Department of Social Systems and Management

Discussion Paper Series

No. 1140

Conditional Minimum Volume Ellipsoid with Applications to Subset Selection for MVE Estimator and Multiclass Discrimination

by

Jun-ya Gotoh and Akiko Takeda

January 2006

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Conditional Minimum Volume Ellipsoid with Applications to Subset Selection for MVE Estimator and Multiclass Discrimination

Jun-ya Gotoh JGOTO@SK.TSUKUBA.AC.JP
Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573 Japan

Akiko Takeda TAKEDA@IS.TITECH.AC.JP
Department of Mathematical and Computing Sciences
Tokyo Institute of Technology
2-12-1 Oh-Okayama, Meguro-ku, Tokyo, 152-8552 Japan

Abstract

In this paper, we present a new formulation for constructing an ellipsoid which generalizes the computation of the minimum volume covering ellipsoid, based on the CVaR minimization technique proposed by Rockafellar and Uryasev (2002). The proposed ellipsoid construction is formulated as a convex optimization and an interior point algorithm for the solution can be developed. In addition, the optimization gives an upper bound of the volume of the ellipsoid associated with the MVE robust estimator, which fact can be exploited for approximate computations of the estimator.

Also, potential applicability of the new ellipsoid construction is discussed through two statistical problems: 1) robust statistics computations including outlier detection and the computation of the MVE estimator; 2) a multiclass discrimination problem, where the maximization of the normal likelihood function is characterized in the context of the ellipsoid construction. Numerical results are given, showing the nice computational efficiency of the proposed interior point algorithm and the capability of the proposed generalization.

Keywords: conditional value-at-risk (CVaR) optimization, minimum volume ellipsoid (MVE) estimator, minimum volume covering ellipsoid, multiclass discriminant analysis, interior point algorithm

1. Introduction

In various contexts concerned with statistics and multivariate analysis, there are several important estimators associated with the n-dimensional ellipsoid of the form

$$E(\boldsymbol{Q},\boldsymbol{\gamma}) := \{\, \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{Q}\boldsymbol{x}-\boldsymbol{\gamma}\|^2 \le n \,\}, \tag{1}$$

where $\boldsymbol{Q}$ is an $n \times n$ real symmetric positive definite matrix and $\boldsymbol{\gamma}$ is a vector in $\mathbb{R}^n$. Clearly, there is a one-to-one correspondence between $E(\boldsymbol{Q},\boldsymbol{\gamma})$ and another ellipsoid of the form

$$\hat{E}(\boldsymbol{D},\boldsymbol{c}) := \{\, \boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{x}-\boldsymbol{c}, \boldsymbol{D}(\boldsymbol{x}-\boldsymbol{c})\rangle \le n \,\}, \tag{2}$$

with change of variables as $\boldsymbol{D} = \boldsymbol{Q}^2$ and $\boldsymbol{c} = \boldsymbol{Q}^{-1}\boldsymbol{\gamma}$, where $\boldsymbol{c} \in \mathbb{R}^n$ represents the location or, equivalently, center of the ellipsoid and the positive definite matrix $\boldsymbol{D}$ represents the covariate structure. The volume of these ellipsoids is then given by

$$\frac{(n\pi)^{n/2}}{\Gamma(n/2+1)}\frac{1}{\det[\boldsymbol{Q}]} = \frac{(n\pi)^{n/2}}{\Gamma(n/2+1)}\frac{1}{\sqrt{\det[\boldsymbol{D}]}}, \tag{3}$$

where Γ is the gamma function.

An example of such an ellipsoid satisfies the following conditions:

(i) for a designated value $\beta \in (0, 1]$, it contains (at least) 100β percent of given n-dimensional points $\{\boldsymbol{x}^i : i \in I := \{1, \ldots, m\}\}$;

(ii) it attains the minimal volume.

This ellipsoid is known as the *minimum volume ellipsoid* with parameter β, denoted by β-MVE, and it is characterized by an optimal solution of the following optimization problem with a combinatorial constraint:

$$\text{(MVE}(\beta)\text{)} \quad \left| \begin{array}{ll} \underset{\boldsymbol{Q},\,\boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \left|\left\{\, i \in I \,:\, \|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2 \le n \right\}\right| \ge \lceil \beta m \rceil \\ & \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{4}$$

where $\lceil w \rceil$ denotes the minimum integer greater than or equal to w, and $\boldsymbol{Q} \succ \boldsymbol{O}$ means that $\boldsymbol{Q}$ is positive definite. The combinatorial constraint of (4) represents the above condition (i), whereas the objective implies the minimization of the volume (3) of ellipsoid $E(\boldsymbol{Q}, \boldsymbol{\gamma})$, corresponding to the condition (ii).

When β is set to be less than $1 - 1/m$, the β-MVE $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ provides robust estimators on the location of the data cloud as $\boldsymbol{Q}^{-1}\boldsymbol{\gamma}$ and the related scatter matrix as $\boldsymbol{Q}^2$. In particular, when β is set to be $(1 - \lceil -m/2 \rceil)/m$, the resulting estimator is shown to have the highest breakdown point, which implies that the resulting ellipsoid is immune against outliers consisting of at most a half of the data set (see Rousseeuw and Leroy, 1987, for details). In order to obtain such estimates, many algorithms for approaching a solution of Problem (4) have been developed. Global optimality of the problem is, however, hard to be assured because the combinatorial constraint in (4) possesses nonconvexity in general. Most of researches including Woodruff and Rocke (1993) applied heuristic algorithms for solving this problem since enumeration algorithms such as Cook et al. (1993) are computationally impractical. Hawkins (1993) proposes a two-phase framework for obtaining an ellipsoid which satisfies a necessary condition to be the β-MVE. Though this framework may work better than the enumeration algorithms, it will also be caught in a bind of the explosive increase of the computation time as the size of the data set grows.

On the other hand, when β is set to be 1.0, Problem (4) is equivalently reformulated as the following convex optimization problem

$$\left| \begin{array}{ll} \underset{\boldsymbol{Q},\,\boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 \le n, \quad (i \in I), \\ & \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{5}$$

and the resulting ellipsoid is called the *minimum volume covering ellipsoid*. This ellipsoid is known to be very useful in various contexts. For example, it is encountered in an experimental design problem (*e.g.*, Boyd and Vandenberghe, 2004; Titterington, 1975) as well as computational geometry (*e.g.*, Ben-Tal and Nemirovski, 2001). In contrast to Problem (4), the formulation (5) has a convex structure, and numerous algorithms have been developed. Among such algorithms are those of Barnes (1982), Khachiyan and Todd (1993), Sun and Freund (2004), Zhang and Gao (2003), Welzl (1991), and Gärtner and Schönherr (1997). Titterington (1975) also provides an algorithm based on the dual form of Problem (5) and it is employed as a subroutine in the implementation of the algorithms of Cook et al. (1993) and Hawkins (1993) so as to approach the β-MVE.

Associated with another important ellipsoid construction is a parameter estimation of elliptical distributions including the normal distribution, and they are characterized by a simultaneous density function of the form $p(\boldsymbol{x}) := \kappa \det[\boldsymbol{Q}]\, q(\|\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{\gamma}\|^2)$, where $\kappa > 0$ is a constant, and q is a function on $\mathbb{R}$. Clearly, its contour is concentric with $E(\boldsymbol{Q}, \boldsymbol{\gamma})$, where parameters $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$ are

often determined through the maximization of the likelihood function. For the normal distribution, $q(x) = \exp\{-x/2\}$ is adopted, and the maximum likelihood estimators $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$ can be obtained explicitly by using the covariance matrix and mean vector, respectively, of the given data points.

In this paper, we present a new formulation for constructing an ellipsoid which is achieved via a convex optimization. This formulation is based on the conditional value-at-risk (CVaR) minimization technique which is developed by Rockafellar and Uryasev (2002) and prevailing in the area of financial risk management. The CVaR of a random variable can be approximately viewed as a conditional expectation of the upper $100(1-\beta)$ percent of the variable. By using this property, the proposed ellipsoid is shown to give a generalization of the minimum volume covering ellipsoid through a parameterization with β. Though the β-MVE can also be considered to provide another generalization of the minimum volume covering ellipsoid through a similar parameterization, its exact computation is intractable as mentioned above. On the other hand, the generalized ellipsoid proposed in this paper is achieved via a solution of a convex optimization problem and an interior point algorithm can be applied. In addition, the formulation also generates the ellipsoid which is determined by the covariance matrix and mean vector of the data points $\boldsymbol{x}_i$, $i \in I$, as a special case with $\beta = 0$, while the β-MVE does not.

The structure of this paper is as follows. In Section 2, we set forth an optimization formulation for computing the generalized minimum volume covering ellipsoid. Section 3 is devoted to developing an interior point algorithm for solving the optimization problem, by following Sun and Freund (2004). In the two succeeding sections, we discuss the potential of the generalization by applying the proposed ellipsoid construction to two statistical problems. The first application discussed in Section 4 is associated with computation of the β-MVE. To approach a good robust estimator, we employ the proposed formulation for selecting a good subset of the data points. The second application discussed in Section 5 is a multiclass discrimination problem. It is shown that the proposed formulation also gives a generalization of the maximum likelihood estimation with the multivariate normal density. For both of the applications, numerical results are presented, showing the potential of the proposed formulation. Finally, Section 6 concludes the paper with some remarks.

2. Formulation of the Conditional Minimum Volume Ellipsoid

In this section, we introduce a generalized minimum volume ellipsoid by extending the problem formulation (5) of the usual minimum volume covering ellipsoid. Let $\{ \boldsymbol{x}^1, ..., \boldsymbol{x}^m \}$ be a given set of points in n-dimensional Euclidean space, and let us denote its index set by $I = \{ 1, ..., m \}$. As in Sun and Freund (2004), we suppose the following assumption throughout the paper.

Assumption 1 *The affine hull of* $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$ *spans* $\mathbb{R}^n$.

The ellipsoid proposed in this paper is then defined by an optimal solution of a nonlinear, but convex optimization problem formulated as follows:

$$\text{(CMVE}(\beta)\text{)} \left| \begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \phi_\beta(\boldsymbol{Q},\boldsymbol{\gamma}) \le n, \\ & \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{6}$$

where $\beta \in [0, 1)$ is a constant,

$$\phi_\beta(\boldsymbol{Q},\boldsymbol{\gamma}) := \min_\alpha \left\{ F_\beta(\boldsymbol{Q},\boldsymbol{\gamma},\alpha) := \alpha + \frac{1}{(1-\beta)m} \sum_{i\in I} \left[\left\| \boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma} \right\|^2 - \alpha \right]^+ \right\}, \tag{7}$$

and $[w]^+ := \max\{w, 0\}$. For an optimal solution $(\boldsymbol{Q},\boldsymbol{\gamma})$ of Problem (CMVE(β)), we define the *β-conditional minimum volume ellipsoid (β-CMVE)* as $E(\boldsymbol{Q},\boldsymbol{\gamma})$. The objective of Problem (6)

implies the minimization of the volume (3) of ellipsoid $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ as well as Problems (4) and (5), and $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$ in the left-hand side of the inequality constraint corresponds to the *β-conditional value-at-risk (β-CVaR)* proposed by Rockafellar and Uryasev (2002) in the context of financial risk management. In the following, let us first explain the meaning of this quantity in a brief manner by reviewing several results of Rockafellar and Uryasev (2002).

Let us first define the *ellipsoidal score* of data point i with respect to $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ by

$$f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) := f(\boldsymbol{x}^i | \boldsymbol{Q}, \boldsymbol{\gamma}) := \|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2, \quad i \in I.$$

For given $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$, we introduce the empirical distribution function of the score $f(\boldsymbol{x}|\boldsymbol{Q}, \boldsymbol{\gamma})$ as

$$\Phi(\,\alpha \,|\, \boldsymbol{Q}, \boldsymbol{\gamma}) := \frac{1}{m} \left| \left\{\, i \in I \,:\, f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \le \alpha \,\right\} \right|,$$

and the β-quantile of the scores for $\beta \in [0, 1)$ by

$$\begin{aligned} \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) &:= \min \{\, \alpha \ge 0 \,:\, \Phi(\,\alpha \,|\, \boldsymbol{Q}, \boldsymbol{\gamma}) \ge \beta \,\} \\ &= \min \left\{\, \alpha \ge 0 \,:\, \left| \left\{\, i \in I \,:\, f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \le \alpha \,\right\} \right| \ge \lceil \beta m \rceil \,\right\}. \end{aligned}$$

It should be noted that α_0 can be well defined by this definition since $f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \ge 0$, $i \in I$ for any $(\boldsymbol{Q}, \boldsymbol{\gamma})$.

According to Rockafellar and Uryasev (2002), ϕ_β is then shown to be equal to the mean of the ellipsoidal score under the β-tail distribution Φ_β which is defined by

$$\Phi_\beta(\eta|\boldsymbol{Q}, \boldsymbol{\gamma}) := \begin{cases} 0 & \text{for} \quad \eta < \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}), \\ (\Phi(\eta|\boldsymbol{Q}, \boldsymbol{\gamma}) - \beta)/(1 - \beta) & \text{for} \quad \eta \ge \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}), \end{cases}$$

and the following relation holds:

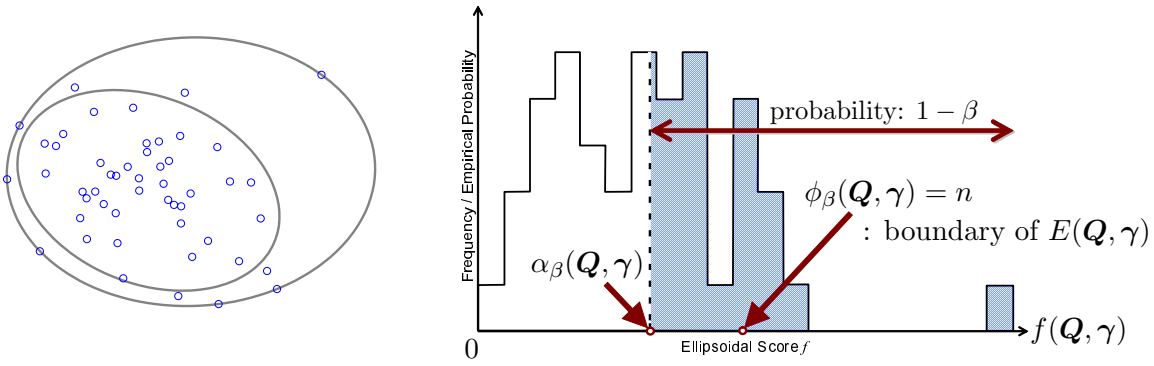
$$0 \;\le\; \alpha_\beta \;\le\; \mathrm{E}\,[\, f \,|\, f \ge \alpha_\beta \,] \;\le\; \phi_\beta \;\le\; \mathrm{E}\,[\, f \,|\, f > \alpha_\beta \,], \tag{8}$$

where E[·] denotes expectation operator under Φ, and $(\boldsymbol{Q}, \boldsymbol{\gamma})$ is omitted in (8) for notational simplicity. From these facts, we can see that the quantity ϕ_β is approximately equal to the expected value of the scores on the subset of data points whose score f^i ranks in the top $100(1-\beta)$ percent of all. Further, the inequality constraint in (6) is satisfied with equality at optimality as shown later via Theorem 2. Therefore, the ellipsoid obtained by solving Problem (6) is the minimum volume ellipsoid determined so that the mean ellipsoidal score of the higher $100(1-\beta)$ percent of the given data points will be located on the boundary of the ellipsoid. Figure 1(a) shows two-dimensional examples of the minimum volume ellipsoid covering fifty points and the β-CMVE with $\beta = 0.5$, and we see that an outlying data can affect the shape and the location of these two ellipsoids in a different manner. Figure 1(b) shows the histogram of the ellipsoidal scores of the points for the 0.5-CMVE in Figure 1(a). In this figure, the boundary of the ellipsoid corresponds to the $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$ which is approximately equal to the expected value of the ellipsoidal scores larger than the β-quantile $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$.

The following proposition clarifies an interpretation of the formulation (6).

Proposition 1 *For $\beta > 1 - \frac{1}{m}$, Problem (6) is equivalent to the minimum volume covering ellipsoid problem formulated as Problem (5). For $\beta = 0$, Problem (6) is equivalent to the following problem:*

$$\left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \dfrac{1}{m} \displaystyle\sum_{i \in I} \left\| \boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma} \right\|^2 \le n, \\ & \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{9}$$



(a) Two Ellipsoids (larger: the Minimum Volume Covering Ellipsoid; smaller: the 0.5-CMVE)

(b) Histogram of the Ellipsoidal Score f of the β-CMVE with $\beta = 0.5$

Figure 1: Geometric Interpretation of the Ellipsoid Construction

Since the left-hand side of the inequality constraint in (9) means the mean ellipsoidal score of all the points, we can view that when $\beta = 0$, Problem (CMVE(β)) determines the minimum volume ellipsoid so that the isoquant surface of the mean ellipsoidal score will form the boundary of the ellipsoid. In addition, by exploring the optimality condition, the unique solution $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$ is obtained via covariance matrix and mean vector as

$$\boldsymbol{Q}^* = \left(\frac{1}{m} \sum_{i \in I} (\boldsymbol{x}^i - \bar{\boldsymbol{x}})(\boldsymbol{x}^i - \bar{\boldsymbol{x}})^\top \right)^{-1/2} \quad ; \quad \boldsymbol{\gamma}^* = \boldsymbol{Q}^* \bar{\boldsymbol{x}}, \quad \text{where} \quad \bar{\boldsymbol{x}} := \frac{1}{m} \sum_{i \in I} \boldsymbol{x}^i. \tag{10}$$

Theorem 16 of Rockafellar and Uryasev (2002) ensures that under the existence of a solution, Problem (6) is equivalent to the following convex optimization problem:

$$\text{(CMVE}(\beta)\text{)} \left| \begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\alpha}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \alpha + \dfrac{1}{(1-\beta)m} \displaystyle\sum_{i \in I} \left[\left\| \boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma} \right\|^2 - \alpha \right]^+ \leq n, \\ & \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{11}$$

in the sense that $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*)$ solves (11) if and only if $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$ solves (6) and the inequality $F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*) \leq n$ holds. The existence of a solution to Problem (11) is ensured via the following theorem, which is shown in Appendix A.

Theorem 2 *Suppose that Assumption 1 holds. Problem (11) has a solution, and the inequality constraint $F_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha) \leq n$ of (11) is satisfied with equality at optimality.*

By combining Theorem 2 and the result of Rockafellar and Uryasev (2002), $F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*) = \phi_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*) = n$ holds at optimality, which implies that α^* is in $\operatorname{argmin}_\alpha F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha)$ and that α^* gives an approximate value of the β-quantile $\alpha_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$ for $\beta \in (0, 1)$ since $\operatorname{argmin}_\alpha F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha)$ is shown to be identical to the closed interval $[\alpha_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*), \alpha_\beta^+(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)]$ where $\alpha_\beta^+(\boldsymbol{Q}, \boldsymbol{\gamma}) := \inf\{\alpha : \Phi(\alpha | \boldsymbol{Q}, \boldsymbol{\gamma}) > \beta\}$. For another properties from optimization viewpoints, readers are referred to Rockafellar and Uryasev (2002).

3. An Algorithm for Solving Problem (CMVE(β))

It is clear that Problem (11) is rewritten as the following convex problem:

$$\text{(CMVE}(\beta)\text{)} \left| \begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\alpha,\boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \alpha + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} \le n, \\ & z_i \ge \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 - \alpha, \quad (i \in I), \\ & \boldsymbol{z} \ge \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{12}$$

where $\boldsymbol{e} = (1, \ldots, 1)^\top$ denotes the vector consisting of ones. Furthermore, the problem is transformed into

$$\text{(CMVE}(\beta)\text{)} \left| \begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & z_i \ge \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} - n, \quad (i \in I), \\ & \boldsymbol{z} \ge \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{13}$$

since Theorem 2 implies that the variable α can be deleted from (12) via $\alpha = n - \frac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m}$. Sun and Freund (2004) have proposed the "dual reduced Newton algorithm" for solving Problem (5), while they also utilized SDPT3 solver (see Toh et al., 1999) to solve the same problem and verified that the computational burden induced from the input form requirement of SDPT3 becomes prohibitive. Therefore, we propose in this section a Newton method to solve Problem (13) in a similar manner to Sun and Freund (2004). In order to apply the method, we add a logarithmic barrier function to (13) and obtain the formulation

$$\left| \begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\boldsymbol{z},\boldsymbol{t}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] - \theta_t \sum_{i\in I} \ln t_i - \theta_z \sum_{i\in I} \ln z_i \\ \text{subject to} & \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 - z_i + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \\ & \boldsymbol{z} > \boldsymbol{0}, \quad \boldsymbol{t} > \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{14}$$

We set positive values on the parameters θ_t and θ_z, and parameterized solutions to Problem (14) varying over $\theta_t \in (0, \infty)$ and $\theta_z \in (0, \infty)$ form the central trajectory of (14). Introducing Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ for the equality constraints in (14), the optimality conditions of (14) are as follows:

$$\sum_{i\in I} \lambda_i \{(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})\boldsymbol{x}^{i\top} + \boldsymbol{x}^i(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})^\top\} = \boldsymbol{Q}^{-1}, \tag{15}$$

$$\sum_{i\in I} \lambda_i(\boldsymbol{\gamma} - \boldsymbol{Q}\boldsymbol{x}^i) = \boldsymbol{0}, \tag{16}$$

$$(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})^\top(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}) - z_i + \frac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \tag{17}$$

$$\boldsymbol{\Lambda}\boldsymbol{t} = \theta_t \boldsymbol{e}, \tag{18}$$

$$\{\frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{E} - \boldsymbol{\Lambda}\}\boldsymbol{z} = \theta_z \boldsymbol{e}, \tag{19}$$

$$\boldsymbol{z} \ge \boldsymbol{0}, \quad \boldsymbol{t} \ge \boldsymbol{0}, \quad \boldsymbol{0} \le \boldsymbol{\lambda} \le \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \tag{20}$$

where $\boldsymbol{E}$ indicates $m \times m$ identity matrix and $\boldsymbol{\Lambda}$ is $m \times m$ diagonal matrix with diagonal elements $\boldsymbol{\lambda}$. It should be noted that the constraint $\boldsymbol{\lambda} \leq \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}$ is described as $\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\}\boldsymbol{\lambda} \geq \boldsymbol{0}$ where $\boldsymbol{U} := \boldsymbol{e}\boldsymbol{e}^\top$ is $m \times m$ matrix of ones.

Sun and Freund (2004) have introduced a logarithmic barrier function concerning slack variables $\boldsymbol{t}$ where $t_i = n - \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2$, into Problem (5) and have shown the optimality conditions as (15) through (18) with $\boldsymbol{z} = \boldsymbol{0}$. The conditions related to the variables $\boldsymbol{z}$ such as (19) and $\boldsymbol{\lambda} \leq \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}$ are additionally introduced for our problem.

The paper Sun and Freund (2004) proved that if $\boldsymbol{\lambda} > \boldsymbol{0}$, the matrix $\boldsymbol{Q}$ and vector $\boldsymbol{\gamma}$ are described with $\boldsymbol{\lambda}$ from (15) and (16), respectively, as

$$\boldsymbol{Q} = \left[2\left(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}\right)\right]^{-1/2}; \quad \boldsymbol{\gamma} = \frac{\boldsymbol{Q}\boldsymbol{X}\boldsymbol{\lambda}}{\boldsymbol{e}^\top\boldsymbol{\lambda}}, \tag{21}$$

where $\boldsymbol{X} := \left[\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m\right]$ denotes an $n \times m$ matrix which consists of a given set of vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$. By using (21), $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$ are deleted from the above optimality conditions, and (17) is rewritten as

$$h_i(\boldsymbol{\lambda}) - z_i + \frac{\boldsymbol{e}^\top\boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \tag{22}$$

where

$$h_i(\boldsymbol{\lambda}) := (\boldsymbol{x}^i - \frac{\boldsymbol{X}\boldsymbol{\lambda}}{\boldsymbol{e}^\top\boldsymbol{\lambda}})^\top \left[2\left(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}\right)\right]^{-1} (\boldsymbol{x}^i - \frac{\boldsymbol{X}\boldsymbol{\lambda}}{\boldsymbol{e}^\top\boldsymbol{\lambda}}).$$

Now we consider (18), (19), (20) and (22) as the optimality conditions for Problem (14). Note that the equation (22) indicates the feasibility of Problem (14) while (18) and (19) correspond to complementarity conditions for optimality. The Newton direction $(\Delta\boldsymbol{\lambda}, \Delta\boldsymbol{t}, \Delta\boldsymbol{z})$ for (18), (19) and (22) at a feasible solution $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{t}}, \bar{\boldsymbol{z}})$ is obtained by solving

$$\begin{cases} \nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})\Delta\boldsymbol{\lambda} + \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\Delta\boldsymbol{z} + \Delta\boldsymbol{t} = \boldsymbol{r}_1 := n\boldsymbol{e} - \boldsymbol{h}(\bar{\boldsymbol{\lambda}}) - \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\bar{\boldsymbol{z}} - \bar{\boldsymbol{t}}, \\ \bar{\boldsymbol{\Lambda}}\Delta\boldsymbol{t} + \bar{\boldsymbol{T}}\Delta\boldsymbol{\lambda} = \boldsymbol{r}_2 := \theta_t\boldsymbol{e} - \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{t}}, \\ \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}\Delta\boldsymbol{z} + \bar{\boldsymbol{Z}}\left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\Delta\boldsymbol{\lambda} = \boldsymbol{r}_3 := \theta_z\boldsymbol{e} - \frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\bar{\boldsymbol{z}} + \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{z}}, \end{cases} \tag{23}$$

where $\bar{\boldsymbol{\Lambda}}$, $\bar{\boldsymbol{T}}$ and $\bar{\boldsymbol{Z}}$ are $m \times m$ diagonal matrices with diagonal elements $\bar{\boldsymbol{\lambda}}$, $\bar{\boldsymbol{t}}$ and $\bar{\boldsymbol{z}}$, respectively. The formula to compute $\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})$ is described in Proposition 5 of Sun and Freund (2004) as

$$\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\boldsymbol{\lambda}) = -2\left(\frac{\Sigma(\boldsymbol{\lambda})}{\boldsymbol{e}^\top\boldsymbol{\lambda}} + \Sigma(\boldsymbol{\lambda}) \circ \Sigma(\boldsymbol{\lambda})\right),$$

where $\boldsymbol{A} \circ \boldsymbol{B}$ denotes the Hadamard product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, *i.e.*, $(\boldsymbol{A} \circ \boldsymbol{B})_{ij} := A_{ij}B_{ij}$ for all i, j, and $\Sigma(\boldsymbol{\lambda})$ is defined by

$$\Sigma(\boldsymbol{\lambda}) := (\boldsymbol{X} - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{e}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}})^\top \left[2\left(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}\right)\right]^{-1} (\boldsymbol{X} - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{e}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}).$$

The last two equalities of (23) lead to

$$\begin{cases} \Delta\boldsymbol{t} = \bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{r}_2 - \bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{T}}\Delta\boldsymbol{\lambda}, \\ \Delta\boldsymbol{z} = \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\boldsymbol{r}_3 - \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\bar{\boldsymbol{Z}}\left(\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right)\Delta\boldsymbol{\lambda}, \end{cases} \tag{24}$$

since the inverse matrices $\bar{\boldsymbol{\Lambda}}^{-1}$ and $\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E}-\bar{\boldsymbol{\Lambda}}\}^{-1}$ exist when we set $\theta_t>0$ and $\theta_z>0$. Then, by using (24), the first equality of (23) is transformed into

$$\Delta\boldsymbol{\lambda}=\boldsymbol{R}^{-1}\left[\boldsymbol{r}_1-\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E}-\bar{\boldsymbol{\Lambda}}\right\}^{-1}\boldsymbol{r}_3-\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{r}_2\right], \tag{25}$$

when the inverse matrix of

$$\boldsymbol{R}:=\left[\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})-\bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{T}}-\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E}-\bar{\boldsymbol{\Lambda}}\right\}^{-1}\bar{\boldsymbol{Z}}\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\right]$$

exists. Indeed, we have the inverse matrix $\boldsymbol{R}^{-1}$. $(\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})-\bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{T}})\prec O$ is ensured by Corollary 6 of Sun and Freund (2004), and

$$\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E}-\bar{\boldsymbol{\Lambda}}\right\}^{-1}\bar{\boldsymbol{Z}}\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\succ\boldsymbol{O}$$

is proved since $\bar{\boldsymbol{Z}}\succ\boldsymbol{O}$ and $\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E}-\bar{\boldsymbol{\Lambda}}\}\succ\boldsymbol{O}$.

We are now in a position to describe the modified dual reduced Newton algorithm.

Algorithm DRN. (A Modified Version of the Dual Reduced Newton Algorithm)

Step 0: (Initialization) Choose initial values of $(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})>\boldsymbol{0}$ satisfying $\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}\boldsymbol{\lambda}>\boldsymbol{0}$.

Step 1: (Stopping Criteria) Compute $OBJ:=-\ln\det[\boldsymbol{Q}]$ using (21). If the following inequalities

$$\|\,n\boldsymbol{e}-\boldsymbol{h}(\boldsymbol{\lambda})-\left\{\tfrac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\boldsymbol{z}-\boldsymbol{t}\,\|\le\epsilon_1;\quad\frac{\boldsymbol{\lambda}^\top\boldsymbol{t}}{OBJ}\le\epsilon_2;\quad\frac{\{\frac{\boldsymbol{e}^\top\boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}-\boldsymbol{\lambda}\}^\top\boldsymbol{z}}{OBJ}\le\epsilon_3$$

are satisfied, terminate the algorithm with an optimal solution $(\boldsymbol{Q},\boldsymbol{\gamma},\boldsymbol{z},\boldsymbol{t})$ of (14).

Step 2: (Newton Direction) Set $\theta_t\leftarrow\frac{\boldsymbol{\lambda}^\top\boldsymbol{t}}{10m}$ and $\theta_z\leftarrow\frac{\{\frac{\boldsymbol{e}^\top\boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}-\boldsymbol{\lambda}\}^\top\boldsymbol{z}}{10m}$. Compute $(\Delta\boldsymbol{z},\Delta\boldsymbol{t},\Delta\boldsymbol{\lambda})$ using (24) and (25).

Step 3: (Step-Size Computation) Compute

$$\bar{\beta}\leftarrow\max\left\{\beta:(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})+\beta(\Delta\boldsymbol{z},\Delta\boldsymbol{t},\Delta\boldsymbol{\lambda})\ge\boldsymbol{0},\ \{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}(\boldsymbol{\lambda}+\beta\Delta\boldsymbol{\lambda})\ge\boldsymbol{0}\right\}$$

and $\tilde{\beta}\leftarrow\min\{0.99\bar{\beta},1\}$. Set $(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})\leftarrow(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})+\tilde{\beta}(\Delta\boldsymbol{z},\Delta\boldsymbol{t},\Delta\boldsymbol{\lambda})$ and go to Step 1.

The Newton method presented above can be started from any point $(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})$ satisfying $(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})>\boldsymbol{0}$ and $\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}\boldsymbol{\lambda}>\boldsymbol{0}$. However, Sun and Freund (2004) pointed out that it is preferable to choose an initial point which guarantees the feasibility of Problem (14), that is, the point $(\boldsymbol{z},\boldsymbol{t},\boldsymbol{\lambda})$ satisfying (22). Such a point can be obtained as follows: Let $\bar{\boldsymbol{z}}>\boldsymbol{0}$ be any positive vector and $\bar{\boldsymbol{\lambda}}$ be some positive vector satisfying $\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}\bar{\boldsymbol{\lambda}}>\boldsymbol{0}$, for example, $\bar{\boldsymbol{\lambda}}=\frac{1}{m}\boldsymbol{e}$. From (22), we compute

$$\bar{\boldsymbol{t}}=n\boldsymbol{e}-\boldsymbol{h}(\bar{\boldsymbol{\lambda}})-\left\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\right\}\bar{\boldsymbol{z}}.$$

If $\boldsymbol{h}(\bar{\boldsymbol{\lambda}})+\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}\bar{\boldsymbol{z}}\le(0.95)n\boldsymbol{e}$, we set $\boldsymbol{z}=\bar{\boldsymbol{z}}$, $\boldsymbol{t}=\bar{\boldsymbol{t}}$ and $\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}$ as an initial point for Algorithm DRN. Otherwise, we see that some element of the vector $\boldsymbol{h}(\bar{\boldsymbol{\lambda}})+\{\frac{1}{(1-\beta)m}\boldsymbol{U}-\boldsymbol{E}\}\bar{\boldsymbol{z}}$ exceeds $0.95n$, and using a scaling parameter γ (>1) defined by

$$\gamma:=\frac{\max_{i\in I}\{h_i(\bar{\boldsymbol{\lambda}})+\frac{\boldsymbol{e}^\top\bar{\boldsymbol{z}}}{(1-\beta)m}-\bar{z}_i\}}{0.95n},$$

we obtain a strictly feasible solution of Problem (14) as

$$\boldsymbol{z} = \frac{1}{\gamma}\bar{\boldsymbol{z}}; \quad \boldsymbol{\lambda} = \gamma\bar{\boldsymbol{\lambda}}; \quad \boldsymbol{t} = n\boldsymbol{e} - \boldsymbol{h}(\boldsymbol{\lambda}) - \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\boldsymbol{z}.$$

Noting that $\boldsymbol{h}(\gamma\bar{\boldsymbol{\lambda}}) = \frac{1}{\gamma}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})$ holds, $\boldsymbol{t} = n\boldsymbol{e} - \frac{1}{\gamma}[\boldsymbol{h}(\bar{\boldsymbol{\lambda}}) + \{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\}\bar{\boldsymbol{z}}] \geq (0.05)n\boldsymbol{e}$ follows.

4. Application 1: Outlier Detection and Subset Selection for Computing β-MVE Estimator

In this and next sections, applications of the β-CMVE to problems arising from robust statistics and multiclass discrimination are presented. First of all, let us begin with a direct application to outlier detection problem.

Ellipsoidal peeling is one of the fundamental approaches for removing outliers (see Rousseeuw and Leroy, 1987). This method first computes the ordinary minimum volume covering ellipsoid by solving Problem (5) for an n-dimensional data set, and then removes the points located on the boundary of the ellipsoid as outliers. By replacing the covering ellipsoid by the β-CMVE, another simple approach for outlier detection can be constructed in a straightforward manner.

Algorithm OD. (Simple Outlier Detection)

Step 0. Choose $\beta \in [0, 1)$.

Step 1. Solve Problem (CMVE(β)), and let $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*)$ be an optimal solution.

Step 2. Remove a set of points whose ellipsoidal score $f^i(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$ is the largest as outliers.

A drawback of the original ellipsoidal peeling based on the ordinary minimum volume covering ellipsoid is that data points which are not outlying are removed unnecessarily because at least $n+1$ points lie on the boundary of the ellipsoid. However, the points to be removed by the above algorithm are expected to be fewer since the boundary of the optimal ellipsoid is determined by the mean value of ellipsoidal scores which are approximately greater than α^*, and only points with the largest score are to be deleted. Thus, the drawback of the original ellipsoidal peeling can be relaxed by introducing the β-CMVE.

The removal criterion in Step 2 of the above algorithm can be replaced by the other ones. For example, if we consider the data points with scores larger than $\phi_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*) = n$ as outliers, we can adopt the following step in place of the Step 2:

Step 2'. Remove a set of points defined by $\{\, i \,:\, f^i(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*) > n \,\}$, as outliers.

The number of points to be deleted by the revised algorithm with Step 2' is larger than or equal to that by the above algorithm with Step 2. Further, if we intend to remove (approximately) $100(1-\beta)$ percent of data points as outliers, the deleted subset may be defined as $\{\, i \,:\, f^i(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*) > \alpha^* \,\}$ since α^* gives a good approximate of the β-quantile of the ellipsoidal score.

Another interesting robust procedure is constructing the β-MVE estimator. As mentioned in Section 1, the exact computation of the β-MVE estimator is very hard, and many heuristic algorithms are proposed. Hawkins (1993) decomposes the problem into two phases: The first is selecting a subset of the data points, and the second is computing the minimum volume ellipsoid covering the selected subset of points. As listed in Section 1, numerous algorithms are proposed for the second phase problem. On the contrary, for the first phase subset selection problem, the number of researches is limited except random sampling algorithms.

In the remaining part of this section, we present three deterministic algorithms of the subset selection for approaching the β-MVE estimator by exploiting the β-CMVE computation. In order to obtain the β-MVE estimator with the highest breakdown point, we have to solve Problem (4)

with $\beta = (1 - \lceil -m/2 \rceil)/m$ which converges to 0.5 as $m \to \infty$. To consolidate the relation between the β-MVE and the β-CMVE, we introduce the following relation.

Lemma 3 $\left\{ (\boldsymbol{Q}, \boldsymbol{\gamma}) : \left| \{ i \in I : f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n \} \right| \geq \lceil \beta m \rceil \right\} = \{ (\boldsymbol{Q}, \boldsymbol{\gamma}) : \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n \}.$

This equivalence is obvious from the definition of $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$. From Lemma 3, Problem (MVE(β)) is equivalently rewritten as follows:

$$\text{(MVE}(\beta)\text{)} \quad \left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n, \\ & \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{26}$$

Though the expression has now become simpler, the difficulty of the problem remains almost the same. It is worthwhile that a solution of Problem (CMVE(β)) provides a feasible solution of Problem (26) and, accordingly, the optimal value is an upper bound of the volume of the β-MVE by noting the relation (8). In addition, an ellipsoid obtained by solving Problem (CMVE(β)) is intuitively a good approximate of the β-MVE when β is close to 1. Motivated by these facts, we describe three algorithms in which Problem (CMVE(β)) is iteratively optimized.

The basic strategy of the first algorithm is to gather a set of points according to the optimal ellipsoidal scores obtained by solving Problem (CMVE(β')) once. The first algorithm, denoted as SS1, is described as follows:

Algorithm SS1.

Input: $\beta, \beta' \in [0, 1)$, the set of data points $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^m\}$ and its index set $I = \{1, ..., m\}$.

Output: $G \subset I$

Step 0. Set $G \leftarrow \phi$.

Step 1. Compute the β'-CMVE by solving (CMVE(β')), and let $(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{\gamma}}, \hat{\alpha}, \hat{\boldsymbol{z}})$ be its optimal solution.

Step 2. Sort the ellipsoidal scores of all the points in ascending order as $f^{i_1}(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{\gamma}}) \leq \cdots \leq f^{i_m}(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{\gamma}})$, where $(i_1, ..., i_m)$ is the index vector after sorting.

Step 3. Set $G \leftarrow \{ i_j : j \leq \lceil \beta m \rceil \}$.

In the above description, we distinguish between the β for the MVE we want to compute and the β' for the CMVE computation. Intuitively, in the case of β close to 1, the optimal value of Problem (CMVE(β)) gives a good approximate of that of Problem (26) from Equation (8), so setting $\beta = \beta'$ is natural. However, for smaller β, *e.g.*, 0.5, we can expect better approximation by using different βs, *i.e.*, $\beta \neq \beta'$.

The second algorithm, denoted as SS2, applies the b-CMVE computation for different bs so as to obtain a subset G of size $\lceil \beta m \rceil$.

Algorithm SS2.

Input: $\beta \in [0, 1)$, the set of data points $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^m\}$ and its index set $I = \{1, ..., m\}$.

Output: $G \subset I = \{1, ..., m\}$

Step 0. Set $b \leftarrow \beta$, $H \leftarrow I$ and $G \leftarrow \phi$.

Step 1. Solve

$$\left|\begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\alpha,\boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \alpha+\dfrac{\boldsymbol{e}^\top\boldsymbol{z}}{(1-b)|H|}\le n, \\ & z_i\ge\left\|\boldsymbol{Q}\boldsymbol{x}^i-\boldsymbol{\gamma}\right\|^2-\alpha,\quad(i\in H), \\ & z_i\ge 0,\quad(i\in H), \\ & \boldsymbol{Q}\succ\boldsymbol{O},\ \boldsymbol{z}\in\mathbb{R}^{|H|}, \end{array}\right. \tag{27}$$

and let $(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}},\hat{\alpha},\hat{\boldsymbol{z}})$ be its optimal solution.

Step 2. Sort the ellipsoidal scores of all the points in ascending order as $f^{i_1}(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}})\le\cdots\le f^{i_m}(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}})$ where $(i_1,...,i_m)$ is the index vector after sorting.

Step 3. If $\{\,i:f^i(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}})>n\,\}\cap H\neq\phi$, then setting $H\leftarrow H\setminus\{\,i:f^i(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}})>n\,\}$ and $b\leftarrow\beta m/|H|$, repeat **Step 1.** Otherwise, set $G\leftarrow\{\,i_j:j\le\lceil\beta m\rceil\,\}$.

This algorithm is inspired by the ones proposed in Larsen et al. (2002), in which the minimization of the conditional value-at-risk (CVaR) is solved iteratively with various βs in order to obtain an approximate solution to the associated quantile (VaR) minimization whose exact solution is also difficult to be found. Their problem contains the CVaR measure in the objective while ours does in constraint. In addition, they impose the constraints on the ordering of the scores in their CVaR minimization problem, we do not impose such constraints in the optimization problem at Step 1. If such constraints are additionally imposed on (27) as in Larsen et al. (2002), the problem would result in having a nonconvex structure because of the nonlinearity of the ellipsoidal score f. By avoiding imposing such constraints, we can simply apply the interior point algorithm described in the previous section so as to solve Problem (27).

The third strategy, denoted as SS3, is a simple application of the modified ellipsoidal peeling stated at the beginning of this section.

Algorithm SS3.

Input: $\beta,\beta'\in[0,1)$, the set of data points $\{\boldsymbol{x}^1,...,\boldsymbol{x}^m\}$.

Output: $G\subset I=\{1,...,m\}$

Step 0. Set $G\leftarrow I$.

Step 1. Solve Problem (CMVE(β)) for G, and let $(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}},\hat{\alpha},\hat{\boldsymbol{z}})$ be its optimal solution.

Step 2. Let $i'\in\arg\max_i f^i(\hat{\boldsymbol{Q}},\hat{\boldsymbol{\gamma}})$ and set $G\leftarrow G\setminus\{\,i'\,\}$. Repeat Step 1 until $|G|<\lceil\beta m\rceil$ is satisfied.

This subset selection strategy solves Problem (CMVE(β)) for $m-\lceil\beta m\rceil$ times, which means the number of computations of Problem (CMVE(β)) increases proportionally to the sample size m. Conversely, since the number of points to be deleted is only one at each iteration, the resulting subset is selected in a prudent manner. Thus, we can expect that this strategy selects a better subset than the previous two strategies, SS1 and SS2.

In order to examine the performance of the above three strategies, computational experiments are conducted by applying eight data sets contained in Rousseeuw and Leroy (1987). Table 1 summarizes the abbreviated names of the test data to be used, and their size, *i.e.*, dimension, n, and the number, m, of samples. Though the size of each data set seems to be rather small in the context of contemporary data analysis, it is reasonable to use these data sets as a first step

Table 1: List of Test Data

name	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
dim. n	4	2	5	2	3	3	2	5
num. m	23	28	20	25	50	28	47	20

because they have been used as a benchmark in the researches dealing with algorithms for the MVE computation. All computations were performed on a personal computer (Dell Precision 370, CPU: Pentium 4, 3.80GHz, RAM: 2GB, OS: Windows) and implemented with MATLAB (R14SP2). For simple presentation, we apply $\beta' = \beta$ for SS1 and SS3 in the following experiments.

Tables 2 (a), (b) and (c) compare the log-volume, $-\ln\det[\boldsymbol{Q}]$, obtained by several strategies for $\beta = 0.9, 0.7$ and 0.5, respectively. Each table consists of four pairs of rows. The upper row of the first to third pairs shows the log-volume achieved by each strategy (SS1, SS2, or SS3, respectively), whereas the lower row of those shows the log-volume achieved by the combination of each proposed strategy and the pairwise swap strategy called the "feasible solution algorithm (FSA)" in Hawkins (1993).

The FSA guarantees that the resulting ellipsoid cannot be improved by any pairwise swap of two sample points, and such an ellipsoid is called a "feasible solution." In our experiments, we repeat the following steps until any "feasible solution" is found: 1. We first apply each subset selection strategy proposed above and then compute the minimum volume ellipsoid covering the selected subset of the data points by employing Titterington's algorithm (Titterington, 1975); 2. We next check if the obtained ellipsoid satisfies the necessary condition to be the β-MVE by checking all possible pairwise swaps, and replace the subset by better one if any. In order to reduce the computation time for evaluating each minimum volume covering ellipsoid, we employ an aborting strategy proposed in Hawkins (1993) which quits the computation as soon as the volume of ellipsoid is found to be larger than the incumbent value.

The upper row of the fourth pair of rows in Tables 2 (a) to (c) shows the smallest log-volume achieved through fifty FSA trials, each starting from a randomly selected subset of the data points, while the lower shows the number of the random subsets which attain the best achieved volume, indicating easiness for finding the minimum volume ellipsoid. As is easily expected, when β gets closer to 0.5, Problem (MVE(β)) is likely to have many nonoptimal local solutions and accordingly, local search algorithms including the FSA may get stuck in such a solution. This can be observed especially for instances with larger number m of samples, *e.g.*, Edu and Sta, and larger dimensions n, *e.g.*, Col and Woo.

From Table 2, we see that when $\beta = 0.9$, the subset selection achieving smaller ellipsoid is not so hard and, accordingly, any combination of each of three proposed strategies and the FSA results in the smallest volume of the fifty trials, except Sta data which is mentioned as a hard problem in Hawkins (1993). In addition, we cannot find any significant difference among the initial subset selections via the three proposed strategies. As β gets closer to 0.5 and the subset selection becomes more cumbersome, the three proposed strategies show some difference. The volume of the initial subset chosen by SS1 is constantly larger than those by the other two, though the smaller initial ellipsoid does not imply the smaller ellipsoid via the combination with the FSA. As for SS2 and SS3, we cannot conclude which strategy is superior.

Tables 3 (a) through (c) summarize the number of updates required until reaching a "feasible solution" for $\beta = 0.5, 0.7$ and 0.9, respectively. The upper three rows in each table show the number of subset updates during the FSA phase starting from the subset obtained by the proposed strategies, whereas the lower four rows show the average, the (unbiased) standard deviation, the maximum, and the minimum, respectively, of the number of the subset updates among fifty trials of the random subset selection. From the tables (b) and (c), we see that the number of the updates

Table 2: Log-Volume, $-\ln\det[\boldsymbol{Q}]$, of Achieved Ellipsoids

(a) $\beta = 0.9$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1	19.0069	0.8367	3.0536	7.7110	14.9312	3.1013	-1.0673	-16.0943
SS1+FSA	17.4473	0.7781	2.6407	7.2108	14.4952	2.7992	-1.0983	-16.0943
SS2	19.0069	0.8367	3.0536	7.7110	14.7597	3.1013	-1.0063	-16.0943
SS2+FSA	17.4473	0.7781	2.6407	7.2108	14.4952	2.7992	-1.0983	-16.0943
SS3	19.0069	0.9330	3.3848	7.9213	14.6100	2.9353	-1.0983	-15.9399
SS3+FSA	17.4473	0.7781	2.6407	7.2108	14.4952	2.7992	-1.0983	-16.0943
min. 50-FSA	17.4473	0.7781	2.6407	7.2108	14.4952	2.7992	-1.6741	-16.0943
#(attained)	50	50	50	50	50	50	2	50

(b) $\beta = 0.7$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1	17.1917	0.3187	1.8023	6.8647	13.8013	2.7896	-2.0487	-16.6535
SS1+FSA	16.1918	-0.7144	1.4295	6.2597	13.7714	1.8795	-2.9991	-17.1574
SS2	16.5934	-0.6920	1.6807	6.7613	13.7629	1.8795	-2.9798	-16.6535
SS2+FSA	16.1918	-0.7144	1.4295	6.2597	13.7407	1.8795	-2.9991	-17.1574
SS3	16.5432	-0.6920	1.4660	6.7849	13.8190	1.9431	-3.0526	-17.1103
SS3+FSA	16.3948	-0.7144	1.4295	6.2597	13.4906	1.8795	-3.0851	-17.1574
min. 50-FSA	16.1810	-0.7647	1.3185	6.2597	13.4906	1.8795	-3.0851	-17.1574
#(attained)	28	4	33	33	11	39	36	26

(c) $\beta = 0.5$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1	16.3823	-1.2902	0.5839	6.6183	13.3759	1.7692	-2.8204	-17.7550
SS1+FSA	14.7993	-1.8291	-0.6165	5.6395	12.7972	1.0899	-3.4529	-19.6479
SS2	15.5503	-1.4027	0.1447	5.8444	12.9566	1.4876	-3.4222	-17.7550
SS2+FSA	14.4443	-1.8291	-0.3937	5.6395	12.7212	0.7952	-3.4529	-19.6479
SS3	15.3103	-1.3551	0.1928	5.7727	12.9867	1.3623	-3.4724	-19.5037
SS3+FSA	14.4443	-1.8291	-0.6165	5.6395	12.7897	1.1479	-3.4724	-19.6479
min. 50-FSA	14.4443	-1.8291	-0.8752	5.6395	12.7212	0.7952	-3.5418	-20.2611
#(attained)	28	33	11	39	2	15	5	2

Table 3: Number of Updates up to a Feasible Solution

(a) $\beta = 0.9$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	2	1	2	1	3	1	1	0
SS2+FSA	2	1	2	1	2	1	1	0
SS3+FSA	2	2	2	2	1	1	0	1
ave.(50-FSA)	1.82	1.92	1.92	1.88	4.66	1.92	3.52	1.78
stdev.(50-FSA)	(0.44)	(0.27)	(0.53)	(0.33)	(0.69)	(0.27)	(0.68)	(0.42)
max.(50-FSA)	2	2	3	2	6	2	4	2
min.(50-FSA)	0	1	1	1	3	1	2	1

(b) $\beta = 0.7$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	3	5	2	3	1	4	4	2
SS2+FSA	2	1	2	2	1	0	2	2
SS3+FSA	1	1	1	2	3	1	2	1
ave.(50-FSA)	4.40	6.90	4.46	5.28	11.28	6.02	12.54	3.36
stdev.(50-FSA)	(1.18)	(1.78)	(1.43)	(1.23)	(2.27)	(1.76)	(1.91)	(1.01)
max.(50-FSA)	7	10	8	8	17	10	16	6
min.(50-FSA)	2	3	2	3	7	3	7	1

(c) $\beta = 0.5$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	4	1	2	4	7	2	4	3
SS2+FSA	5	2	1	2	4	5	1	3
SS3+FSA	3	2	3	1	2	3	0	1
ave.(50-FSA)	5.98	8.08	4.42	6.54	15.42	7.10	15.36	4.34
stdev.(50-FSA)	(1.83)	(2.04)	(1.81)	(2.01)	(3.35)	(1.97)	(2.78)	(1.24)
max.(50-FSA)	10	12	9	12	26	12	22	8
min.(50-FSA)	3	4	2	3	8	3	9	2

via the proposed strategies is stably smaller than that by the random ones. This is emphasized in case of huge number of samples as we will see later, even though the FSA schemes can save computation time by making use of the aborting strategy

Table 4 shows the computational CPU time on the second time scale. The second row of first to three pairs of rows reports the CPU time spent for obtaining the initial subset by each strategy. From this, we see that single application of each strategy is very efficient. The fourth pair of rows shows the average CPU time of fifty trials starting from randomly selected subsets, and the number in parentheses is its standard deviation. We see that the FSA starting from a subset determined by the proposed strategies reaches a "feasible solution" in a more efficient manner on average than that from a randomly selected subset. In particular, when the number of samples is large, *e.g.*, Edu or Sta, the computation time for the FSA increases remarkably, while the time for solving (11) increases gradually.

In order to confirm this aspect, we further apply our algorithms to randomly generated data of larger size with $(n, m) = (5, 100)$ and $(5, 300)$. Tables 5 and 6 report the computational performance, showing the average, the standard deviation, the maximum and the minimum of the CPU time among ten different instances drawn from a five-dimensional composite distribution of two normal ones with different variances. When $(n, m) = (5, 300)$ and the subset update starts from the randomly generated subset, result is not given in Table 6 because the first problem out

Table 4: CPU Time for Small Data Sets [sec.]

(a) $\beta = 0.9$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	1.83	0.91	1.78	0.44	4.11	1.56	7.58	0.47
SS1	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.00
SS2+FSA	1.91	0.89	1.80	0.48	5.95	1.66	3.42	0.45
SS2	0.02	0.02	0.02	0.03	0.05	0.03	0.03	0.02
SS3+FSA	2.00	1.05	1.45	0.70	0.72	1.61	1.91	0.83
SS3	0.05	0.03	0.03	0.03	0.09	0.05	0.06	0.02
ave. (50 FSA)	2.00	2.70	1.72	0.88	64.05	3.06	8.47	2.75
stdev.	(0.61)	(0.97)	(0.49)	(0.30)	(25.86)	(0.64)	(2.22)	(1.37)

(b) $\beta = 0.7$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	2.56	3.55	0.75	3.83	13.31	4.38	7.69	2.58
SS1	0.03	0.02	0.02	0.03	0.05	0.03	0.05	0.00
SS2+FSA	1.13	0.36	0.64	3.25	2.58	0.38	3.61	2.56
SS2	0.02	0.05	0.05	0.03	0.06	0.05	0.05	0.00
SS3+FSA	0.25	0.36	0.48	3.17	10.58	0.75	3.86	1.25
SS3	0.05	0.06	0.03	0.05	0.19	0.06	0.11	0.02
ave. (50 FSA)	7.60	9.81	3.15	5.40	186.86	16.87	102.29	7.40
stdev.	(3.99)	(9.46)	(2.15)	(2.10)	(118.29)	(12.07)	(63.34)	(8.19)

(c) $\beta = 0.5$	Air	Bra	Col	Del	Edu	Sal	Sta	Woo
SS1+FSA	3.00	0.52	0.42	2.84	13.94	0.80	23.91	0.52
SS1	0.03	0.05	0.02	0.05	0.05	0.03	0.03	0.00
SS2+FSA	2.67	0.58	0.17	2.70	5.45	1.72	0.91	0.53
SS2	0.03	0.05	0.03	0.05	0.08	0.06	0.06	0.00
SS3+FSA	2.69	0.39	0.70	2.50	7.45	0.72	0.33	0.22
SS3	0.06	0.09	0.05	0.08	0.25	0.09	0.09	0.02
ave. (50 FSA)	8.09	11.10	1.23	10.99	227.67	13.15	146.99	1.92
stdev.	(5.04)	(9.01)	(0.92)	(7.77)	(137.70)	(9.75)	(93.76)	(1.53)

Table 5: CPU Time for 10 Random Data with $(n, m) = (5, 100)$ [sec.]

$\beta = 0.5$	ave.	stdev.	max.	min.
SS1+FSA	36.3	(18.0)	73.1	16.1
SS2+FSA	39.8	(18.7)	79.1	11.9
SS3+FSA	38.0	(13.8)	55.8	12.2
random FSA	1753.4	(593.6)	2719.2	962.7

Table 6: CPU Time for 10 Random Data with $(n, m) = (5, 300)$ [sec.]

$\beta = 0.5$	ave.	stdev.	max.	min.
SS1+FSA	810.38	(296.19)	1407.58	419.19
SS1	0.83	(0.06)	0.92	0.73
SS2+FSA	1063.50	(281.92)	1494.94	635.13
SS2	2.09	(0.16)	2.36	1.92
SS3+FSA	883.09	(412.83)	1712.50	403.98
SS3	58.39	(1.36)	60.41	55.61
random+FSA	-	-	-	-

of ten could not reach optimality within 36 hours. From theses tables, we see that every proposed strategy achieves a "feasible subset" more efficiently than the randomly selected strategy, and that the time spent for obtaining an initial candidate is much smaller than that for the updating. Tables 7 and 8 show the average number of updates starting from the initial subset to any "feasible solution." From these tables, we confirm that every proposed strategy takes smaller number of updates compared to the random FSA. This is the reason why the computational performances of the three proposed strategies are superior to the randomly generated FSA application.

5. Application 2: Multiclass Discrimination

In this section, we turn to apply the β-CMVE to a multiclass discriminant analysis motivated by the second statement of Proposition 1.

Under the normality assumption, the log-likelihood function with observations $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^m\}$ is defined by

$$\ell(\boldsymbol{Q}, \boldsymbol{\gamma}) = m \ln \det[\boldsymbol{Q}] - \frac{1}{2} \sum_{i \in I} f^i(\boldsymbol{Q}, \boldsymbol{\gamma}). \tag{28}$$

Table 7: Number of Updates through FSA with $(n, m) = (5, 100)$

$\beta = 0.5$	ave.	stdev.	max.	min.
SS1+FSA	4.4	(2.0)	8	2
SS2+FSA	4.8	(2.0)	8	1
SS3+FSA	4.3	(2.2)	8	1
random FSA	35.4	(3.2)	40	30

Table 8: Number of Updates through FSA with $(n, m) = (5, 300)$

$\beta = 0.5$	ave.	stdev.	max.	min.
with SS1	6.8	(1.87)	10	4
with SS2	7.7	(2.11)	11	5
with SS3	6.6	(4.22)	16	2
random	-	-	-	-

Associated with the maximization of the normal log-likelihood function (28), let us consider the following optimization problem:

$$\left| \underset{\boldsymbol{Q} \succ \boldsymbol{O}, \gamma, \alpha}{\text{maximize}} \quad m \ln \det [\boldsymbol{Q}] - \frac{1}{2} \sum_{i \in I} \left(\alpha + \frac{1}{1-\beta} \left[f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) - \alpha \right]^+ \right). \right. \tag{29}$$

The difference between the normal log-likelihood function (28) and the objective of Problem (29) is found in their second terms. From Proposition 1 and the equivalence between (6) and (11), the second term in the former can be viewed as a special case of that of the latter with $\beta = 0$.

More directly, the following proposition shows the equivalence between the generalized log-likelihood maximization (29) and Problem (CMVE(β)). The proof is provided in Appendix A.

Proposition 4 *Problem (29) and Problem (12) have the same optimal solution.*

From this proposition, Problem (CMVE(β)) can be regarded as the maximization of the generalized log-likelihood function (29). In other words, optimal ellipsoid via Problem (CMVE(β)) is determined so that the conditional normal likelihood of data points whose ellipsoidal score ranks in the top $100(1-\beta)$ percent of all, would be maximal. This fact is consistent with that Problem (CMVE(β)) with $\beta = 0$ characterizes the covariance matrix and the mean vector as in (10).

Once $(\boldsymbol{Q}, \boldsymbol{\gamma})$ is obtained in any way, a multivariate normal density can be identified. One of the most useful applications of this estimation is discriminant analysis. For multiclass discrimination, one can use the normal likelihood function (28) as in Fisher's discriminant model. More specifically, we first estimate $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ by solving Problem (12) for each class $k \in K$, and then, for a new sample $\bar{\boldsymbol{x}}$, we assign its class label $\bar{k}$ as

$$\bar{k} \in \arg\max_{k \in K} \left\{ \ln \det[\boldsymbol{Q}_k] - \frac{1}{2} f(\bar{\boldsymbol{x}} | \boldsymbol{Q}_k, \boldsymbol{\gamma}_k) \right\}, \tag{30}$$

where K is the index set of classes.

Another possibly promising criterion is the comparison of the modified Mahalanobis distances, *i.e.*, for a new sample $\bar{\boldsymbol{x}}$, its class label $\bar{k}$ is assigned as

$$\bar{k} \in \arg\min_{k \in K} \ (\bar{\boldsymbol{x}} - \boldsymbol{c}_k)^\top \boldsymbol{D}_k (\bar{\boldsymbol{x}} - \boldsymbol{c}_k), \tag{31}$$

where the metric of the squared distance is given by $\boldsymbol{D}_k = (\boldsymbol{Q}_k)^2$ and the origin by $\boldsymbol{c}_k = \{\boldsymbol{Q}_k\}^{-1} \boldsymbol{\gamma}_k$.

For both of the criteria, each estimate $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ depends on only one β, say, β_k, and, therefore, different βs can be applied to different classes.

In order to examine the potential of the proposed multiclass classification model, ten-fold cross-validation is carried out using two famous data sets in Blake and Merz (1998): (i) Iris (three classes) and (ii) WDBC Cancer (two classes). Table 9 shows the total learning (in-sample) and testing (out-sample) error rates with two different criteria: (a) via the normal likelihood (30) and (b) via the modified Mahalanobis distance (31) when the Iris data is applied. For ease of presentation, β is set to be common in all three classes. From this table, we see that nicely low test error rate is achieved

Table 9: Learning and Testing Error Rates of 10-Fold Cross-Validation for Iris data

		(a) Likelihood		(b) Mahalanobis	
		learn	test	learn	test
β	0.05	2.15%	3.33%	4.81%	6.67%
	0.10	2.00%	3.33%	4.22%	6.67%
	0.15	1.93%	3.33%	3.63%	6.67%
	0.20	1.85%	2.00%	3.33%	5.33%
	0.25	1.78%	2.00%	3.11%	4.00%
	0.30	1.70%	2.67%	2.81%	4.00%
	0.35	1.63%	2.67%	2.52%	4.00%
	0.40	1.63%	2.00%	2.37%	4.00%
	0.45	1.56%	2.67%	2.15%	3.33%
	0.50	1.41%	2.67%	1.93%	2.67%
	0.55	1.63%	2.67%	1.63%	2.67%
	0.60	1.48%	2.67%	1.41%	2.67%
	0.65	1.33%	2.67%	1.56%	3.33%
	0.70	1.33%	2.67%	1.70%	3.33%
	0.75	1.48%	3.33%	2.00%	2.67%
	0.80	1.70%	3.33%	2.07%	2.67%
	0.85	1.93%	4.67%	1.93%	4.67%
	0.90	3.48%	5.33%	2.00%	2.00%
	0.95	3.48%	5.33%	2.00%	2.00%

by searching β over $(0, 1)$. For the two criteria, the lowest rate is achieved at different β, and we see that the selection of criterion is crucial for the results, and more elaborative analysis is needed for the selection.

Tables 10 (i) to (iii) show the total learning and testing error rates via the two criteria and Fisher's methods when the WDBC cancer data is applied. Apart from the case of the iris data, we computed with various combinations of βs for two classes, *i.e.*, β_1 and β_2. It is worth noting that the required number of times for solving Problem (CMVE(β)) is only $2N$ (not N^2) where N is the number of subdivisions of each β, because each ellipsoid is determined based on only one parameter β, though each error rate is evaluated by the comparison between two ellipsoids with different βs.

Mangasarian et al. (1995) formulated a linear programming discriminant model and achieves fairly low predictive error with this data set. One of the authors applied an extended quadratic model of their linear model, and sees it difficult to outperform their model (Konno et al., 2002). From Table 10 (ii)-(d), we see that a nicely small misclassification result which is comparable to that in Mangasarian et al. (1995) is achieved via the modified Mahalanobis distance criterion when $(\beta_1, \beta_2) = (0.93, 0.76)$ and $(0.93, 0.77)$. In addition, comparing with the Fisher's discrimination methods which may be viewed as special cases in our model, we see that from Table 10 (iii), ours is more predictive than the Fisher's classical models.

6. Concluding Remarks

In this paper, we provide a new formulation for constructing an ellipsoid from a set of given data points in $\mathbb{R}^n$, based on the CVaR technique proposed by Rockafellar and Uryasev (2002). This formulation yields a generalized notion of both the minimum volume ellipsoid covering all the data points and the ellipsoid characterized via the maximum likelihood estimators of the normal distribution. Computation of the generalized ellipsoid is accomplished through a convex optimization and a modified version of an interior point algorithm developed by Sun and Freund (2004) can solve it in a fairly efficient manner.

Motivated by such generalizing property and computational accessibility, we exploit this ellipsoid construction in two statistical applications. We first apply it to selecting the subset of given data points for approaching the MVE estimator, which is defined via the minimum volume ellipsoid containing inside a certain proportion of the given data points. Though seeking the MVE in a direct manner is known to be very tough, preliminary experiments demonstrate that the proposed algorithms achieve as small an ellipsoid as by the random heuristic algorithm proposed in Hawkins

Table 10: Learning and Testing Error Rates of 10-Fold Cross-Validation for WDBC Cancer data

(i) via the Normal Log-Likelihood (30)

(a) learning error

β_1 \ β_2	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.05	3.91%	4.04%	4.24%	4.35%	4.57%	4.96%	5.23%	5.23%	5.29%	5.70%
0.15	3.91%	3.85%	3.93%	4.06%	4.30%	4.55%	4.69%	4.98%	5.10%	5.94%
0.25	3.77%	3.73%	3.73%	3.81%	4.12%	4.14%	4.22%	4.51%	5.14%	5.88%
0.35	3.53%	3.59%	3.73%	3.79%	3.85%	4.02%	4.26%	4.51%	5.21%	6.17%
0.45	3.53%	3.48%	3.59%	3.65%	3.81%	3.91%	4.16%	4.63%	5.14%	6.62%
0.55	3.51%	3.61%	3.61%	3.57%	3.81%	4.00%	4.28%	4.78%	5.49%	7.46%
0.65	3.77%	3.55%	3.55%	3.67%	3.96%	4.18%	4.61%	5.00%	5.47%	7.85%
0.75	3.75%	3.61%	3.55%	3.50%	3.96%	4.41%	4.92%	5.35%	5.76%	8.40%
0.85	4.20%	4.06%	3.91%	3.91%	4.16%	4.92%	5.33%	5.60%	6.82%	9.18%
0.95	4.55%	4.35%	4.37%	4.35%	4.28%	4.80%	5.74%	6.29%	8.06%	10.56%

(b) testing error

β_1 \ β_2	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.05	4.22%	4.39%	4.39%	4.39%	4.57%	5.27%	5.80%	5.80%	5.27%	5.80%
0.15	3.87%	4.04%	4.04%	4.22%	4.57%	4.57%	5.10%	5.10%	5.45%	5.98%
0.25	3.69%	3.69%	4.04%	3.87%	4.22%	4.39%	4.22%	4.39%	5.27%	5.98%
0.35	3.69%	3.69%	3.87%	4.04%	4.22%	4.22%	4.22%	4.75%	5.27%	6.33%
0.45	3.69%	3.69%	3.87%	4.04%	3.87%	4.22%	4.39%	4.75%	5.45%	6.68%
0.55	3.87%	4.04%	3.69%	3.51%	4.04%	4.04%	4.57%	5.27%	5.62%	7.73%
0.65	3.87%	4.04%	4.04%	4.04%	4.39%	4.39%	4.92%	5.10%	5.80%	7.91%
0.75	4.22%	4.22%	4.22%	3.87%	4.22%	4.75%	5.10%	5.45%	5.80%	8.44%
0.85	4.39%	4.22%	4.04%	3.87%	4.22%	5.10%	5.45%	5.80%	7.03%	9.31%
0.95	4.75%	4.57%	4.75%	4.75%	4.57%	5.10%	5.98%	6.50%	8.61%	10.72%

(ii) via the Modified Mahalanobis' Distance (31)

(a) learning error

β_1 \ β_2	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.05	8.51%	10.08%	11.89%	13.47%	14.88%	15.93%	18.08%	21.44%	25.56%	25.60%
0.15	6.35%	8.18%	9.78%	11.50%	13.26%	14.80%	16.19%	19.12%	22.16%	22.57%
0.25	4.92%	6.29%	7.85%	9.51%	11.09%	12.89%	15.04%	17.22%	19.61%	20.19%
0.35	4.18%	4.73%	5.99%	7.36%	8.92%	10.29%	12.69%	15.50%	17.97%	17.87%
0.45	3.75%	4.10%	4.57%	5.84%	7.05%	8.49%	10.29%	13.83%	16.17%	15.60%
0.55	3.30%	3.53%	3.83%	4.22%	5.35%	6.35%	8.16%	10.76%	14.02%	13.42%
0.65	3.03%	3.05%	3.28%	3.51%	3.83%	4.59%	5.98%	8.47%	10.94%	12.09%
0.75	3.51%	3.24%	3.14%	3.03%	3.48%	3.55%	4.37%	5.98%	8.44%	9.88%
0.85	4.24%	4.00%	3.83%	3.38%	3.14%	2.97%	2.95%	3.91%	5.17%	6.91%
0.95	6.05%	5.21%	4.75%	4.37%	4.08%	3.91%	3.85%	3.01%	3.03%	4.43%

(b) testing error

β_1 \ β_2	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.05	8.79%	10.37%	12.30%	13.53%	15.11%	16.34%	18.63%	21.97%	25.83%	25.83%
0.15	6.33%	8.44%	10.19%	11.95%	13.53%	15.11%	16.70%	19.16%	23.02%	23.02%
0.25	5.80%	6.15%	8.26%	9.84%	11.60%	13.36%	15.47%	17.75%	19.68%	20.39%
0.35	4.22%	5.10%	6.15%	7.56%	9.14%	10.54%	13.18%	15.82%	17.57%	18.45%
0.45	3.69%	4.04%	5.10%	6.15%	7.21%	8.61%	10.54%	14.06%	16.34%	16.70%
0.55	3.16%	3.34%	3.87%	4.39%	5.62%	5.98%	7.91%	10.72%	13.88%	14.06%
0.65	3.34%	3.87%	3.16%	3.51%	4.22%	4.75%	5.98%	8.79%	11.60%	11.95%
0.75	3.87%	3.51%	3.69%	3.34%	3.16%	4.22%	4.57%	6.15%	8.79%	10.37%
0.85	4.75%	4.22%	4.04%	3.69%	3.69%	3.34%	3.51%	4.22%	5.10%	6.85%
0.95	6.50%	5.27%	4.92%	4.75%	4.39%	4.57%	4.04%	3.51%	2.99%	5.27%

(c) learning error (detailed)

β_1 \ β_2	0.75	0.76	0.77	0.78	0.79	0.80
0.90	3.03%	3.05%	3.22%	3.38%	3.55%	3.69%
0.91	2.95%	3.01%	3.07%	3.16%	3.22%	3.38%
0.92	2.83%	2.83%	2.91%	2.93%	2.99%	3.10%
0.93	2.79%	2.69%	2.69%	2.73%	2.83%	2.99%
0.94	2.79%	2.71%	2.71%	2.71%	2.73%	2.83%
0.95	3.01%	2.89%	2.89%	2.87%	2.77%	2.79%

(d) testing error (detailed)

β_1 \ β_2	0.75	0.76	0.77	0.78	0.79	0.80
0.90	3.51%	3.34%	3.51%	3.87%	3.87%	4.04%
0.91	3.34%	3.34%	3.16%	3.34%	3.51%	3.69%
0.92	2.99%	2.99%	2.81%	2.99%	2.99%	3.34%
0.93	2.99%	2.64%	2.64%	2.81%	2.81%	3.16%
0.94	3.34%	3.16%	2.99%	3.16%	3.34%	3.51%
0.95	3.51%	3.34%	3.34%	3.16%	3.16%	3.16%

(iii) Fisher's Discriminant Analysis†

model type	(a) learning error	(b) testing error
Linear	3.14%	4.22%
Quadratic	2.68%	4.39%
Mahalanobis	9.86%	12.83%

† The function 'classify' in MATLAB Statistics toolbox is used.

(1993) in far less computation time. Comparative study with the other methods including some deterministic ones such as the effective independence distribution (EID) method proposed in Poston et al. (1997) is in progress and will be reported elsewhere in the future.

The second application of the proposed ellipsoid is multiclass discrimination. Since the covariance matrix and the mean vector are characterized as a solution of the maximum likelihood estimator of the multidimensional normal distribution, we also generalize the Fisher's discriminant methods through a parameterization with β. From preliminary computational experiments, we see that the proposed methods can achieve better predictive accuracy on the class index of unseen data. Further elaborative experiments remain to be done as a future work as well as analysis on statistical properties of the proposed methods.

Acknowledgments

Research of the first author is supported by MEXT Grant-in-Aid for Young Scientists (B) 17710125. Research of the second author is supported by MEXT Grant-in-Aid for Young Scientists (B) 16710110.

Appendix A

A.1 Proof of Proposition 1

With fixed $(\boldsymbol{Q}, \boldsymbol{\gamma})$, we sort the ellipsoidal scores $f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) := \|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2$, $i \in I$, in ascending order. If ℓ different data points $\boldsymbol{x}^{j_1}, ..., \boldsymbol{x}^{j_\ell}$ have the same score, say, $f^i(\boldsymbol{Q}, \boldsymbol{\gamma})$, we consider those ℓ points as a single point $\boldsymbol{x}^i$ and assign the value of $\frac{\ell}{m}$ to it as its empirical probability p^i instead of $\frac{1}{m}$ to each point. Then, we denote the sorted scores as $g^1(\boldsymbol{Q}, \boldsymbol{\gamma}) < \ldots < g^{m'}(\boldsymbol{Q}, \boldsymbol{\gamma})$, $m' \leq m$, with the underlying probability p^i, $i \in I' := \{1, \ldots, m'\}$. Proposition 8 of Rockafellar and Uryasev (2002) evaluates the β-quantile (VaR) of $g^i(\boldsymbol{Q}, \boldsymbol{\gamma}), i \in I'$, as $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) = g^K(\boldsymbol{Q}, \boldsymbol{\gamma}) = \|\boldsymbol{Q}\boldsymbol{x}^K - \boldsymbol{\gamma}\|^2$, where K is the unique index such that $\sum_{i=1}^{K} p^i \geq \beta > \sum_{i=1}^{K-1} p^i$. Hence, $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) = \phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) = \max_{i \in I} f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) = g^{m'}(\boldsymbol{Q}, \boldsymbol{\gamma})$ holds for $\beta > 1 - 1/m \geq \sum_{i=1}^{m'-1} p^i$, and the constraint $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n$ of (6) can be replaced by $f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n$ for all $i \in I$. For the case of $\beta = 0$, one has $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) = \min_{\alpha} \{\frac{1}{m} \sum_{i \in I} \max\{\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2, \alpha\}\} = \frac{1}{m} \sum_{i \in I} \|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2$. □

A.2 Proof of Theorem 2

At first, we show that Problem (11) has an optimal solution. Since the equivalence between (11) and (12) is obvious, it suffices to show that (12) has an optimal solution. The Lagrangian dual of (12) is given as

$$\left| \begin{array}{ll} \underset{\boldsymbol{\lambda},\, \eta > 0}{\text{maximize}} & \frac{n}{2} + \frac{1}{2} \ln \det \left[2(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{1}{\boldsymbol{e}^\top \boldsymbol{\lambda}} \boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top \boldsymbol{X}^\top) \right] - n\eta \\ \text{subject to} & \boldsymbol{e}^\top \boldsymbol{\lambda} = \eta,\ \boldsymbol{0} \leq \boldsymbol{\lambda} \leq \frac{\eta}{(1-\beta)m} \boldsymbol{e}. \end{array} \right.$$

By replacing $\boldsymbol{\lambda}/\eta$ by $\tilde{\boldsymbol{\lambda}}$, the dual turns out to be

$$\left| \begin{array}{ll} \underset{\boldsymbol{\lambda},\, \eta > 0}{\text{maximize}} & \frac{n}{2} + \frac{1}{2} \ln \det \left[\boldsymbol{X}\tilde{\boldsymbol{\Lambda}}\boldsymbol{X}^\top - \boldsymbol{X}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{X}^\top \right] - n\eta + \frac{1}{2} \ln (2\eta)^n \\ \text{subject to} & \boldsymbol{e}^\top \tilde{\boldsymbol{\lambda}} = 1,\ \boldsymbol{0} \leq \tilde{\boldsymbol{\lambda}} \leq \frac{1}{(1-\beta)m} \boldsymbol{e}. \end{array} \right.$$

This problem is optimized with $\eta = \frac{1}{2}$, and one reaches

$$\left| \begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & \frac{1}{2} \ln \det \left[\boldsymbol{X}\tilde{\boldsymbol{\Lambda}}\boldsymbol{X}^\top - \boldsymbol{X}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^\top \boldsymbol{X}^\top \right] \\ \text{subject to} & \boldsymbol{e}^\top \tilde{\boldsymbol{\lambda}} = 1,\ \boldsymbol{0} \leq \tilde{\boldsymbol{\lambda}} \leq \frac{1}{(1-\beta)m} \boldsymbol{e}, \end{array} \right. \tag{32}$$

with corresponding primal solution

$$\boldsymbol{Q} = (\boldsymbol{X}\tilde{\boldsymbol{\Lambda}}\boldsymbol{X}^\top - \boldsymbol{X}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^\top\boldsymbol{X}^\top)^{-1/2}; \quad \boldsymbol{\gamma} = \sum_{i\in I}\tilde{\lambda}_i\boldsymbol{Q}\boldsymbol{x}^i. \tag{33}$$

We observe that the dual (32) has a feasible solution $\tilde{\boldsymbol{\lambda}} = \boldsymbol{e}/m$ with finite objective value, *i.e.*, $\det[\boldsymbol{X}\tilde{\boldsymbol{\Lambda}}\boldsymbol{X}^\top - \boldsymbol{X}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^\top\boldsymbol{X}^\top] > 0$ under Assumption 1, so it has a finite optimal solution. By noting that the complementarity condition is fulfilled, (33) is a solution of Problem (12).

Next we show that $F_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha) = n$ holds at optimality. Let $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*)$ be an optimal solution of (11), and let

$$U^* := F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*) = \alpha^* + \frac{1}{(1-\beta)m}\sum_{i\in I}\left[\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2 - \alpha^*\right]^+.$$

Now we show $U^* > 0$ for any $\beta \in [0, 1)$. Note that the strict inequality $\sum_{i\in I}\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2 > 0$ holds, since assuming on the contrary that $\sum_{i\in I}\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2 = 0$, we have $\boldsymbol{x}^i = (\boldsymbol{Q}^*)^{-1}\boldsymbol{\gamma}^*$ for all $i \in I$, which contradicts Assumption 1. Therefore, we see that when $\beta = 0$, $U^* = \frac{1}{m}\sum_{i\in I}\max\{\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2, \alpha^*\}$ is positive. When $\beta > 0$, $\alpha^* \geq 0$ follows and hence, $U^* > 0$ is shown. Indeed, assuming on the contrary that $\alpha^* < 0$, the constraint $F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*) \leq n$ of (11) is expressed as

$$\frac{1}{(1-\beta)m}\sum_{i\in I}\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2 \leq n - (1 - \frac{1}{1-\beta})\alpha^*$$

and one then finds a feasible solution $(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha) = \{\frac{n}{n-(1-\frac{1}{1-\beta})\alpha^*}\}^{1/2}(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, 0)$ with smaller objective value, which contradicts the optimality of $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*)$. The strict inequalities $0 < n-(1-\frac{1}{1-\beta})\alpha^* < n$ are ensured since $\sum_{i\in I}\left\|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\right\|^2 > 0$ and $(1 - \frac{1}{1-\beta})\alpha^* > 0$.

Suppose that the inequality constraint of (11) is not binding, *i.e.*, $U^* < n$. Then, one finds a better feasible solution $((\frac{n}{U^*})^{1/2}\boldsymbol{Q}^*, (\frac{n}{U^*})^{1/2}\boldsymbol{\gamma}^*, (\frac{n}{U^*})\alpha^*)$ with the objective value $(-\ln\det[\boldsymbol{Q}^*] + n/2\ln(\frac{U^*}{n})) < -\ln\det[\boldsymbol{Q}^*]$, which contradicts the optimality of $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$. □

A.3 Proof of Proposition 4

We note that Problem (29) is equivalent to the following minimization problem:

$$\left|\begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma},\alpha,\boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] + \frac{1}{2}\{\alpha + \frac{\boldsymbol{e}^\top\boldsymbol{z}}{(1-\beta)m}\} \\ \text{subject to} & z_i \geq \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 - \alpha, \quad (i \in I), \\ & \boldsymbol{z} \geq \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}. \end{array}\right. \tag{34}$$

The Lagrangian dual of (34) becomes

$$\left|\begin{array}{ll} \underset{\boldsymbol{\lambda}}{\text{maximize}} & \frac{n}{2} + \frac{1}{2}\ln\det\begin{bmatrix} \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top & \boldsymbol{X}\boldsymbol{\lambda} \\ \boldsymbol{\lambda}^\top\boldsymbol{X}^\top & 1 \end{bmatrix} \\ \text{subject to} & \boldsymbol{e}^\top\boldsymbol{\lambda} = 1, \ \boldsymbol{0} \leq \boldsymbol{\lambda} \leq \frac{1}{(1-\beta)m}\boldsymbol{e}, \end{array}\right. \tag{35}$$

and the corresponding primal solution is given by

$$\boldsymbol{Q} = (\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top)^{-1/2}; \quad \boldsymbol{\gamma} = \sum_{i\in I}\lambda_i\boldsymbol{Q}\boldsymbol{x}^i,$$

which is in common with the Lagrangian dual (32) of (12) as shown in the proof of Theorem 2. □

References

Barnes, E.R. (1982), "An Algorithm for Separating Patterns by Ellipsoids," *IBM Journal of Research and Development*, vol.26, No.6, 759–764.

Ben-Tal, A. and Nemirovski, A. (2001), *Lectures on Modern Convex Optimization*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, USA.

Blake, C.L. and Merz, C.J. (1998), UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press, Cambridge, UK.

Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters*, 16, 213-218.

Gärtner, B. and Schönherr, S. (1997), "Smallest Enclosing Ellipses –Fast and Exact–," *Proc. 13. Annual ACM Symposium on Computational Geometry*, 430–432.

Hawkins, D.M. (1993), "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics*, 8, 95-107.

Khachiyan, L. and Todd, M. (1993), "On the Complexity of Approximating the Maximal Inscribed Ellipsoid for a Polytope," *Mathematical Programming*, 61, 137–159.

Konno, H., Gotoh, J., Uryasev, S. and Yuki,A. (2002), "Failure Discrimination by Semi-Definite Programming," *Financial Engineering, Supply Chain and E-commerce*, edited by P.Pardalos and V. Tsitsiringos, Kluwer Academic Publisher, Netherland.

Larsen, N., Mausser H., and Uryasev, S. (2002), "Algorithms for Optimization of Value-at-Risk," P. Pardalos and V.K. Tsitsiringos, (Eds.) *Financial Engineering, e-Commerce and Supply Chain*, Kluwer Academic Publishers, 129-157.

Mangasarian, O.L., Street, W.N., and Wolberg, W.H. (1995), "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, 43, 570-577.

Poston, W.L., Wegman, E.J., Priebe, C.E. and Solka, J.L. (1997), "A deterministic method for robust estimation of multivariate location and shape," *Journal of Computational and Graphical Statistics*, 6, 300-313.

Rockafellar, T.R. and Uryasev, S. (2002), "Conditional Value-at-Risk for General Loss Distributions," *Journal of Banking and Finance*, 26, 1443-1471.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, USA.

Sun, P. and Freund, R.M. (2004), "Computation of Minimum Volume Covering Ellipsoids," *Operations Research* 52, no.5, 690-706.

Titterington, D.M. (1975), "Optimal Design: Some Geometrical Aspects of D-Optimality," *Biometrika*, 62, 313-320.

Toh, K., Todd, M. and Tütüncü, R. (1999), "Sdpt3 – a matlab software package for semidefinite programming," *Optimization Methods and Software*, 11, 545-581.

Welzl, E. (1991), "Smallest Enclosing Disks (Balls and Ellipsoids)," In H.Maurer, editor, *New Results and New Trends in Computer Science*, vol.555 of Lecture Notes in Computer Science, 359–370, Springer-Verlag, Berlin.

Woodruff, D.L. and Rocke, D.M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, 2, 69-95.

Zhang, Y. and Gao, L. (2003), "On Numerical Solution of the Maximum Volume Ellipsoid Problem," *SIAM Journal of Optimization*, 14, 1, 53–76.