrt{n}|\hat{\mu}_i - (\mu_0)_i| = O_P(1) \text{ for all } i = 1, \dots, p \tag{4.3}$$

$$\sqrt{n}|\hat{\Sigma}_{ij} - (\Sigma_0)_{ij}| = O_P(1) \text{ for all } i, j = 1, \dots, p \tag{4.4}$$

Another alternative is Tyler's M-estimate of Multivariate scatter given in [50], which also gives a $\sqrt{n}$ consistent estimate of Σ.

4.1.2. *High-dimensional case.*

(a) Sample mean, thresholded mean or LASSO estimator can be used to estimate μ

(b) Ledoit-Wolf estimator of covariance and Precision matrix [37], which gives distribution-free consistent estimators of covariance and precision matrix in high-dimensions as $p, n \to \infty$ or graphical lasso estimators [21] can be used to estimate covariance, Σ and precision matrix, Ω.

4.2. *Estimation of Density Generator Using Estimates $\hat{\mu}_n$ and $\hat{\Sigma}_n$.* The difference between the approach in Section 3 and this one is that $Y_i = (X_i - \mu)^T \Sigma^{-1} (X_i - \mu)$ is replaced by $\hat{Y}_i = (X_i - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1} (X_i - \hat{\mu}_n)$. But, if we use $\hat{Y}_i$ instead of Y_i in finding the estimate of $g_p(y)$, then, we shall show that we have a new rate of convergence depending on behavior of $||\Omega - \hat{\Omega}||$ and $||\mu - \hat{\mu}||$.

Lemma 10. *Under conditions of Lemma 9 and ϕ' being bounded and*

$$\begin{aligned} \mathbb{P}[||\mu - \hat{\mu}||_2 > t] &\leq J_1(t, n, p) \\ \mathbb{P}[||\Omega - \hat{\Omega}||_F > t] &\leq J_2(t, n, p). \end{aligned}$$

Then as $n \to \infty$, for some constants $c, d > 0$

$$\mathbb{P}\left[\left|\hat{\phi}_n(\hat{Y}_i) - \phi(Y_i)\right| > t\right] \leq c \exp(-dt) J_1(t, p, n) J_2(t, p, n) \tag{4.5}$$

Proof. Let us consider $\hat{\phi}_{(\hat{\mu}, \hat{\Omega})}$ as the estimate of log-density generator using $(\hat{Y}_i)_{i=1}^n$ as the data and $\hat{\phi}_{(\mu, \Omega)}$ as the estimate of log-density generator using $(Y_i)_{i=1}^n$ as the data. By applying Lemma 7 and 9, we get, for some constants, $k_1, k_2, k_3, k_4 > 0$,

$$\begin{aligned} \mathbb{P}\left[\left|\hat{\phi}_{(\hat{\mu}, \hat{\Omega})}(\hat{Y}_i) - \phi_{(\hat{\mu}, \hat{\Omega})}(\hat{Y}_i)\right| \geq t\right] &\leq k_1 \exp(-k_2 t) \\ \mathbb{P}\left[\left|\hat{\phi}_{(\mu, \Omega)}(Y_i) - \phi_{(\mu, \Omega)}(Y_i)\right| \geq t\right] &\leq k_3 \exp(-c_4 t) \end{aligned}$$

Now, $\hat{\phi}_n(\hat{Y}_i) \equiv \hat{\phi}_{(\hat{\mu},\hat{\Omega})}$, we want to prove, that for some constants $c, d > 0$,

$$\mathbb{P}\left[\left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi(Y_i)\right| > t\right] \leq c\exp(-dt)$$

Now,

$$\begin{aligned}
\left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right| \quad &\leq \quad \left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i)\right| \\
&\quad + \left|\hat{\phi}_{(\mu,\Omega)}(Y_i) - \phi_{(\mu,\Omega)}(Y_i)\right| \\
&\quad + \left|\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right|
\end{aligned}$$

Since, we already have bounds for first and second term, we only have to bound the third term. Now,

$$\begin{aligned}
\left(\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right) &= \phi\left((X_i - \hat{\mu})^T\hat{\Omega}(X_i - \hat{\mu})\right) - \phi\left((X_i - \mu)^T\Omega(X_i - \mu)\right) \\
&\leq |\phi'|\Big((X_i - \hat{\mu})^T\hat{\Omega}(X_i - \hat{\mu}) - (X_i - \mu)^T\Omega(X_i - \mu)\Big)
\end{aligned}$$

$$\begin{aligned}
(X_i - \mu)^T\Omega(X_i - \mu) &= (X_i - \hat{\mu} + \hat{\mu} - \mu)^T\Omega(X_i - \hat{\mu} + \hat{\mu} - \mu) \\
&= (X_i - \hat{\mu})^T\Omega(X_i - \hat{\mu}) + 2(X_i - \hat{\mu})^T\Omega(X_i - \hat{\mu}) + (\mu - \hat{\mu})^T\Omega(\mu - \hat{\mu}) \\
&= (X_i - \hat{\mu})^T\hat{\Omega}(X_i - \hat{\mu}) + (X_i - \hat{\mu})^T(\Omega - \hat{\Omega})(X_i - \hat{\mu}) \\
&\quad +2(X_i - \hat{\mu})^T\Omega(X_i - \hat{\mu}) + (\mu - \hat{\mu})^T\Omega(\mu - \hat{\mu})
\end{aligned}$$

So, from the assumptions on the estimators $\hat{\mu}$ and $\hat{\Omega}$ and if ϕ' is bounded, we get that for some constant $k_5, k_6 > 0$,

$$\mathbb{P}\left[\left|\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right| \geq t\right] \leq k_5\exp(-k_6 t)J_1(p,n,t)J_2(p,n,t)$$

and so the lemma follows. □

4.3. *Maximum Likelihood Estimation of μ and Ω.* From the Section 4.2, we have the data in the form $\hat{Y}_i \equiv (X_i - \hat{\mu})^T\hat{\Omega}(X_i - \hat{\mu})$, which we use to get an estimate of the density generator function g_p in a log-linear spline form like in Eq(3.14) in Section 3.1.1. So, the linear spline estimate of $\log g$ takes the form

$$\log\hat{g}_p(x) \quad = \quad -\sum_{i=1}^{n-1}(a_i x + b_i)\mathbf{1}\left((\hat{Y}_{(i)}, \hat{Y}_{(i+1)}]\right) - (a_{n+1}x + b_{n+1})\mathbf{1}\left((\hat{Y}_{(n)}, \hat{Y}_{(n+1)})\right) \tag{4.6}$$

where, $\hat{Y}_{(i)}$ is the i^{th} order statistic for $\{\hat{Y}_i\}_{i=0}^{n+1}$ with $\hat{Y}_0 \equiv -\infty$ and $\hat{Y}_{n+1} \equiv \infty$ and $(a_i, b_i)_{i=1}^n$ as define in Eq. (3.14).

Now, if we define $Y_i = (X_i - \mu)^T \Omega (X_i - \mu)$, then, the likelihood function of $\theta = (\mu, \Omega)$ given data $(X_1, \ldots, X_n)$ and density generator g_p becomes -

$$\mathcal{L}(\theta|\boldsymbol{Y}, g_p) = \prod_{i=1}^{n} |\Omega|^{1/2} g_p \left((X_i - \mu)^T \Omega (X_i - \mu) \right) = \prod_{i=1}^{n} |\Omega|^{1/2} g_p(Y_i) \tag{4.7}$$

and the log-likelihood function of $\theta = (\mu, \Omega)$ given data $(X_1, \ldots, X_n)$ and density generator g_p becomes -

$$\ell(\theta|\boldsymbol{Y}, g_p) = \frac{n}{2} \log|\Omega| + \sum_{i=1}^{n} \log g_p(Y_i)$$

Now, we can plug-in the a variant of estimate of $\log g_p$ from Eq (4.6), in the form

$$\log \tilde{g_p}(x) \quad = \quad - \sum_{i=1}^{n-1} (a_i x + b_i) \mathbf{1} \left((Y_{(i)}, Y_{(i+1)}] \right) - (a_{n+1} x + b_{n+1}) \mathbf{1} \left((Y_{(n)}, Y_{(n+1)}) \right)$$

where, $Y_{(i)}$ is the i^{th} order statistic for $\{Y_i\}_{i=0}^{n+1}$ with $Y_0 \equiv -\infty$ and $Y_{n+1} \equiv \infty$ and plug in $\log \tilde{g_p}$ in place of $\log g_p$, to get the approximated log-likelihood - Then, we can write

$$\begin{aligned} \ell(\theta|\boldsymbol{X}) &= \frac{n}{2} \log|\Omega| - \sum_{i=1}^{n} a_i \left((X_i - \mu)^T \Omega (X_i - \mu) \right) + Constant \\ &= \frac{n}{2} \log|\Omega| - \operatorname{tr}(S^* \Omega) + Constant \end{aligned}$$

where, $S^* = \sum_{i=1}^{n} a_i (X_i - \mu)(X_i - \mu)^T$. By maximizing the approximated log-likelihood $\tilde{\ell}(\theta)$ -

$$\tilde{\ell}(\theta|\boldsymbol{X}) = \frac{n}{2} \log|\Omega| - \operatorname{tr}(S^* \Omega) \tag{4.8}$$

we will get estimates of μ and Ω as

$$(\tilde{\mu}, \tilde{\Omega}) = \arg \max_{\mu, \Omega \succ 0} \tilde{\ell}(\mu, \Omega) \tag{4.9}$$

which we call **robust regularized estimators** of Euclidean parameters. See beginning of Section 2 for the notation.

4.3.1. *Penalized ML Estimation of μ and Ω: High-dimensional Case.* Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values s indicates *sparsity.* In the high-dimensional case, we have the dimension of the Euclidean parameters, p, growing with number of samples, n. We consider the penalized approximated log-likelihood function under assumption of sparsity to be

$$\tilde{\ell}(\mu|\boldsymbol{X}) = -\sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) \tag{4.10}$$

$$\tilde{\ell}(\Omega|\boldsymbol{X}, \tilde{\mu}) = \frac{n}{2} \log|\Omega| - \mathrm{tr}\,(S^*\Omega) + \nu|\Omega^-|_1 \tag{4.11}$$

where, $S^* \equiv \sum_{i=1}^{n} a_i (X_i - \tilde{\mu})^T (X_i - \tilde{\mu})$ and

$$\tilde{\mu} = \arg\max_{\mu} \tilde{\ell}(\mu) \tag{4.12}$$

$$\tilde{\Omega} = \arg\max_{\Omega \succ 0} \tilde{\ell}(\Omega) \tag{4.13}$$

are the **robust regularized estimators** of Euclidean parameters. See Section 2.1 for the notation.

Note that, if we had known Ω or if the proper form of elliptic density was known, we could have used penalization in the mean parameter too. So, if Ω is known, then, the penalized likelihood for μ becomes -

$$\tilde{\ell}(\mu|\boldsymbol{X}) = \sum_{i=1}^{n} -a_i (X_i - \mu)^T (X_i - \mu) + \nu_1 ||\mu||_1 \tag{4.14}$$

$$\tag{4.15}$$

and the penalized likelihood estimate, $\tilde{\mu}$ is

$$\tilde{\mu} = \arg\max_{\mu} \tilde{\ell}(\mu) \tag{4.16}$$

5. Inference III: Combined Approach and Theory.

Now, we can summarize the estimation procedure based on the steps suggested in Section 3 and 4. We shall provide the estimation procedure in this section and we shall provide the theoretical justification for the method.

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p\left((x - \mu_0)^T \Omega (x - \mu_0)\right) \tag{5.1}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x)dx = 1$$

$\mathbb{R}^+ = [0, \infty)$.

5.1. *Fixed Dimension Case.* We first consider the fixed dimensional case, that is when we do not have the dimension of the Euclidean parameters, p, not growing with number of samples, n. The estimation steps are as follows -

(1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\begin{aligned} \mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] &\leq J_1(t, n, p) \\ \mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] &\leq J_2(t, n, p). \end{aligned}$$

So, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(n))$, where, $\omega(n) \to 0$ as $n \to \infty$ with given tail bounds.
There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [25] is a good source

(2) Define $\hat{Y}_i = (X_i - \hat{\mu})^T \hat{\Omega}(X_i - \hat{\mu})$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (3.5). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (3.14).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ, given by equation (4.8)

$$\tilde{\ell}(\theta|\boldsymbol{X}) = \frac{n}{2} \log|\Omega| - \mathrm{tr}\left(S^* \Omega\right).$$

where, $S^* = \sum_{i=1}^n a_i (X_i - \mu)(X_i - \mu)^T$. We maximize $\tilde{\ell}(\theta)$ with respect to θ, to get the robust estimates of θ_0.

(4) *(Optional)* We can use the estimates to obtained in Step 4 and repeat Steps 1-3, to get an estimate of g_p. But, both in theory and practice, that does not improve the error rates of the new estimate of g_p.

5.2. *High-dimensional Case.* Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0= s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small value of s indicates *sparsity*. We first consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n. The estimation steps are as follows -

(1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates around zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\begin{aligned} \mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] &\leq J_1(t, n, p) \\ \mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] &\leq J_2(t, n, p). \end{aligned}$$

So, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(p, n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.
There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [25] is a good source

(2) Define $\hat{Y}_i = (X_i - \hat{\mu})^T \hat{\Omega}(X_i - \hat{\mu})$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (3.5). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (3.14).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (4.10)

$$\begin{aligned} \tilde{\ell}(\mu|\boldsymbol{X}) &= \sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) \\ \tilde{\ell}(\Omega|\boldsymbol{X}, \tilde{\mu}) &= \frac{n}{2} \log|\Omega| - \mathrm{tr}\,(S^*\Omega) + \nu|\Omega^-|_1 \end{aligned}$$

where, $S^* \equiv \sum_{i=1}^{n} a_i (X_i - \tilde{\mu})^T (X_i - \tilde{\mu})$ and

$$\begin{aligned} \tilde{\mu} &= \arg\max_{\mu} \tilde{\ell}(\mu) \\ \tilde{\Omega} &= \arg\max_{\Omega} \tilde{\ell}(\Omega) \end{aligned}$$

are the *robust regularized estimators* of Euclidean parameters of θ_0.

Note that, if we had known Ω or if the proper form of elliptic density was known, we could have used penalization in the mean parameter too. So, if Ω is known, then, the penalized likelihood for μ becomes -

$$\tilde{\ell}(\mu|\boldsymbol{X}) = -\sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) + \nu_1 ||\mu||_1$$

and the penalized likelihood estimate, $\tilde{\mu}$ is

$$\tilde{\mu} \quad = \quad \arg\max_{\mu} \tilde{\ell}(\mu)$$

The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. One way of solving the optimization problem for μ is by using LARS algorithm of [17]. The optimization problem for Ω can be solved by using the graphical LASSO algorithms provided in [21], [54] and [28].

(4) *(Optional)* We can use the estimates to obtained in Step 4 and repeat Steps 1-3, to get an estimate of g_p. But, both in theory and practice, that does not improve the error rates of the new estimate of g_p.

In Section 5.3, we give proof of Theorem 6, by which we show that regularized estimators of the Euclidean parameters θ in the high-dimensional case is also robust to tail behavior of the underlying elliptical distribution.

5.3. *Theory.* We have described the estimation procedure of the Euclidean parameters and the non-parametric component of the elliptical density in Section 5.2. We now try to show that the estimators have nice behavior in the case of fixed and high dimension. The main theorem in this section is Therem 5 and Theorem 6given in Section 2.2. However, to prove the Theorems we first need to discuss the setup and the conditions.

We have $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p \left((x - \mu_0)^T \Omega (x - \mu_0) \right) \tag{5.2}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter. We shall also consider that g_p possess **consistency** property.

Now, according to the *consistency* condition of Elliptical distribution mentioned in Lemma 4, for independent random variables $(\xi_1, \ldots, \xi_n)$ with $\xi_i \overset{d}{=} \xi$, $\forall i$,

$$(\sqrt{\xi_1}\Omega_0^{1/2}(X_1 - \mu_0), \ldots, \sqrt{\xi_n}\Omega_0^{1/2}(X_n - \mu_0)) \overset{d}{=} (Z_1, \ldots, Z_n)$$

where, $(Z_1, \ldots, Z_n)$ are independent p-variate standard Gaussian random variables. If we define, $W_i \equiv \Omega_0^{1/2}(X_i - \mu_0)$, then,

$$(\sqrt{\xi_1}W_1, \ldots, \sqrt{\xi_n}W_n) \overset{d}{=} (Z_1, \ldots, Z_n) \tag{5.3}$$

Now, according to the estimation procedure we have proposed, after the estimation of the density generator by a log-lear spline, according to log-likelihood equation Eq (4.7), the resulting log-likelihood for θ becomes of the form

$$\ell(\theta|\boldsymbol{X}) = \frac{n}{2}\log|\Omega| - \sum_{i=1}^{n} a_i \left((X_i - \mu)^T \Omega (X_i - \mu)\right) + Constant$$

which is like the log-likelihood if estimated density $\hat{f}$ with parameters (μ_0, Ω_0) which has the form -

$$\hat{f}(X_1, \ldots, X_n|\theta_0) = C|\Omega|^{1/2}\exp\left(-\sum_{i=1}^{n} a_i (X_i - \mu_0)^T \Omega_0 (X_i - \mu_0)\right).$$

that means that as if the data $W_i \equiv \Omega_0^{1/2}(X_i - \mu_0)$ has the following distributional form -

$$(\sqrt{a_1}W_1, \ldots, \sqrt{a_n}W_n) \overset{d}{=} (Z_1, \ldots, Z_n)$$

So, we can see that by our estimation of the non-parametric component, we have got an estimate of the latent scale variable ξ inherent to the consistent elliptical distribution. Our results on rate will thus depend on the tail behavior of ξ.

We wish to prove the Theorem 6 and Theorem 5 now. Recall the assumptions (A1)-(A5) given in Section 2.2, which preceded Theorem 6 and Theorem 5.

Before proving Theorem 6, we shall state and prove two lemma on concentration inequalities which are vital for the proof of Theorem 6. Lemma 11 is variant of the Lemma B.1 of [7]. Lemma 12 is variant of the Lemma 3 of [5] or Lemma 1 of [45].

Concentration inequality around the mean parameter μ_0 will be goal of our first Lemma.

LEMMA 11. *Let X_i be i.i.d elliptically distributed random variables having elliptic distribution with parameters (μ_0, Ω_0) and a_i's are as stated in Eq. (5.3) and Ω be an estimate of Ω_0 satisfying Assumption (A2). Then,*

$$\mathbb{P}\left[\max_j \sum_{i=1}^n \sqrt{a_i}|(\Omega_0^{1/2}(X_i - (\mu_0)))_j|> 2nt\right] \leq p(c\exp(-dt)J_1(t,n,p)J_2(t,n,p))^n \tag{5.4}$$

where, c, d are some constants.

So, the rate of convergence is controlled by $\sigma_1(p,n)$, which is the function such that the above inequality satisfies if we replace $t = O(\sigma_1(p,n))$

PROOF. We have $(\sqrt{\xi_i}\Omega_0^{1/2}(X_i - \mu_0) \overset{d}{=} Z_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$. So, we have, by Gaussian tail inequality

$$\mathbb{P}\left[\xi_i(X_i - (\mu_0))^T\Omega_0(X_i - \mu_0) > t^2\right] \quad \leq \quad c_1\exp(-c_2t^2)$$

From Lemma 9 and Lemma 10, we have that,

$$\mathbb{P}\left[(a_i - \xi_i)(X_i - (\mu_0))^T\Omega_0(X_i - \mu_0) > t^2\right] \quad \leq \quad (k_1\exp(-k_2t^2)J_1(t^2,n,p)J_2(t^2,n,p))$$

Combining the above two equations, we have for some constants $c_3, c_4 > 0$,

$$\begin{aligned}\mathbb{P}\left[a_i(X_i - (\mu_0))^T\Omega_0(X_i - \mu_0) > t^2\right] \quad &\leq \quad c_3\exp(-c_4t^2)J_1(t^2,n,p)J_2(t^2,n,p) \\ \Rightarrow \mathbb{P}\left[\sqrt{a_i}|(\Omega_0^{1/2}(X_i - (\mu_0)))_j|> t\right] \quad &\leq \quad c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p)\end{aligned}$$

So, we get that,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n \sqrt{a_i}|(\Omega_0^{1/2}(X_i - (\mu_0)))_j|> nt\right] \quad &\leq \quad (c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p))^n \\ \mathbb{P}\left[\max_j \sum_{i=1}^n \sqrt{a_i}|(\Omega_0^{1/2}(X_i - (\mu_0)))_j|> 2nt\right] \quad &\leq \quad p(c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p))^n\end{aligned}$$

Let us consider that $\sigma_1(p,n)$ to be the function such that the above inequality satisfies if we replace $t = O(\sigma_1(p,n))$. □

So, we can see that depending on the behavior of $H_1(t)$, either $H_1(t)$ or $\exp(-c_2t^2)$ controls the rate in the above Lemma.

Concentration inequality around the covariance matrix parameter Σ_0 will be goal of our next Lemma.

LEMMA 12. *Let X_i be i.i.d elliptically distributed random variables having elliptic distribution with parameters (μ_0, Σ_p) and a_i's are as stated in Eq.* (5.3) *and $\Omega_p \equiv \Sigma_p^{-1}$. We also have, $\lambda_{\max}(\sigma_p) \leq \overline{k} < \infty$. Then, if $(\Sigma_p)_{ab} = \sigma_{ab}$,*

$$\mathbb{P}\left[\max_{j\neq k}\sum_{i=1}^{n}|a_i\hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}|> nt\right] \\ \leq d_1p^2(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(\exp(-nd_3t^2)) \text{ for } |t|\leq \delta \tag{5.5}$$

where, $\hat{\mu}$ is an of μ_0, d_1, d_2, d_3 and δ depend on $\overline{k}$ only.

So, the rate of convergence is controlled by $\sigma_2(p, n)$, which is the function such that the above inequality satisfies if we replace $t = O(\sigma_2(p, n))$.

PROOF. Consider $W_i = X_i - \mu_0$ and $\hat{W}_i = X_i - \hat{\mu}$ for $i = 1, \ldots, n$. We have $\Omega_0^{1/2}\sqrt{\xi_i}W_i \overset{d}{=} Z_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$. So, we have, by Lemma 3 of [5], for $|t|\leq \delta$, for some constants $c_1, c_2 > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n}|\xi_iW_{ij}W_{ik} - (\Sigma_0)_{jk}|> nt\right] \leq c_1\exp(-nc_2t^2) \tag{5.6}$$

We have $a_i > 0$ and $\xi_i > 0$. Now,

$$|a_iW_{ij}W_{ik} - \xi_iW_{ij}W_{ik}|\leq |a_i||W_i||_2^2-\xi_i||W_i||_2^2|.\frac{|W_{ij}W_{ik}|}{\sum_j W_{ij}^2} \leq |a_i||W_i||_2^2-\xi_i||W_i||_2^2|$$

Since, a_i from $\hat{\phi}$ is the slope estimate of the log density generator ϕ, whose slope is ξ, conditional on ξ. So, we have from Lemma 9 and Lemma 10 for $c_3, c_4 > 0$, that,

$$\mathbb{P}\left[|a_i||W_i||_2^2-\xi_i||W_i||_2^2|> t\right] \leq c_3\exp(c_4t)J_1(t, n, p)J_2(t, n, p)$$

which implies,

$$\mathbb{P}\left[|a_iW_{ij}W_{ik} - \xi_iW_{ij}W_{ik}|> t\right] \leq c_3\exp(c_4t)J_1(t)J_2(t) \tag{5.7}$$

Now,

$$|a_i\hat{W}_{ij}\hat{W}_{ik} - a_iW_{ij}W_{ik}| \leq a_i\left|-X_{ij}(\hat{\mu}_k - (\mu_0)_k) - X_{ik}(\hat{\mu}_j - (\mu_0)_j) + (\hat{\mu}_j\hat{\mu}_k - (\mu_0)_j(\mu_0)_k)\right|$$

So, we have, for some constants, $c_5, c_6 > 0$,

$$\mathbb{P}\left[|a_i\hat{W}_{ij}\hat{W}_{ik} - a_iW_{ij}W_{ik}|> t\right] \leq H(t)c_5\exp(c_6t) \tag{5.8}$$

So, combining the above equations (5.6) (5.7) and (5.8), for constants $d_1, d_2, d_3 > 0$, we get that,

$$\mathbb{P}\left[\sum_{i=1}^{n} |a_i \hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}| > 3nt\right] \leq d_1(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(d_3\exp(-nd_3t^2))$$

$$\mathbb{P}\left[\max_{j\neq k}\sum_{i=1}^{n} |a_i \hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}| > 3nt\right] \\ \leq p(p-1)d_1(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(\exp(-nd_3t^2))$$

Let us consider that $\sigma_2(p,n)$ to be the function such that the above inequality satisfies if we replace $t = O(\sigma_2(p,n))$. □

So, we can see that depending on the behavior of $H(t), J_1(t), J_2(t)$ and these rates along with $\exp(-c_2t^2)$ controls the rate in the above Lemma.

5.3.1. *Proof of Theorem 6.*

(a) We consider

$$Q(\mu) = \sum_{i=1}^{n} a_i(X_i-\mu)^T\Omega_0(X_i-\mu) - \sum_{i=1}^{n} a_i(X_i-\mu_0)^T\Omega_0(X_i-\mu_0)$$

Our estimate $\tilde{\mu}$ given in Eq. (4.12) minimizes $Q(\mu)$ or equivalently $\hat{\delta} = \tilde{\mu} - \mu_0$ minimizes $G(\delta) \equiv Q(\mu_0+\delta)$ given an estimate Ω of Ω_0. Consider the set

$$\Theta_n(M) = \{\delta : ||\delta||_2 = Mr_n\}$$

where,

$$r_n = \sqrt{p}O(\sigma_1(p,n)) \to 0$$

where, $\sigma_1(p,n)$ is taken from statement of Lemma 11. Note that $G(\delta)$ is a convex function and

$$G(\hat{\delta}) \leq G(0) = 0$$

Then, if we can show that

$$\inf\{G(\delta) : \delta \in \Theta(M)\} > 0$$

then the minimizer $\hat{\delta}$ must be inside the sphere $\Theta_n(M)$ and hence

$$||\hat{\delta}||_2 \leq Mr_n$$

Now,

$$\begin{aligned}
&\sum_{i=1}^{n} a_i(X_i-\mu)^T\Omega_0(X_i-\mu) \\
&\quad = \sum_{i=1}^{n} a_i(X_i-\mu_0)^T\Omega_0(X_i-\mu_0) + 2\sum_{i=1}^{n} a_i(X_i-\mu_0)^T\Omega_0(\mu_0-\mu) \\
&\qquad\qquad + \sum_{i=1}^{n} a_i(\mu_0-\mu)^T\Omega_0(\mu_0-\mu)
\end{aligned}$$

So,

$$G(\delta) = 2\sum_{i=1}^{n} a_i(X_i-\mu_0)^T\Omega_0\delta + \sum_{i=1}^{n} a_i\delta^T\Omega_0\delta$$

Now, by applying Cauchy-Scwartz inequality and Lemma 11, and $\bar{a}_1 = \sum_{i=1}^n \sqrt{a_i}$ and $\bar{a}_2 = \sum_{i=1}^n a_i$, we get that,

$$\begin{aligned}
G(\delta) &\geq -2\sum_{i=1}^{n} \sqrt{a_i}\sigma_1(p,n)||\Omega_0^{1/2}\delta||_2 + \sum_{i=1}^{n} a_i\delta^T\Omega_0\delta \\
&\geq -2\bar{a}_1 M\sqrt{s}(\sigma_1(p,n))^2 + \bar{a}_2\underline{k}M^2 s(\sigma_1(p,n))^2 \\
&\geq \bar{a}_2(1/(\bar{k}+o_P(1)))M^2 s(\sigma_1(p,n))^2 - 2\bar{a}_1 M\sqrt{s}(\sigma_1(p,n))^2 \\
&> 0
\end{aligned}$$

for large enough $M > 0$. So, for large enough $M > 0$,

$$G(\delta) > 0$$

So, our proof follows.

(b) The proof closely follows proof of Theorem 1 in [45]. We do not repeat the proof as essentially the same proof follows. The only difference are
   (i) Take S^* in stead of $\hat{\Sigma}$ in the whole proof.
   (ii) Take $r_n = \sqrt{p+s}O(\sigma_2(p,n)) \to 0$, where,$\sigma_2(p,n)$ is taken from statement of Lemma 12
   (iii) Use Lemma 12 instead of Lemma 1 after the equations (12) and (13) of the proof.
   (iv) Use regularization parameter $\nu_2 = \frac{C_1}{\epsilon}O(\sigma_2(p,n))$, where, $\sigma_2(p,n)$ is taken from statement of Lemma 12

(c) Follows fro Lemma 7.

5.3.2. *Proof of Theorem 5.* Proof of Theorem 5 becomes a special case of proof of Theorem 6 as we do not have dependence on p in rate anymore.

6. Application to Special Problems.

6.1. *Application to Covariance and Precision Matrix Estimation.* The general method of estimation given in Section 5 can be used in *robust regularized* estimation of covariance and inverse covariance matrices. Class of Elliptical densities contain densities having tails both thicker and thinner than sub-Gaussian random variables. So, adaptive estimation of covariance matrix from a class of elliptical densities lead to covariance matrix estimators, which are robust to tail-behavior of the under distribution of the random variable.

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mathbf{0}, \Omega)$. Then the density function $f(\cdot; \mu, \Omega)$ is of the form

$$f(x; \mathbf{0}, \Omega) = |\Omega|^{1/2} g_p \left(x^T \Omega x \right) \tag{6.1}$$

where $\theta = (\mathbf{0}, \Omega) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and covariance parameters respectively with $\Omega \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x) dx = 1$$

$\mathbb{R}^+ = [0, \infty)$. $\Omega \equiv \Sigma^{-1}$ is the inverse covariance parameter and Σ is the covariance parameter.

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity.* We consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

6.1.1. *Method.* If we follow the estimation procedure suggested in Section 5, we shell get the **robust regularized** estimate, $\tilde{\Omega}$ of Ω with nice theoretical properties given in Theorem 6. This procedure can be performed for any other additional structure on the parameters and for any other form of penalization on parameters. As a special case, assume *sparsity condition*: $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form and ℓ_1 penalty on off-diagonal elements of Ω. The steps of estimation procedure can be stated as -

(1) Assume that we have an initial consistent estimators $\hat{\Omega}$, such that, $||\hat{\Omega} - \Omega_0||_F$ concentrates around zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\Omega - \hat{\Omega}||_F > t] \quad \leq \quad J_2(t, n, p).$$

So, we have, $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.
There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [25] is a good source.

(2) Define $\hat{Y}_i = X_i^T \hat{\Omega} X_i$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (3.5). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (3.14).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (4.10)

$$\tilde{\ell}(\Omega|\boldsymbol{X}, \tilde{\mu}) \quad = \quad \frac{n}{2} \log|\Omega| - \mathrm{tr}\,(S^* \Omega) + \nu |\Omega^-|_1$$

where, $S^* \equiv \sum_{i=1}^n a_i X_i^T X_i$ and

$$\tilde{\Omega} \quad = \quad \arg\max_{\Omega} \tilde{\ell}(\Omega)$$

are the *robust regularized estimators* of Euclidean parameter Ω.
The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. The optimization problem for Ω can be solved by using the graphical LASSO algorithms provided in [21], [54] and [28].

We can get *robust regularized* estimate of covariance matrix Σ by this method.

6.1.2. *Theoretical Performance.* We start with the following assumption -

(B1) Assume that we have an initial consistent estimators $\hat{\Omega}$, such that, $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] \quad \leq \quad J(t, n, p). \tag{6.2}$$

We have, $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega(p,n))$, where, $\omega(p,n) \to 0$ as $p, n \to \infty$ with given tail bounds.

In the high-dimensional situation we assume the additional structure of sparsity imposed on Ω_0. Density generator g comes from a consistent family elliptical distributions and so by Lemma 4, $\Omega_0^{1/2}(X) \overset{d}{=} Z_p/\sqrt{\xi}$. Let us consider the probability density function of X_j to be h $(j = 1, \ldots, p)$ and

$$H(u) \equiv \int_u^\infty h(v)dv \tag{6.3}$$

We consider the following assumptions -

(B2) Suppose that $||\Omega^-||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity.*

(B3) $\lambda_{\min}(\Sigma_0) \geq \underline{k} > 0$, or equivalently $\lambda_{\max}(\Omega_0) \leq 1/\underline{k}$. $\lambda_{\max}(\Sigma_0) \leq \overline{k}$.

(B4) The function $H(t)$ defined in Eq. (6.3) and $J(t,n,p)$ defined in Eq. (6.2), satisfies the following conditions -
there exists a function $\sigma(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $d_1, d_2, d_3 > 0$, for $t = O(\sigma(p,n))$,

(6.4)
$$d_1 p^2 (J(t,n,p))^n (\exp(-nd_2 t))(d_3 \exp(-nd_3 t^2)) \to 0 \text{ as } p, n \to \infty$$

Let us consider that we have obtained estimators of the Euclidean parameters $\tilde{\Omega}$ and the nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 5. We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow.

THEOREM 13. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (B1)-(B4),*

$$||\Omega_0 - \tilde{\Omega}||_F = O_P\left(\sqrt{(p+s)}\sigma(p,n)\right) \textit{ for } \nu = O\left(\sigma(p,n)\right).$$

PROOF. The proof follows from Theorem 6 with $\mu_0 = 0$. □

So, we get a consistent estimator $\tilde{\Omega}$ with computable rates of convergence, which is robust against tail behavior. Similarly, we can also get estimates of the covariance matrix Σ and correlation matrix by extending the methods suggested in [45] and [36] in our setup.

6.2. *Application to Regression with Elliptical Errors.* The general method of estimation given in Section 5 can be used in *robust regularized* regression. Class of Elliptical densities contain densities having tails both thicker and thinner than sub-Gaussian random variables. So, if we have linear regression with error variables having elliptical distributions, then, adaptive estimation from a class of elliptical densities lead to regression estimators, which are robust to tail-behavior of the error variables. Thus, we shall be able to get **robust regularized** regression estimators.

Let $((Y_1, X_1), \ldots, (Y_n, X_n))$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ are predictor and response variable such that

$$Y_i = \beta_0^T X_i + \epsilon_i$$

where, ϵ_i are independent elliptically distributed random variables with density $f(\cdot; \mathbf{0}, \mathbf{I})$. So, the density function of Y_i is $f(\cdot; \beta_0^T X_i, \mathbf{I})$, where, f is the elliptical density. So, the problem of estimation of β boils down to the problem of estimation of mean parameter of the elliptical density.

6.2.1. *Method.* For high-dimensional regression, we consider the most vanilla situation and method. We consider the high-dimensional situation under the additional structure of sparsity imposed on the regression coefficients β_0. We consider that $||\beta||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity.* We consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

(1) Assume that we have an initial consistent estimators $\hat{\beta}$, such that, $||\hat{\beta} - \beta_0||_2$ concentrates around zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\beta_0 - \hat{\beta}||_2 > t] \quad \leq \quad J(t, n, p)$$

So, we have, $||\hat{\beta} - \beta_0||_F = O_P((\omega(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.
There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [25] is a good source.

(2) Define $\hat{E}_i = (Y_i - \hat{\beta}^T X_i)^2$ and based on $(\hat{E}_1, \ldots, \hat{E}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (3.5). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (3.14).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (4.10)

$$\tilde{\ell}(\mu|\boldsymbol{X}) \quad = \quad \sum_{i=1}^{n} a_i(Y_i - \beta^T X_i)^T(Y_i - \beta^T X_i) + \nu||\beta||_1$$

and

$$\tilde{\beta} \quad = \quad \arg\max_{\beta} \tilde{\ell}(\beta)$$

are the *robust regularized estimators* of Euclidean parameters of β_0. The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. One way of solving the optimization problem for β is by using LARS algorithm of [17].

6.2.2. *Theoretical Performance.* Thus following the estimation procedure suggested above, we shell get the **robust regularized** estimate, $\tilde{\beta}$ of β_0 with nice theoretical properties under restrictions on design matrix and coefficient parameters such as given in [7], [55] and [42]. Let us just give one such example of conditions on design matrix and coefficient parameters called *Restricted Eigenvalues* conditions given in [7]. The condition is stated as -

(C1) Assume

$$\kappa(s,p,c_0) \equiv \min_{J_0 \subseteq [p], |J_0| \leq s} \; \min_{\delta \neq 0, |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_{01}}|_1} \frac{||X\delta_2||_2}{\sqrt{n}||\delta_{J_{01}}||_2} > 0 \tag{6.5}$$

where, for integers s, m such that $1 \leq s \leq p/2$ and $m \geq s, s+m \leq p$, a vector $\delta \in \mathbb{R}^p$ and a set of indices $J_0 \subseteq \{1,\ldots,p\}$ with $|J_0| \leq s$; denote by J_1 the subset of $\{1,\ldots,p\}$ corresponding to the m largest in absolute value coordinates of δ outside of J_0, and define $J_{01} \equiv J_0 \cup J_1$.

Also,

(C2) Assume that we have an initial consistent estimators $\hat{\beta}$, such that, $||\hat{\beta} - \beta_0||_2$ concentrates to zero with tail bounds given by functions $J(t,n,p)$ such that,

$$\mathbb{P}[||\beta_0 - \hat{\beta}||_2 > t] \quad \leq \quad J(t,n,p) \tag{6.6}$$

We have, $||\hat{\beta} - \beta_0||_F = O_P((\omega(p,n))$, where, $\omega(p,n) \to 0$ as $p, n \to \infty$ with given tail bounds.

We consider the high-dimensional situation. So, in this case, the dimension of Euclidean parameters grows with n. In the high-dimensional situation we assume the additional structure of sparsity imposed on the coefficient parameters β_0. Density generator g comes from a consistent family elliptical distributions and so by Lemma 4, $\epsilon_i \overset{d}{=} Z/\sqrt{\xi}$. Let us consider the probability density function of ϵ_i to be h $(j = 1, \ldots, p)$ and

$$H(u) \quad \equiv \quad \int_u^\infty h(v)dv \tag{6.7}$$

We consider the following assumptions -

(C3) Suppose that $||\beta||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*.

(C4) The function $H(t)$ defined in Eq. (6.7) and $J(t, n, p)$ defined in Eq. (6.2), satisfies the following conditions -
there exists a function $\sigma(p, n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $c, d > 0$, for $t = O(\sigma(p, n))$,

$$p(c\exp(-dt)J(t, n, p))^n \to 0 \text{ as } p, n \to \infty \tag{6.8}$$

Let us consider that we have obtained estimators of the coefficient parameters $\tilde{\beta}$ and nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 5. We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow.

Theorem 14. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (C1)-(C4), with $c_0 = 3$ in (C1),*

$$||\beta_0 - \tilde{\beta}||_2 = O_P\left(\sqrt{s}\sigma(p, n)\right) \tag{6.9}$$

Proof. The proof follows the same steps as proof of Theorem 7.2 of [7] with only the concentration inequality portion replaced by the concentration inequality of Theorem 6 and Lemma 11. □

This procedure can be performed for any other additional structure on the regression parameters and for any other form of penalization of loss function. These gives a lot of scope for future work.

6.3. *Mixture of Elliptic Distributions.* Let $X_1, \dots, X_n \sim \sum_{k=1}^{K} p_k f_k(x; \mu_k, \Omega_k)$, where, $X_i \in \mathbb{R}^p$ and $f_k(\cdot; \mu_k, \Omega_k)$ is the density of elliptic distribution of the form

$$f(x; \mu_k, \Omega_k) = |\Omega_k| g_k \left((x - \mu_k)^T \Omega_k (x - \mu_k)\right) \tag{6.10}$$

where, $g_k(\cdot)$ is a density generator, that is, a non-negative function on $[0, \infty)$ such that the spherically symmetric (around zero) function $g_k(x^T x)$, $x \in \mathbb{R}^p$ integrates to 1 and g_k is non-increasing in $[0, \infty)$ so that the density is unimodal.

Our goal is the estimation of Euclidean parameters $\boldsymbol{\theta} = (\mu_k, \Omega_k)_{k=1}^K$ or $\boldsymbol{\theta} = (\mu_k, \Omega_k)_{k=1}^K$ in the high-dimensional setting as well as the infinite-dimensional parameters $\boldsymbol{G} = (g_1, \dots, g_K)$

6.3.1. *EM Algorithm.* We have data $\boldsymbol{X} = (X_1, \dots, X_n)$, where, $X_i \overset{i.i.d}{\sim} p(x; \boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{\pi})$. The complete data vector is given by $\boldsymbol{X_c} = (\boldsymbol{Z}, \boldsymbol{X})$, where, $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ and each $Z_i \in \{0, 1\}^K$ indicates component label. The complete data log likelihood is given by

$$\begin{aligned} \ell_c(\boldsymbol{\phi}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k + \sum_{k=1}^{K} \lambda_{1k} ||\mu_k||_1 + \sum_{k=1}^{K} \lambda_{2k} ||\Omega^-||_1 \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \left(\frac{1}{2} \log |\Omega_k| + \log g \left(tr\left((x - \mu_k)^T (x - \mu_k)\right) \Omega_k\right)\right) \end{aligned}$$

The conditional log likelihood is given by

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\phi}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \pi_k + \sum_{k=1}^{K} \lambda_{1k} ||\mu_k||_1 + \sum_{k=1}^{K} \lambda_{2k} ||\Omega^-||_1 \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \left(\frac{1}{2} \log |\Omega_k| + \log g \left(tr\left((x - \mu_k)^T (x - \mu_k)\right) \Omega_k\right)\right) \end{aligned}$$

where, $\tau_{ik} = \mathbb{E}[Z_{ik} | \boldsymbol{X}, \boldsymbol{\phi}]$

For m^{th} iteration of the algorithm, we start with $\boldsymbol{\phi}^{(m)}$ and
E step We estimate $\tau_{ik}^{(m+1)} = \mathbb{E}[Z_{ik} | \boldsymbol{X}, \boldsymbol{\phi}^{(m)}]$.

M step We maximize $\mathcal{Q}(\boldsymbol{\phi})$ with $\boldsymbol{\tau}^{(m+1)}$ and the m^{th} iterate estimates are

$$\begin{aligned}
\pi_k^{(m+1)} &= \frac{\sum_{i=1}^n \tau_{ik}^{(m+1)}}{\sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(m+1)}} \\
\mu_k^{(m+1)} &= \arg\min_{\mu} \sum_{i=1}^n \tau_{ik} \left(\log g_k^{(m)} \left(tr\left((X_i-\mu)(X_i-\mu)^T\right)\Omega_k^{(m)}\right)\right) + \lambda_{1k}||\mu||_1 \\
\Omega_k^{(m+1)} &= \arg\min_{\Omega \succ 0} \sum_{i=1}^n \tau_{ik} \left(\log g_k^{(m)} \left(tr\left((X_i-\mu)(X_i-\mu)^T\right)\Omega\right) - \frac{1}{2}\log|\Omega|\right) \\
&+ \lambda_{2k}||\Omega^-||_1 \\
g_k^{(m+1)} &= \text{log-linear spline estimate of } g_k \text{ based on} \\
& (X_i - \mu_k^{(m+1)})^T \Omega_k^{(m+1)} (X_i - \mu_k^{(m+1)}) \text{ for } i = 1, \ldots, n
\end{aligned}$$

6.3.2. *Theoretical Results.* The family of distributions $\boldsymbol{P} = \{P_{(\boldsymbol{\theta},\boldsymbol{G})}\}$ becomes identifiable under the conditions given in [27]. The condition states that for elliptical distributions with consistency property, that is, elliptical distributions of the form $X_p \stackrel{d}{=} \frac{Z_p}{\xi}$ given in Lemma 4, we have identifiability of Euclidean parameters for mixtures of such elliptic distributions if density of ξ, h exists and satisfies

$$\lim_{r \to 0} \frac{h(r)}{h(ar)} = 0 \text{ for } a > 1 \tag{6.11}$$

The result is given in Theorem 4 of [27].

THEOREM 15. *Assume that we have a mixture of elliptical distributions $\boldsymbol{P} = \{P_{(\boldsymbol{\theta},\boldsymbol{G})}\}$ with consistency property. Also, the scale parameter, ξ as defined in Lemma 4 of each elliptical distribution component satisfies 6.11. Also, the Euclidean parameters of each elliptical distribution component satisfies the conditions mentioned in Theorem 6 and they lie within a compact set. Then, we have -*
The EM algorithm converges to a stationary point or local maxima of the penalized likelihood function.

PROOF. We are maximizing the penalized likelihood at each **M Step** of the EM algorithm by following the steps of penalized maximum likelihood inference in Section 5. Thus, we get a hill-climbing algorithm. Also, the penalized likelihood function is bounded from above. So, the sequence of estimators obtained by EM iterations converges to a stationary point of the penalized likelihood function, since we are within a compact set. □

7. Simulation Examples.

We simulate from high-dimensional Gaussian and t-distribution and try to estimate the inverse covariance matrix.

7.1. *Estimation of High-dimensional Covariance Matrix and Density Generator.*

We have number of samples $n = 400$ and dimension of the data vectors as $p = 400$. We generate the Gaussian distribution with mean $\mathbf{0}$ and banded covariance matrix. The estimated covariance matrices are given in Figure 2. The top row of Figure 2 are the original covariance matrix, empirical covariance matrix and Graphical LASSO estimate from left to right. The bottom row of Figure 2 are the banded estimated covariance matrix, robust estimated covariance matrix and robust regularized estimated covariance matrix (our method) from left to right.

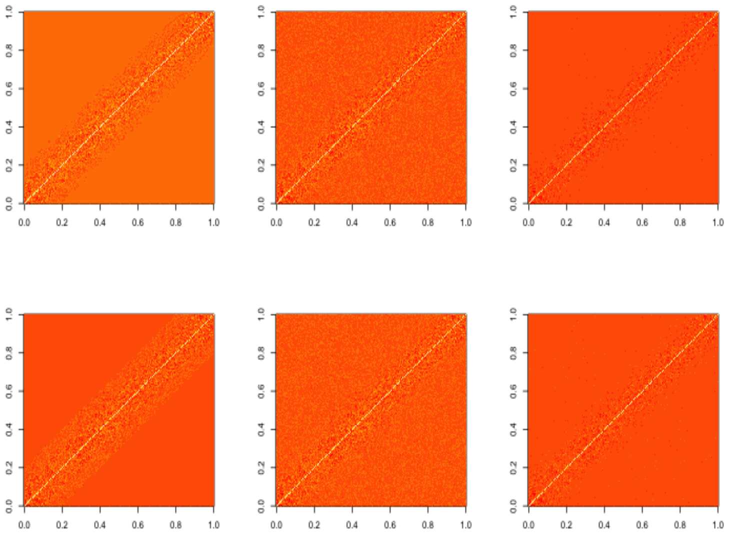


FIGURE 2. *For $n = 400$ and $p = 400$, we get different covariance estimators for a banded covariance matrix of normal distribution.*

We have number of samples $n = 400$ and dimension of the data vectors as $p = 400$. We generate the t distribution having 2 degree of freedom and a banded covariance matrix. The estimated covariance matrices are given in Figure 6. The top row of Figure 6 are the original covariance matrix, empirical covariance matrix and Graphical LASSO estimate from left to right. The bottom row of Figure 6 are the banded estimated covariance matrix, robust estimated covariance matrix and robust regularized estimated covariance matrix (our method) from left to right. We also give the estimated

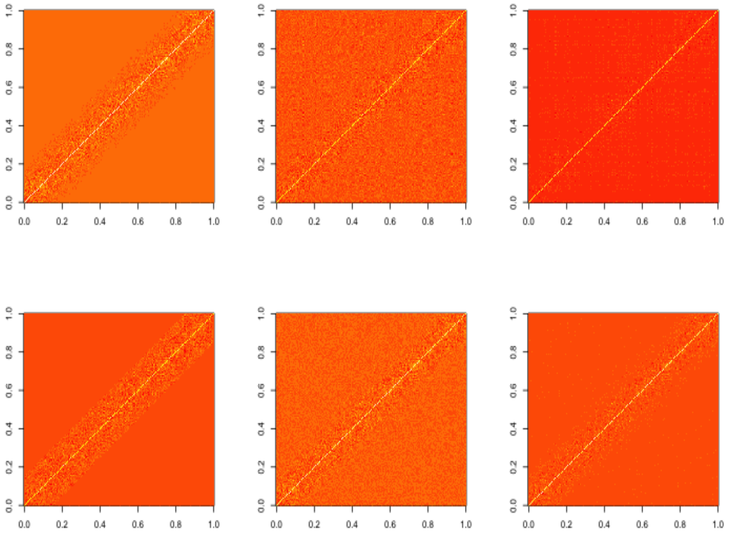


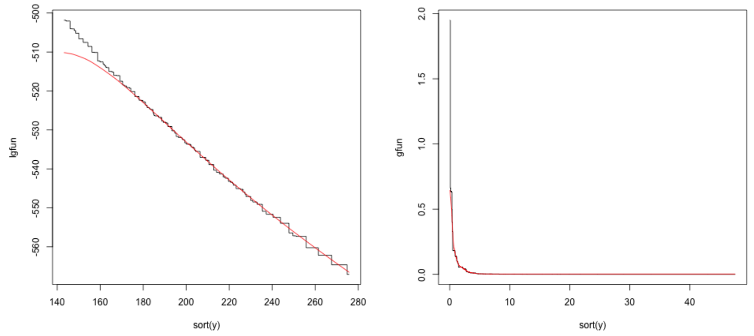
FIGURE 3. *For $n = 400$ and $p = 400$, we get different covariance estimators for a banded covariance matrix. of t-dist*

density generators for the Gaussian distribution and t distribution cases.

8. Real Data Examples.

8.1. *High-dimensional Covariance Estimation in Breast Cancer Data.* In the biological data set, we focus on selecting gene expression profiling as a potential tool to predict them breast cancer patients who may achieve pathologic Complete Response (pCR), which is defined as no evidence of viable, invasive tumor cells left in surgical specimen. pCR after neoadjuvant chemotherapy has been described as a strong indicator of survival, justifying its use as a surrogate marker of chemosensitivity. Consequently, considerable interest has been developed in finding methods to predict which patients will have a pCR to preoperative therapy. In this study, we use the normalized gene expression data of 130 patients with stage I-III breast cancers analyzed by Hess et al. (2006) [26]. Among the 130 patients, 33 of them are from class 1 (achieved pCR), while the other 97 belong to class 2 (did not achieve pCR).

To evaluate the performance of the penalized precision matrix estimation



(a) (b)

Figure 4. *For $n = 400$, we get estimator of density generator g_p for normal and t.*

using three different penalties, we randomly divide the data into training and testing sets of sizes 109 and 21, respectively, and repeat the whole process 100 times. To maintain similar class proportion for the training and testing datasets, we use a stratified sampling: each time we randomly select 5 subjects from class 1 and 16 subjects from class 2 (both are roughly 1/6 of their corresponding total class subjects) and these 21 subjects make up the testing set; the remaining will be used as the training set. From each training data, we first perform a two-sample t-test between the two groups and select the most significant 120 genes that have the smallest p-values. In this case, the dimensionality p = 120 is slightly larger than the sample size n = 109 for training datasets in our classification study. Due to the noise accumulation demonstrated in Fan and Fan (2008), p = 120 may be larger than needed for optimal classification, but allows us to examine the performance when $p > n$. Second, we perform a gene-wise standardization by dividing the data with the corresponding standard deviation, estimated from the training dataset. Finally, we estimate the precision matrix and covariance matrix for both the classes for the training data using our method and standard graphical LASSO estimates. We find that our method gives sparser estimates of both inverse covariance and covariance matrices. The mean number of non zeros are given in the following table -

9. Conclusion.

We have developed adaptive estimation procedure for estimation of elliptical distribution for both low and high-dimensional cases. The method of estimation is novel and it gives us a way to move from fixed-dimensional to high-dimensional case quite naturally. We have developed

	Graphical LASSO	Our Method
Covariance Matrix	5312	4616
Inverse Covariance Matrix	486	412

Table 1
The number of non-zero elements in estimators of covariance and inverse covariance matrix in Breast Cancer Data.

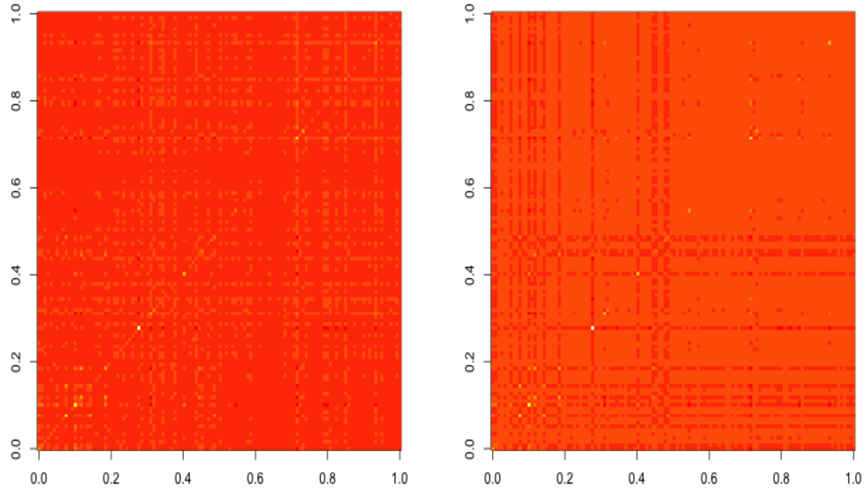


Figure 5. *Breast Cancer Data covariance matrix estimators using Graphical Lasso (Left) and our method (Right).*

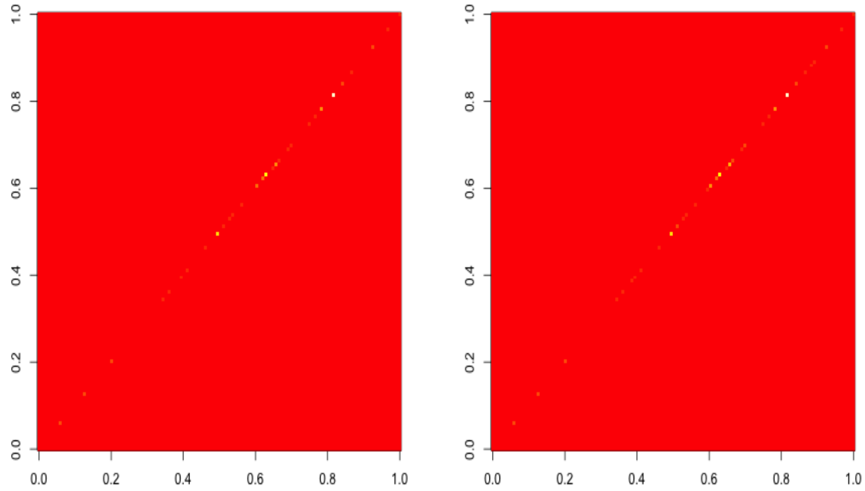


FIGURE 6. *Breast Cancer Data inverse covariance matrix estimators using Graphical Lasso (Left) and our method (Right).*

estimation procedure for the density generator function of the elliptical distribution in a log-linear spline form in Section 3 and derive respective error bounds. We use the estimate of density generator function of elliptical distribution to adaptively estimate Euclidean parameters of elliptical distribution.

For the estimation of Euclidean parameters, we devise a weighted loss function, where, the weights come from the slopes of estimated density generator function. As a result we have a very natural extension of squared error loss function and with the help of this weighted squared error loss function, we are able to estimate mean and covariance matrix parameters coming from distributions with widely varying tail behavior. So, we get robust estimates of the mean and covariance matrix.

Now, for the high-dimensions case too, weighted least squares loss function is a natural generalization of the least squares loss function, but with this simple generalized loss function, we are able to handle random variables coming from widely varying tail behaviors. As a result we can obtain estimators which are both regularized and robust in high-dimensions. Our approach is not the only approach in statistics literature which can produce estimators that have this dual property of being both robust to changing tail conditions and regularized to constrained parameter spaces in high-dimensions, but it is a quite natural one.

We have indicated three special cases, for which our method can be independently developed -(a) Estimation of Covariance and Precision matrix (b) Regression with Elliptical errors and (c) Clustering via mixtures of elliptical distribution in Section 6. For all of these cases, our method can give robust estimates, which can be regularized in high-dimensions.

Feasible algorithms are quite easily obtainable for our method, as most algorithms that work on least squares loss function also work for weighted least square loss functions too. So, we give an easy approximation to a hard optimization problem and try to solve an optimization problem, which is much easier to handle. As a result our method can borrow strength from existing optimization literature.

So, we have provided an estimation procedure of mean and covariance matrix parameters of elliptic distributions, which is adaptive to the tail behavior and given some theoretical justification for the estimators. The procedure can be extended to use in several classical statistical problems of regression, classification and clustering, thus making our method a very important stepping stone for developing future natural robust regularized estimators.

References.

[1] Anderson, T., Hsu, H. and Fang, K.-T. (1986). Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. *Canadian Journal of Statistics* **14** 55–59.

[2] Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: limit distribution theory and the spline connection. *Ann. Statist.* **35** 2536–2564. MR2382657 (2009b:62077)

[3] Battey, H. and Linton, O. (2012). Nonparametric estimation of multivariate elliptic densities via finite mixture sieves.

[4] Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR663424 (84a:62045)

[5] Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 (2009a:62255)

[6] Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 (2010b:62197)

[7] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 (2010j:62118)

[8] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models. Johns Hopkins Series in the Mathematical Sciences.* Johns Hopkins University Press, Baltimore, MD. MR1245941 (94m:62007)

[9] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data. Springer Series in Statistics.* Springer, Heidelberg. Methods, theory and applications. MR2807761 (2012e:62006)

[10] Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885 (2011h:62215)

[11] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 (2009b:62016)

[12] Chen, Y., Wiesel, A. and Hero, A. O. III (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Trans. Signal Process.* **59** 4097–4107. MR2865971 (2012h:62210)

[13] Cui, H. and He, X. (1995). The consistence of semiparametric estimation of elliptic densities. *Acta Math. Sin. New Ser* **11** 44–58.

[14] Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607. MR2758237

[15] Davis, R. A., Pfaffel, O. and Stelzer, R. (2011). Limit Theory for the largest eigenvalues of sample covariance matrices with heavy-tails. *arXiv preprint arXiv:1108.5464*.

[16] Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798 (2011b:62096)

[17] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166 (2005d:62116)

[18] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011 (2010d:62132)

[19] El Karoui, N. (2009). Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. MR2588248 (2011f:62061)

[20] Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric multivariate and related distributions. Monographs on Statistics and Applied Probability* **36**. Chapman and Hall Ltd., London. MR1071174 (91i:62070)

[21] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[22] Friedman, J. and Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26** 243–250.

[23] Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957). MR0093415 (19,1243c)

[24] Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser.* 539–555. Wadsworth, Belmont, CA. MR822052 (87i:62076)

[25] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust statistics. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics.* John Wiley & Sons Inc., New York. The approach based on influence functions. MR829458 (87k:62054)

[26] Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J. et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology* **24** 4236–4244.

[27] Holzmann, H., Munk, A. and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian journal of statistics* **33** 753–763.

[28] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. and Ravikumar, P. (2013). Sparse inverse covariance matrix estimation using quadratic approximation. *arXiv preprint arXiv:1306.3212*.

[29] Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics. MR606374 (82i:62057)

[30] Jonker, M. A. and van der Vaart, A. W. (2001). A semi-parametric model for censored and passively registered data. *Bernoulli* **7** 1–31. MR1811742 (2002b:62034)

[31] Kano, Y. (1994). Consistency property of elliptical probability density functions. *J. Multivariate Anal.* **51** 139–147. MR1309373 (96a:62057)

[32] Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19** 2102–2119. MR1135166 (92k:62104)

[33] Kiefer, J. and Wolfowitz, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **34** 73–85. MR0397974 (53 ##1829)

[34] Koenker, R. and Mizera, I. (2010). Quasi-concave density estimation. *Ann. Statist.* **38** 2998–3027. MR2722462 (2011j:62108)

[35] Komlós, J., Major, P. and Tusnády, G. (1975). An Approximation of Partial Sums of Independent rv'-s, and the Sample df. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **32** 111–131.

[36] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 (2011d:62064)

[37] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 (2004m:62130)

[38] Liebscher, E. (2005). A semiparametric density estimator based on elliptical distributions. *J. Multivariate Anal.* **92** 205–225. MR2102252 (2005h:62157)

[39] Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *J. Amer. Statist. Assoc.* **90** 330–341. MR1325140 (96e:62042)

[40] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 (2008b:62044)

[41] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

[42] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351 (2010e:62176)

[43] Pal, J. K. and Woodroofe, M. (2007). Large sample properties of shape restricted regression estimators with smoothness adjustments. *Statistica Sinica* **17** 1601.

[44] Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order restricted statistical inference. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. John Wiley & Sons Ltd., Chichester. MR961262 (90b:62001)

[45] Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372 (2010g:62186)

[46] Rufibach, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Stat. Comput. Simul.* **77** 561–574. MR2407642

[47] Srivastava, N. and Vershynin, R. (2011). Covariance Estimation for Distributions with 2+\ epsilon Moments. *arXiv preprint arXiv:1106.2775*.

[48] Stute, W. and Werner, U. (1991). Nonparametric estimation of elliptically contoured densities. In *Nonparametric Functional Estimation and Related Topics* 173–190. Springer.

[49] Tantiyaswasdikul, C. and Woodroofe, M. B. (1994). Isotonic smoothing splines under sequential designs. *Journal of Statistical Planning and Inference* **38** 75–87.

[50] Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR885734 (88g:62112)

[51] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. *Springer Series in Statistics*. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)

[52] Walther, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.* **24** 319–327. MR2757433 (2011j:62110)

[53] Wiesel, A. (2012). Unified framework to regularized covariance estimation in scaled Gaussian models. *IEEE Trans. Signal Process.* **60** 29–38. MR2932100

[54] Witten, D. M., Friedman, J. H. and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20** 892–900.

[55] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Department of Statistics
367 Evans Hall
Berkeley, CA, 94720
E-mail: sharmo@stat.berkeley.edu
bickel@stat.berkeley.edu