Department of Social Systems and Management

Discussion Paper Series

No. 1157

Conditional Minimum Volume Ellipsoid with Application to Multiclass Discrimination

by

Jun-ya Gotoh and Akiko Takeda

Original: January 2006
Revised: September 2006

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Conditional Minimum Volume Ellipsoid with Application to 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

Original: Jan.2006; Revised: Jul.2006

Abstract

In this paper, we present a new formulation for constructing an n-dimensional ellipsoid by generalizing the computation of the minimum volume covering ellipsoid. The proposed ellipsoid construction is associated with a user-defined parameter $\beta \in [0, 1)$, and formulated as a convex optimization based on the CVaR minimization technique proposed by Rockafellar and Uryasev [15]. An interior point algorithm for the solution is developed by modifying the DRN algorithm of Sun and Freund [19] for the minimum volume covering ellipsoid. By exploiting the solution structure, the associated parametric computation can be performed in an efficient manner. Also, the maximization of the normal likelihood function can be characterized in the context of the proposed ellipsoid construction, and the likelihood maximization can be generalized with parameter β. Motivated by this fact, the new ellipsoid construction is examined through a multiclass discrimination problem. Numerical results are given, showing the nice computational efficiency of the 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 discrimination, interior point algorithm

1 Introduction

In various contexts concerned with statistics and multivariate analysis, there are several important ways for constructing ellipsoids with a finite set of data points scattered in the n-dimensional real space. An example of such ellipsoids is called the *minimum volume covering ellipsoid*, which is defined as an ellipsoid having the minimum volume and covering all the given points. A straightforward way of its use in data analysis is to detect outliers out of the given data set, based on whether a point is on the boundary of the ellipsoid (see, e.g., [17]). Moreover, this ellipsoid construction is found useful in various contexts such as experimental design (e.g., [4, 20]) and computational geometry (e.g., [2]), for example. Another important aspect of this minimum volume covering ellipsoid is its high accessibility to the solution methods for the construction.

In fact, its optimization formulation has a convex structure, and numerous algorithms have been developed. Among such algorithms are those of Titterington [20], Barnes [1], Khachiyan and Todd [10], Sun and Freund [19], Zhang and Gao [24], Welzl [22], and Gärtner and Schönherr [7].

A generalized version of the minimum volume covering ellipsoid is known as the *minimum volume ellipsoid* with parameter $\beta \in (0, 1]$, denoted by β-MVE, and defined as an ellipsoid satisfying the following conditions:

(i) for a designated value $\beta \in (0, 1]$, it contains (at least) 100β percent of m points in $\mathbb{R}^n$;

(ii) it attains the minimal volume.

Obviously, when β is set to be in $(1-\frac{1}{m}, 1]$, the β-MVE is equal to the minimum volume covering ellipsoid. For adequate β, the center and the matrix determining the shape of the β-MVE provide estimators less affected by outliers, on the location of the data cloud and the related scatter matrix, respectively. Especially for $\beta > \frac{1}{2}$, the center and the matrix of the β-MVE are called $100(1-\beta)$ percent breakdown estimators, and $1-\beta$ is essentially considered to be its breakdown value, which denotes the fraction of outliers admissible for bounded estimators (see [16, 17] for details). Except the case where $\beta \in (1 - 1/m, 1]$, the computation of the β-MVE results in a nonconvex optimization problem because of a combinatorial constraint corresponding to the condition (i) above. Though many algorithms for approaching a β-MVE have been developed, most of researches including [23] have applied heuristic algorithms for solving this problem since enumeration algorithms such as in [6] are computationally impractical. Hawkins [8], for example, proposes a two-phase framework for obtaining an ellipsoid which satisfies a necessary condition to be the β-MVE. While 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.

Associated with another important ellipsoid construction is a parameter estimation of elliptical distributions, which are characterized by simultaneous density functions of the form $p(\boldsymbol{x}) := \kappa \det[\boldsymbol{Q}]\, q(\|\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{\gamma}\|^2)$, where $\kappa > 0$ is a constant, q is a function on $\mathbb{R}$, and 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 estimates $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$ can be obtained explicitly by using the covariance matrix and the mean vector, respectively, of the given data points.

In this paper, we propose another ellipsoid construction by generalizing the computation of the minimum volume covering ellipsoid. The resulting ellipsoid can be obtained by solving a convex minimization problem where its geometrical interpretation can be given in the context of the conditional value-at-risk (CVaR) minimization technique developed by Rockafellar and Uryasev [15]. Contrary to the β-MVE, which also contains the minimum volume covering ellipsoid as a special case as mentioned above, the new generalization can be achieved via a convex optimization which also has a user-defined parameter $\beta \in [0, 1)$. In addition, we show that the formulation with $\beta = 0$ coincides with the maximization of the normal likelihood function, and it thus provides a geometrical interpretation of the maximum likelihood estimation under the normality assumption. Conversely, a generalization of the maximum likelihood can be considered by extending the proposed ellipsoid construction into the context of the likelihood

maximization, and such a generalization can be applied to analyses where the normality has played an important role so far. In the latter part of this paper, we consider to apply the generalized likelihood maximization to a discriminant analysis. Besides, the formulation is a tractable convex problem that existing algorithmic techniques can solve. In this paper, we propose an interior point algorithm for computing the proposed ellipsoid by modifying the dual reduced Newton (DRN) algorithm developed by Sun and Freund [19]. A parametric computation of the β-CMVE for various values of βs can be efficiently performed by exploiting an optimal solution of the previously solved problem. Indeed, numerical results show that the parametric computation of the ellipsoids with ten βs can reduce the total computation time approximately by a factor of 4.

The structure of this paper is as follows. In Section 2, we provide a new formulation for constructing a minimum ellipsoid, and describe its relations to several ways for constructing an ellipsoid such as the normal likelihood maximization, another generalization of the minimum covering ellipsoid in [19] and the β-MVE. Section 3 is devoted to describing an interior point algorithm for solving the proposed optimization problem by modifying an efficient algorithm proposed by Sun and Freund [19] for solving the minimum volume covering ellipsoid problem. Some techniques for accelerating the associated parametric optimization are also proposed. In Section 4, the generalized ellipsoid computation is employed in a multiclass discrimination so as to examine the potential of the generalization. Numerical results show that the proposed method can improve the Fisher's discriminant analysis, and sometimes outperforms the one-against-one ν-SVM approach implemented in LIBSVM [5]. Finally, we conclude the paper with some remarks.

2 Formulation of the Conditional Minimum Volume Ellipsoid

In this section, we first summarize the preliminary properties of the n-dimensional ellipsoid and the formulation of the minimum volume covering ellipsoid problem. Secondly, we introduce a generalized minimum volume ellipsoid by extending the optimization formulation of the usual minimum volume covering ellipsoid. We then describe relations of the proposed formulation to several problems of constructing an ellipsoid.

2.1 Preliminary

Let $\{ \boldsymbol{x}^1, ..., \boldsymbol{x}^m \}$ be a given set of points in n-dimensional Euclidean space, and let I denote its index set, i.e., $I = \{ 1, ..., m \}$. As in [19], we suppose the following assumption throughout the paper.

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

Without loss of generality, ellipsoids in $\mathbb{R}^n$ are given in the following form:

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

where $\boldsymbol{Q}$ is an $n \times n$ real symmetric positive definite matrix, and $\boldsymbol{\gamma}$ is a vector in $\mathbb{R}^n$. We note that the (rightmost) constant n in (1) can be replaced by any positive number, but for simplicity,

we here adopt n. Also, we refer to the set of points satisfying equality in (1) as the boundary of the ellipsoid.

Clearly, there is a one-to-one correspondence between $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ and another ellipsoid of the form

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

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, the center of the ellipsoid, and the positive definite matrix $\boldsymbol{D}$ represents the covariate structure. The volume of these ellipsoids is then given by

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

where Γ is the gamma function (see, e.g., [19]).

Based on these notations, the minimum volume covering ellipsoid is computed by solving the following convex optimization:

$$\left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 \leq n, \quad (i \in I), \\ & \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{3}$$

The objective of this optimization implies to minimize the volume of the n-dimensional ellipsoid, while the m nonlinear inequality constraints impose that all of the data points are included in the ellipsoid.

2.2 Formulation of the Conditional Minimum Volume Ellipsoid and Its Geometric Interpretation

The ellipsoid proposed in this paper is defined as $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ with an optimal solution $(\boldsymbol{Q}, \boldsymbol{\gamma})$ of a nonlinear, but convex optimization problem formulated as follows:

$$(\mathrm{CMVE}(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & F_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha) \leq n, \\ & \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{4}$$

where

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

and $\beta \in [0, 1)$ is a user-defined constant, and $[w]^+ := \max\{w, 0\}$. It is clear that Problem (4) can be rewritten as the following convex problem:

$$(\mathrm{CMVE}(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha, \boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \alpha + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} \leq n, \\ & z_i \geq \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2 - \alpha, \quad (i \in I), \\ & \boldsymbol{z} \geq \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{5}$$

where $\boldsymbol{e} = (1, \ldots, 1)^\top$ denotes the vector consisting of ones. In this paper, the ellipsoid $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ obtained via Problem (4) or (5) is referred as the *β-conditional minimum volume ellipsoid*, and denoted by β-CMVE.

In the following, we will see the geometric interpretation of these formulations. For given $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$, let $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) := \min_\alpha \{F_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha)\}$, and let us introduce an optimization formulation as follows:

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

Theorem 16 of Rockafellar and Uryasev [15] ensures that under the existence of a solution, Problem (6) is equivalent to Problem (4) in the sense that $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*)$ solves (4) 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 (4) and, accordingly, the equivalence between (4) and (6) are ensured via the following theorem, whose proof is shown in Appendix A.

Theorem 2.1 *Suppose that Assumption 1 holds. Problem (4) has a solution, and the inequality constraint $F_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}, \alpha) \leq n$ of (4) is satisfied with equality at optimality.*

In order to interpret the meaning of (4) or (5) through the equivalence with (6), we below state the geometrical meaning of the function $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$. First of all, let us define the *ellipsoidal score* of data point i with respect to $E(\boldsymbol{Q}, \boldsymbol{\gamma})$ by

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

Let us denote the empirical distribution function of the score $f(\boldsymbol{x}|\boldsymbol{Q}, \boldsymbol{\gamma})$ by $\Phi(\,\alpha \,|\, \boldsymbol{Q}, \boldsymbol{\gamma})$, i.e.,

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

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

$$\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) \quad := \min \{\, \alpha \geq 0 \,:\, \Phi(\,\alpha \,|\, \boldsymbol{Q}, \boldsymbol{\gamma}) \geq \beta \,\}.$$

It should be noted that $\alpha_0(\boldsymbol{Q}, \boldsymbol{\gamma})$ is well defined by this definition since $f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \geq 0$, $i \in I$ for any $(\boldsymbol{Q}, \boldsymbol{\gamma})$.

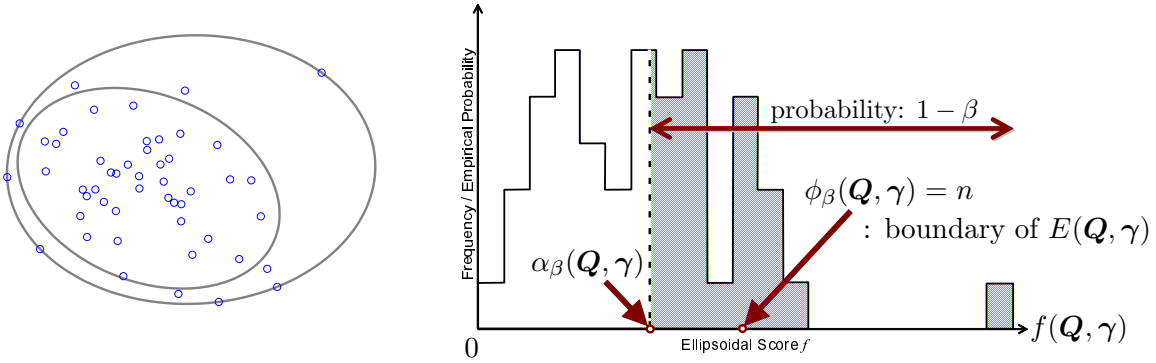
According to Rockafellar and Uryasev [15], $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$ is then shown to be equal to the mean of the ellipsoidal score under the β-tail distribution $\Phi_\beta(\eta|\boldsymbol{Q}, \boldsymbol{\gamma})$, which is defined by

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

More intuitive interpretation of ϕ_β is given by [15] as

$$0 \;\leq\; \alpha_\beta \;\leq\; \mathrm{E}\,[\, f \,|\, f \geq \alpha_\beta \,] \;\leq\; \phi_\beta \;\leq\; \mathrm{E}\,[\, f \,|\, f > \alpha_\beta \,], \tag{7}$$

where $\mathrm{E}[\cdot]$ denotes mathematical expectation operator under Φ, and $(\boldsymbol{Q}, \boldsymbol{\gamma})$ is omitted in (7) for notational simplicity. From these facts, we can see that the quantity ϕ_β is approximately equal



(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

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. 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.

Further, optimal α^* gives an approximate of the β-quantile α_β of the optimal distribution of the ellipsoidal score. More specifically, by combining Theorem 2.1 and the result of Rockafellar and Uryasev [15], $F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*) = \phi_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*) = n$ holds at optimality. These equalities imply that α^* is in $\mathrm{argmin}_\alpha F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha)$, and we see that α^* gives an approximate value of the β-quantile $\alpha_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)$ for $\beta \in (0,1)$ since $\mathrm{argmin}_\alpha F_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha)$ is shown to be identical to the closed interval $[\alpha_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*), \alpha^+_\beta(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*)]$ where $\alpha^+_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) := \inf\{\alpha \geq 0 : \Phi(\alpha|\boldsymbol{Q}, \boldsymbol{\gamma}) > \beta\}$. For additional properties from optimization viewpoints, readers are referred to [15].

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 this figure, the score on the boundary of the ellipsoid corresponds to $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$, which is approximately equal to the expected value of the ellipsoidal scores larger than the β-quantile $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$, as mentioned.

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

Proposition 2.2 *For $\beta > 1 - \frac{1}{m}$, Problem (6) is equivalent to the minimum volume covering ellipsoid problem formulated as Problem (3). For $\beta = 0$, Problem (6) is equivalent to the*

following problem:

$$\left|\begin{array}{ll} \underset{\boldsymbol{Q},\boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \dfrac{1}{m}\displaystyle\sum_{i\in I}\left\|\boldsymbol{Q}\boldsymbol{x}^i-\boldsymbol{\gamma}\right\|^2 \le n, \\ & \boldsymbol{Q}\succ\boldsymbol{O}, \end{array}\right. \tag{8}$$

and has the unique solution $(\boldsymbol{Q}^*,\boldsymbol{\gamma}^*)$ *defined via covariance matrix and mean vector as*

$$\boldsymbol{Q}^*=\left(\frac{1}{m}\sum_{i\in I}(\boldsymbol{x}^i-\bar{\boldsymbol{x}})(\boldsymbol{x}^i-\bar{\boldsymbol{x}})^\top\right)^{-1/2};\quad \boldsymbol{\gamma}^*=\boldsymbol{Q}^*\bar{\boldsymbol{x}},\quad \textit{where}\quad \bar{\boldsymbol{x}}:=\frac{1}{m}\sum_{i\in I}\boldsymbol{x}^i. \tag{9}$$

The first statement of Proposition 2.2 implies that the β-CMVE is a generalization of the minimum volume covering ellipsoid since the former with β sufficiently close to 1 is proved to be equivalent to the latter. On the other hand, the second statement indicates that the β-CMVE generalizes another interesting ellipsoid. Since the left-hand side of the inequality constraint in (8) means the mean ellipsoidal score of all the points, we can see 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. Furthermore, it is worth noting that the solution defined by (9) implies that constructing the β-CMVE with $\beta=0$ is equivalent to maximizing the likelihood function of the point set under the normality assumption since (9) exactly corresponds to the inverse of the sample covariance matrix and the mean vector through the one-to-one correspondence of two ellipsoids (1) and (2).

2.3 Relations to the Other Ellipsoid Constructions

2.3.1 Relation to Maximization of a Generalized Normal Log-Likelihood Function

The second statement of Proposition 2.2 shows that the construction of the β-CMVE also generalizes the maximization of the normal log-likelihood, which is defined with observations $\{\boldsymbol{x}^1,...,\boldsymbol{x}^m\}$ by

$$\ell(\boldsymbol{Q},\boldsymbol{\gamma})=m\ln\det[\boldsymbol{Q}]-\frac{1}{2}\sum_{i\in I}f^i(\boldsymbol{Q},\boldsymbol{\gamma}). \tag{10}$$

Associated with the maximization of the log-likelihood function (10), let us consider the following optimization problem:

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

or its equivalent differentiable optimization:

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

The difference between the function (10) and the objective of Problem (11) is found in their second terms. It is apparent that Problem (11) generalizes the maximization of the normal log-likelihood function (10) since (11) results in the latter by setting $\beta = 0$.

More specifically, we see that Problem (11) is equivalent to the maximization of a conditional likelihood defined by $\prod_{i \in I} \ell^i(\boldsymbol{Q}, \boldsymbol{\gamma})$ where

$$\ell^i(\boldsymbol{Q}, \boldsymbol{\gamma}) := \begin{cases} \det[\boldsymbol{Q}] \exp\{-\frac{1}{2}(\alpha + \frac{1}{1-\beta}(f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) - \alpha))\} & \text{for } i \text{ s.t. } f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) > \alpha \\ \det[\boldsymbol{Q}] \exp\{-\frac{1}{2}\alpha\} & \text{for } i \text{ s.t. } f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq \alpha. \end{cases}$$

Likelihood is assigned at each point dependently on whether $f^i(\boldsymbol{Q}, \boldsymbol{\gamma})$ is greater than α or not, and if $f^i(\boldsymbol{Q}, \boldsymbol{\gamma})$ is smaller than α, it does not contribute to the conditional likelihood.

More directly, the following proposition shows the equivalence between the generalized log-likelihood maximization (11) and Problem (CMVE(β)). The proof is provided in Appendix A.

Proposition 2.3 *Problem (11) and Problem (4) provide the same ellipsoid.*

From this proposition, optimal α approximates the β-quantile, and the optimal value of (11) can be interpreted as the maximized conditional likelihood to which only points with f^i larger than the β-quantile contribute. In this sense, Problem (CMVE(β)) can be regarded as the maximization of the generalized log-likelihood function (11). 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 (9).

2.3.2 Relation to MVCEP in Sun and Freund [19]

Another generalized formulation of the minimum volume covering ellipsoid (3) is described in Sun and Freund [19] as follows:

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

where $P > 0$ is a user-defined parameter. This formulation also intends to relax the dependency on the outlying data as well as Problem (CMVE(β)), and it can be expected that there is a one-to-one relation between (13) and our formulation. This intuition holds true as shown in the following proposition.

Proposition 2.4 *Let $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*, \boldsymbol{z}^*)$ be an optimal solution of Problem (12). When $\alpha^* > 0$, $(\frac{1}{\sqrt{\alpha^*}}\boldsymbol{Q}^*, \frac{1}{\sqrt{\alpha^*}}\boldsymbol{\gamma}^*, \frac{1}{\alpha^*}\boldsymbol{z}^*)$ is an optimal solution of Problem (13) with $P = \frac{\alpha^*}{2(1-\beta)m}$.*

This statement may remind readers of the relation between two formulations, called C-SVM and ν-SVM, for the support vector machines (SVMs) with different parameters C and ν (see,

e.g., [18]). For example, two formulations of the SVM for two-class linear classification can be written as follows:

$$\left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\boldsymbol{e}^\top \boldsymbol{z} \\ \text{subject to} & y^i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq 1 - z_i,\ z_i \geq 0 \quad (i \in I), \end{array} \right. \tag{14}$$

and

$$\left| \begin{array}{ll} \underset{\boldsymbol{w},\, b,\, \boldsymbol{z},\, \rho}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\boldsymbol{e}^\top \boldsymbol{z} \\ \text{subject to} & y^i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) \geq \rho - z_i,\ z_i \geq 0 \quad (i \in I),\ \rho \geq 0, \end{array} \right. \tag{15}$$

where $y^i \in \{-1, +1\}$ indicates which class a data i belongs to. In the context of the SVM, parameter C in (14) can take any positive value and the meaning of the value is vague, while parameter ν in (15) is in $(0, 1)$, and indicates the lower bound of the number of support vectors, which will be defined below.

It is apparent that in our formulation (5), the parameter $(1 - \beta)$ plays a similar role to the parameter ν in (15), whereas the parameter P in (13) corresponds to C in (14). This correspondence provides the CMVE with analogous properties of the SVMs parameterized with ν as in (15). To show the properties, let us define the term *support vectors* and *margin error* with an optimal solution $(\boldsymbol{Q}^*, \boldsymbol{\gamma}^*, \alpha^*, \boldsymbol{z}^*)$ of (CMVE(β)) and $\boldsymbol{\lambda}^*$ of its dual problem (35). The points $\boldsymbol{x}^i$ with $\lambda_i^* > 0$ are called support vectors, and let SV denote the index set of them. Also, the points with $z_i^* > 0$ are called margin errors, and let ERR denote the index set of them, i.e., $ERR := \{i \in I : z_i^* > 0\} = \{i \in I : \|\boldsymbol{Q}^*\boldsymbol{x}^i - \boldsymbol{\gamma}^*\|^2 > \alpha^*\}$.

Proposition 2.5 *An optimal ellipsoid constructed by solving Problem (CMVE(β)) has a β-property in the following sense:*

(i) $1 - \beta$ *is a lower bound on the fraction of support vectors, that is,* $1 - \beta \leq \frac{|SV|}{m}$,

(ii) $1 - \beta$ *is an upper bound on the fraction of margin errors, that is,* $1 - \beta \geq \frac{|ERR|}{m}$.

2.3.3 Relation to the β-MVE

As mentioned in Introduction, the β-MVE is an important ellipsoid providing robust estimates of the center of the data points and the scatter matrix. This ellipsoid construction is characterized by an optimal solution of the following optimization problem with a combinatorial constraint:

$$(\text{MVE}(\beta)) \quad \left| \begin{array}{ll} \underset{\boldsymbol{Q},\, \boldsymbol{\gamma}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & \left|\left\{ i \in I \,:\, \|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\|^2 \leq n \right\}\right| \geq \lceil \beta m \rceil, \\ & \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{16}$$

Here the combinatorial constraint defines a nonconvex feasible region, and this optimization problem is thus a nonconvex optimization problem, in general.

In order to clarify the relation between the formulations (16) and (CMVE(β)) in the minimal ellipsoid context, we observe the following lemma:

Lemma 2.6 $\left\{ (\boldsymbol{Q}, \boldsymbol{\gamma}) \,:\, \left| \{ i \in I \,:\, f^i(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n \} \right| \geq \lceil \beta m \rceil \right\} = \{ (\boldsymbol{Q}, \boldsymbol{\gamma}) \,:\, \alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma}) \leq n \}.$

This equivalence is straightforward from the definition of $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$. Hence, Problem (MVE($\beta$)) is equivalently rewritten as follows:

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

From this formulation and the interpretation of the β-CMVE, the computation of the β-MVE can be considered as the construction of a minimum volume ellipsoid whose boundary corresponds to the 100β percentile in the ellipsoidal score. As indicated in (7), $\phi_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$ is an upper bound of $\alpha_\beta(\boldsymbol{Q}, \boldsymbol{\gamma})$, and for β close to 1, these two quantities take values close to each other. Therefore, when β is close to 1, solutions of the convex program (CMVE(β)) are expected to provide good approximate solutions of the nonconvex program (16).

3 Modification of the Dual Reduced Newton Algorithm

3.1 An Algorithm for Solving Problem (CMVE(β))

Problem (5) can be transformed into

$$(\mathrm{CMVE}(\beta)) \left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{z}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] \\ \text{subject to} & z_i \geq \left\| \boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma} \right\|^2 + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} - n, \quad (i \in I), \\ & \boldsymbol{z} \geq \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \end{array} \right. \tag{17}$$

by deleting α from (5) via $\alpha = n - \frac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m}$, which is implied by Theorem 2.1. Sun and Freund [19] have proposed the "dual reduced Newton algorithm" for solving Problem (3), while they also utilized SDPT3 solver (see [21]) 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 (17) in a similar manner to Sun and Freund [19]. In order to apply the method, we add a logarithmic barrier function to (17) and obtain the formulation

$$\left| \begin{array}{ll} \underset{\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{z}, \boldsymbol{t}}{\text{minimize}} & -\ln\det[\boldsymbol{Q}] - \theta_t \displaystyle\sum_{i \in I} \ln t_i - \theta_z \sum_{i \in I} \ln z_i \\ \text{subject to} & \left\| \boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma} \right\|^2 - z_i + \dfrac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \\ & \boldsymbol{z} > \boldsymbol{0}, \quad \boldsymbol{t} > \boldsymbol{0}, \quad \boldsymbol{Q} \succ \boldsymbol{O}. \end{array} \right. \tag{18}$$

We set positive values on the parameters θ_t and θ_z, and parameterized solutions to Problem (18) varying over $\theta_t \in (0, \infty)$ and $\theta_z \in (0, \infty)$ form the central trajectory of (18). We follow the central trajectory by reducing the parameters θ_t and θ_z down to 0. Introducing Lagrange

multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ for the equality constraints in (18), the optimality conditions of (18) are as follows:

$$\sum_{i \in I} \lambda_i \{(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})\boldsymbol{x}^{i\top} + \boldsymbol{x}^i(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})^\top\} = \boldsymbol{Q}^{-1}, \tag{19}$$

$$\sum_{i \in I} \lambda_i(\boldsymbol{\gamma} - \boldsymbol{Q}\boldsymbol{x}^i) = \boldsymbol{0}, \tag{20}$$

$$(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma})^\top(\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}) - z_i + \frac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \tag{21}$$

$$\boldsymbol{\Lambda}\boldsymbol{t} = \theta_t \boldsymbol{e}, \tag{22}$$

$$\left\{ \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{E} - \boldsymbol{\Lambda} \right\} \boldsymbol{z} = \theta_z \boldsymbol{e}, \tag{23}$$

$$\boldsymbol{z} \geq \boldsymbol{0}, \quad \boldsymbol{t} \geq \boldsymbol{0}, \quad \boldsymbol{0} \leq \boldsymbol{\lambda} \leq \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}, \quad \boldsymbol{Q} \succ \boldsymbol{O}, \tag{24}$$

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. Note that the above equations with $\theta_t = \theta_z = 0$ correspond to the optimality conditions of Problem (17).

Sun and Freund [19] have dealt with the minimum volume covering ellipsoid (3) and introduced a logarithmic barrier function concerning slack variables $t_i = n - \left\|\boldsymbol{Q}\boldsymbol{x}^i - \boldsymbol{\gamma}\right\|^2$ $(i \in I)$ into Problem (3) to apply a Newton method. They have shown the optimality conditions as (19) through (24) with $\boldsymbol{z} = \boldsymbol{0}$. Compared to their conditions, equations related to $\boldsymbol{z}$ such as (23) and $\boldsymbol{\lambda} \leq \frac{\boldsymbol{e}^\top \boldsymbol{\lambda}}{(1-\beta)m}\boldsymbol{e}$ are additionally introduced in our conditions. Sun and Freund [19] proved that the condition $\boldsymbol{\lambda} > \boldsymbol{0}$ together with Assumption 1 ensures positive definiteness of the matrix $\left(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}\right)$, and furthermore, the matrix $\boldsymbol{Q}$ and vector $\boldsymbol{\gamma}$ are described with $\boldsymbol{\lambda}$ from (19) and (20), respectively, as

$$\boldsymbol{Q} = \left[2\left(\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^\top - \frac{\boldsymbol{X}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top\boldsymbol{X}^\top}{\boldsymbol{e}^\top\boldsymbol{\lambda}}\right)\right]^{-1/2}; \quad \boldsymbol{\gamma} = \frac{\boldsymbol{Q}\boldsymbol{X}\boldsymbol{\lambda}}{\boldsymbol{e}^\top\boldsymbol{\lambda}}, \tag{25}$$

where $\boldsymbol{X} := [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m]$ denotes an $n \times m$ matrix which consists of a given set of vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m$. By using (25), $\boldsymbol{Q}$ and $\boldsymbol{\gamma}$ are deleted from the above optimality conditions, and (21) is rewritten as

$$h_i(\boldsymbol{\lambda}) - z_i + \frac{\boldsymbol{e}^\top \boldsymbol{z}}{(1-\beta)m} + t_i = n, \quad (i \in I), \tag{26}$$

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 (22), (23), (24) and (26) as the optimality conditions for Problem (18). Note that the equation (26) indicates the feasibility of Problem (18), while (22) and (23) correspond to complementarity conditions for optimality. At a feasible solution, we first compute Newton direction for the system of equalities (22), (23) and (26), and then compute the step-size so

that the resulting solution satisfies the inequalities (24). The Newton direction $(\Delta\boldsymbol{\lambda}, \Delta\boldsymbol{t}, \Delta\boldsymbol{z})$ for (22), (23) and (26) at a feasible solution $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{t}}, \bar{\boldsymbol{z}})$ is obtained by solving

$$\begin{cases} \nabla_{\boldsymbol{\lambda}} \boldsymbol{h}(\bar{\boldsymbol{\lambda}})\Delta\boldsymbol{\lambda} + \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\Delta\boldsymbol{z} + \Delta\boldsymbol{t} = \boldsymbol{r}_1 := n\boldsymbol{e} - \boldsymbol{h}(\bar{\boldsymbol{\lambda}}) - \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\bar{\boldsymbol{z}} - \bar{\boldsymbol{t}}, \\ \bar{\boldsymbol{\Lambda}}\Delta\boldsymbol{t} + \bar{\boldsymbol{T}}\Delta\boldsymbol{\lambda} = \boldsymbol{r}_2 := \theta_t\boldsymbol{e} - \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{t}}, \\ \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}\Delta\boldsymbol{z} + \bar{\boldsymbol{Z}}\left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\Delta\boldsymbol{\lambda} = \boldsymbol{r}_3 := \theta_z\boldsymbol{e} - \frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\bar{\boldsymbol{z}} + \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{z}}, \end{cases} \tag{27}$$

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. A simple expression of $\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}})$ is given in Proposition 5 of Sun and Freund [19] by

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

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 (27) lead to

$$\begin{cases} \Delta\boldsymbol{t} = \bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{r}_2 - \bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{T}}\Delta\boldsymbol{\lambda}, \\ \Delta\boldsymbol{z} = \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\boldsymbol{r}_3 - \left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\bar{\boldsymbol{Z}}\left(\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right)\Delta\boldsymbol{\lambda}, \end{cases} \tag{28}$$

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 (28), the first equality of (27) is transformed into

$$\Delta\boldsymbol{\lambda} = \boldsymbol{R}^{-1}\left[\boldsymbol{r}_1 - \left\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\right\}\left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\boldsymbol{r}_3 - \bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{r}_2\right], \tag{29}$$

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 $\boldsymbol{R}^{-1}$. For $\bar{\boldsymbol{\lambda}} > \boldsymbol{0}$ and $\bar{\boldsymbol{t}} > \boldsymbol{0}$, $(\nabla_{\boldsymbol{\lambda}}\boldsymbol{h}(\bar{\boldsymbol{\lambda}}) - \bar{\boldsymbol{\Lambda}}^{-1}\bar{\boldsymbol{T}}) \prec O$ is ensured by Corollary 6 of Sun and Freund [19]. Also,

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

is proved, since $\left\{\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m}\boldsymbol{E} - \bar{\boldsymbol{\Lambda}}\right\}^{-1}\bar{\boldsymbol{Z}}$ is a diagonal matrix with diagonal elements of $\bar{z}_i/(\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta)m} - \bar{\lambda}_i)$, $(i \in I)$. Therefore, we see that $\boldsymbol{R}$ is negative definite.

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) Let $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $\epsilon_3 > 0$. Choose initial values of $(\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\lambda})$ satisfying $\boldsymbol{z} \geq \boldsymbol{0}$, $(\boldsymbol{t}, \boldsymbol{\lambda}) > \boldsymbol{0}$ and $\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\}\boldsymbol{\lambda} > \boldsymbol{0}$.

Step 1: (Stopping Criteria) Compute $OBJ := -\ln\det[\boldsymbol{Q}]$ using (25). If the following inequalities

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

are satisfied, terminate the algorithm with $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{z})$.

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 (28) and (29).

Step 3: (Step-Size Computation) Compute

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

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.

When ϵ_1, ϵ_2 and ϵ_3 are sufficiently small, the solution $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{z})$ of Algorithm DRN can be regarded as an optimal solution of Problem (CMVE(β)) in (17).

3.2 Parametric Optimization Method

As will be exhibited in Section 4, when the conditional minimum volume ellipsoid is applied to discrimination problem, the choice of parameter β plays an important role in achieving high predictive accuracy. Hence, solving Problem (CMVE(β)) many times with different β, say, $\beta_1, ..., \beta_N$, is required so as to improve the discrimination results. To this end, we propose to parametrically solve Problem (CMVE(β)) in an efficient manner by utilizing a solution which is already obtained under different β. In what follows, we consider to solve Problem (CMVE(β)) parametrically for $N\beta$s satisfying $\beta_1 := 0.0 < \beta_2 < \cdots < \beta_N < 1$.

There are mainly two devices for speeding up the parametric computation. The first is the downsizing of Problem (CMVE(β_h)) by exploiting the optimality condition of Problem (CMVE(β_{h-1})). The second is the setting of an initial vector $(\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\lambda})$, which is necessary to start Algorithm DRN.

Let $(\boldsymbol{z}_h^*, \boldsymbol{t}_h^*, \boldsymbol{\lambda}_h^*)$ be an optimal solution of (CMVE(β_h)) and its dual problem (35). It should be noted that the unique optimal solution of Problem (CMVE(β_1)) with $\beta_1 = 0$ satisfies $(\boldsymbol{z}_1^*, \boldsymbol{t}_1^*, \boldsymbol{\lambda}_1^*) = (\boldsymbol{h}(\boldsymbol{\lambda}_1^*), \boldsymbol{0}, \frac{1}{2m}\boldsymbol{e})$ as well as (9), which is ensured by comparing (5) and (8).

Downsizing the Problems When some components of $\boldsymbol{\lambda}_{h-1}^*$ take zero, so do the corresponding components of $\boldsymbol{z}_{h-1}^*$ because of the (complementarity) condition (23) with $\theta_z = 0$. We expect that the corresponding components of $\boldsymbol{z}_h^*$ and $\boldsymbol{\lambda}_h^*$ also result in taking zero at optimality since the feasible region of the dual (35) of Problem (CMVE(β_h)), which is given in Appendix, includes that of Problem (CMVE(β_{h-1})). Hence, the number of zeros in $\tilde{\boldsymbol{\lambda}}^*$, which is an optimal solution

of the dual (35), is expected to increase as β becomes larger since the sum of $\tilde{\lambda}_i$s is bounded above by one.

Let us suppose that some components of $\boldsymbol{\lambda}^*_{h-1}$ are zero. We then consider a subproblem of Problem (CMVE(β_h)) by removing such components from variables $\boldsymbol{z}$, $\boldsymbol{t}$ and $\boldsymbol{\lambda}$, and the corresponding data points $\boldsymbol{x}^i$ from $\boldsymbol{X}$ as long as Assumption 1 holds without $\boldsymbol{x}^i$. After solving the reduced version of Problem (CMVE(β_h)), we check whether $t_i = n - h_i(\boldsymbol{\lambda}^*_h) - \frac{\boldsymbol{e}^\top \boldsymbol{z}^*_h}{(1-\beta_h)m}$ is nonnegative or not for all $i \in I$ by using the obtained optimal solution $(\boldsymbol{z}^*_h, \boldsymbol{\lambda}^*_h)$. If $t_i \geq 0$ holds for all $i \in I$, the optimality condition of Problem (CMVE(β_h)) is satisfied and removed components of $\boldsymbol{z}$ and $\boldsymbol{\lambda}$ are proved to be zero. Otherwise, the removed components and data points $\boldsymbol{x}^i$ are added to Problem (CMVE(β_h)) and solve it again.

Now we make sure that the reduction of variables, constraints and data points never induce any troubles numerically during Algorithm DRN, especially in the computation of the Newton direction. More specifically, for the reduced data matrix $\boldsymbol{X}'$ and the corresponding dual variable $\boldsymbol{\lambda}'$, the positive definiteness of the matrix $\boldsymbol{Q} = \left(\boldsymbol{X}'\boldsymbol{\Lambda}'\boldsymbol{X}'^\top - \boldsymbol{X}'\boldsymbol{\lambda}'\boldsymbol{\lambda}'^\top \boldsymbol{X}'^\top/(\boldsymbol{e}^\top\boldsymbol{\lambda}')\right)$ is ensured for the sake of the deletion rule of data point, i.e., $\boldsymbol{x}^i$ is deleted only when Assumption 1 holds without $\boldsymbol{x}^i$. Also, the negative definiteness of the reduced matrix $\boldsymbol{R}'$ of $\boldsymbol{R}$ in (29) is proved as well as $\boldsymbol{R}$.

Initial Vector The Newton method presented above can be started from any solution $(\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\lambda})$ satisfying $\boldsymbol{z} \geq \boldsymbol{0}$, $(\boldsymbol{t}, \boldsymbol{\lambda}) > \boldsymbol{0}$ and $\{\frac{1}{(1-\beta)m}\boldsymbol{U} - \boldsymbol{E}\}\boldsymbol{\lambda} > \boldsymbol{0}$, but it is preferable to choose an initial vector which satisfies the feasibility of Problem (18), that is, the condition (26). Here we try to find a better initial vector of Algorithm DRN for solving Problem (CMVE(β_h)) by utilizing an optimal solution of Problem (CMVE(β_{h-1})).

By removing zero components from an optimal solution $\boldsymbol{\lambda}^*_{h-1}$ of (CMVE(β_{h-1})), we have a positive vector $\boldsymbol{\lambda}'_{h-1} > \boldsymbol{0}$. Similarly, removing the corresponding components from $\boldsymbol{z}^*_{h-1}$ and $\boldsymbol{t}^*_{k-1}$ leads to smaller sized vectors $\boldsymbol{z}'_{h-1}$ and $\boldsymbol{t}'_h$, respectively. In order to construct an initial vector for solving Problem (CMVE(β_h)), let us consider

$$(\bar{\boldsymbol{z}}, \bar{\boldsymbol{t}}, \bar{\boldsymbol{\lambda}}) = (\ \frac{1}{\gamma}\boldsymbol{z}'_{h-1},\ \frac{1}{\gamma}\boldsymbol{t}'_{h-1} + 0.05n\boldsymbol{e},\ \gamma\boldsymbol{\lambda}'_{h-1}\ ), \tag{30}$$

where $\gamma := \frac{1}{0.95n}\left\{n - \frac{\boldsymbol{e}^\top \boldsymbol{z}'_{h-1}}{m}\left(\frac{1}{1-\beta_{h-1}} - \frac{1}{1-\beta_h}\right)\right\}$. Note that $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{\lambda}}) > \boldsymbol{0}$ is clearly satisfied, and $\bar{\boldsymbol{\lambda}}$ is strictly less than $\frac{\boldsymbol{e}^\top\bar{\boldsymbol{\lambda}}}{(1-\beta_h)m}\boldsymbol{e}$ by construction. Moreover, the initial vector satisfies the feasibility (26) of Problem (18). Indeed, using the property $\boldsymbol{h}(\gamma\boldsymbol{\lambda}'_{h-1}) = \frac{1}{\gamma}\boldsymbol{h}(\boldsymbol{\lambda}'_{h-1})$, we have

$$\begin{aligned}
& \boldsymbol{h}(\bar{\boldsymbol{\lambda}}) - \bar{\boldsymbol{z}} + \tfrac{\boldsymbol{e}^\top\bar{\boldsymbol{z}}}{(1-\beta_h)m}\boldsymbol{e} + \bar{\boldsymbol{t}} \\
= & \ \tfrac{1}{\gamma}\left(\boldsymbol{h}(\boldsymbol{\lambda}'_{h-1}) - \boldsymbol{z}'_{h-1} + \tfrac{\boldsymbol{e}^\top\boldsymbol{z}'_{h-1}}{(1-\beta_{h-1})m}\boldsymbol{e} + \boldsymbol{t}'_{h-1}\right) - \tfrac{1}{\gamma}\left(\tfrac{\boldsymbol{e}^\top\boldsymbol{z}'_{h-1}}{(1-\beta_{h-1})m}\boldsymbol{e} - \tfrac{\boldsymbol{e}^\top\boldsymbol{z}'_{h-1}}{(1-\beta_h)m}\boldsymbol{e}\right) + 0.05n\boldsymbol{e} \\
= & \ \tfrac{1}{\gamma}\left\{n - \tfrac{\boldsymbol{e}^\top\boldsymbol{z}'_{h-1}}{m}\left(\tfrac{1}{1-\beta_{h-1}} - \tfrac{1}{1-\beta_h}\right)\right\}\boldsymbol{e} + 0.05n\boldsymbol{e} \\
= & \ n\boldsymbol{e}.
\end{aligned}$$

Though an initial vector of Problem (CMVE(β_h)) can also be obtained as

$$(\bar{\boldsymbol{z}}, \bar{\boldsymbol{t}}, \bar{\boldsymbol{\lambda}}) = (\ \frac{1}{\gamma}\boldsymbol{z}'_1,\ \frac{1}{\gamma}\boldsymbol{t}'_1 + 0.05n\boldsymbol{e},\ \gamma\boldsymbol{\lambda}'_1\ ) \tag{31}$$

Table 1: The Numbers of Iterations and Variables via the Parametric Computation Strategies (WDBC-Cancer $\langle 3 \rangle$ Data)

Data Set A: Benign Group of WDBC-Cancer Data ($m = 357$)									
h	2	3	4	5	6	7	8	9	10
β_h	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
#iter. from (31)	11	15	17	19	19	20	17	20	20
#iter. from (30)	11	10	9	9	9	9	9	10	10
#(data used)	357	322	286	250	216	180	143	110	76
Data Set B: Malignant Group of WDBC-Cancer Data ($m = 212$)									
h	2	3	4	5	6	7	8	9	10
β_h	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
#iter. from (31)	10	13	12	14	15	16	16	18	17
#iter. from (30)	10	9	9	9	9	9	9	9	10
#(data used)	212	191	170	150	128	108	88	65	43

by using the optimal solution $(\boldsymbol{z}_1^*, \boldsymbol{t}_1^*, \boldsymbol{\lambda}_1^*)$ of Problem (CMVE(β_1)) instead of the previous solution $(\boldsymbol{z}_{h-1}^*, \boldsymbol{t}_{h-1}^*, \boldsymbol{\lambda}_{h-1}^*)$, numerical experiments show that the use of the previous solution helps Algorithm DRN converge fast to an optimal solution. Table 1 shows the numbers of iterations and used data points when these parametric optimization techniques are applied to WDBC-Cancer data set of [3]. "WDBC-Cancer $\langle 3 \rangle$" implies the use of the three attributes (no.2, 24 and 25) with which Mangasarian et al. [13] attained about 2.5% cross-validation error rate by applying a linear programming discriminant model. In Table 1, two initial vector settings are compared by showing the number of iterations of the algorithm until the stopping criterion is satisfied. For example, for $\beta = 0.6$, the number of iterations starting from (30) is about half of that from (31). This table also shows the number of data points $\boldsymbol{x}^i$, $i \in I$, used in reduced version of Problem (CMVE(β_h)), and we see that the number of used data points and their corresponding variables $(\boldsymbol{z}', \boldsymbol{t}', \boldsymbol{\lambda}')$ decrease gradually as β becomes large, as expected. In fact, by combining the downsizing strategy, the total computational time for computing the ten problems with $\beta = 0, 0.1, ..., 0.9$ is reduced approximately by a factor of 4. Besides, through these numerical experiments, the downsizing strategy never led to any violation of the removed constraints, and accordingly, we did not need to solve Problem (CMVE(β_h)) by restoring the removed data $\boldsymbol{x}^i$, variables and constraints.

4 Multiclass Discrimination Based on the β-CMVE Computation

From Proposition 2.3, we see that the computation of the β-CMVE provides an alternative way for estimating normal densities by attaching weight to the outlying data points. In particular,

if the outlying data can be considered more important in capturing the shape of the data cloud, estimates obtained by solving Problem (CMVE(β)) may be useful. In order to examine this strategy, we here employ the β-CMVE construction to find a better decision rule for multiclass discrimination along the lines of Fisher's discriminant analysis.

Let K be the index set of classes, and suppose that each data point $\boldsymbol{x}^i$ belongs to a class in K. The index set I of training data is divided into $|K|$ classes such as $I = \cup_{k\in K} I_k$. The purpose of the multiclass discrimination here is to predict the labels of unknown data points, by exploiting the β-CMVE computation. More specifically, we first estimate $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ by solving Problem (5) with index set I_k for each class $k \in K$, and then, for a new sample $\bar{\boldsymbol{x}}$, we assign its class label $\bar{k}$ by one of three criteria described below.

The first criterion is a straightforward modification of the Bayesian decision rule under the normal distribution (see, e.g., [9]). After computing $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ for every $k \in K$, class label of a new sample $\bar{\boldsymbol{x}}$ is determined by

$$\bar{k} \in \arg\max_{k\in K} \left\{ \ln\det[\boldsymbol{Q}_k] - \frac{1}{2} f(\bar{\boldsymbol{x}}|\boldsymbol{Q}_k, \boldsymbol{\gamma}_k) + \ln m_k \right\}, \tag{32}$$

where m_k is the cardinality of training data which belongs to class $k \in K$, i.e., $m_k = |I_k|$.

As the second criterion, it is natural to consider the comparison of each normal likelihood, i.e.,

$$\bar{k} \in \arg\max_{k\in K} \left\{ \ln\det[\boldsymbol{Q}_k] - \frac{1}{2} f(\bar{\boldsymbol{x}}|\boldsymbol{Q}_k, \boldsymbol{\gamma}_k) \right\}. \tag{33}$$

The difference between the two criteria is the term $\ln m_k$, which corresponds to the prior probability information.

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

$$\bar{k} \in \arg\min_{k\in K} \{f(\bar{\boldsymbol{x}}|\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)\}. \tag{34}$$

The difference between the previous two criteria is the term $\ln\det[\boldsymbol{Q}_k]$.

Here, it should be noted that different βs can be applied to different classes since each estimate $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ depends on only one β, say, β_k. In this sense, the solution $(\boldsymbol{Q}_k, \boldsymbol{\gamma}_k)$ should be denoted as $(\boldsymbol{Q}_k(\beta_k), \boldsymbol{\gamma}_k(\beta_k))$, but we omit to denote its dependency on β_k for notational simplicity. Also, it is worth noting that the required number of times for solving Problem (CMVE(β)) is only $|K|N$ (not $N^{|K|}$) where N is the number of subdivisions of each β. 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. From this viewpoint, β-CMVE computation is preferable to the other multiclass classification approaches which require a considerable number of computations in the associated optimization problems. For example, the decision rule based on voting on all possible two-class classification results, such as the multiclass classification algorithm provided in LIBSVM [5], requires $\binom{K}{2}$ solutions of optimization problems for the voting.

If $\boldsymbol{Q}_k$ and $\boldsymbol{\gamma}_k$ are estimated by the unbiased covariance matrix and the mean vector as

$$\boldsymbol{Q}_k = \left(\frac{1}{m_k - 1} \sum_{i\in I_k} (\boldsymbol{x}^i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}^i - \bar{\boldsymbol{x}}_k)^\top \right)^{-\frac{1}{2}} ; \quad \boldsymbol{\gamma}_k = \boldsymbol{Q}_k \bar{\boldsymbol{x}}_k, \quad \text{where} \quad \bar{\boldsymbol{x}}_k := \frac{1}{m_k} \sum_{i\in I_k} \boldsymbol{x}^i,$$

Table 2: List of Data Sets

Name of Data Set	#Class	#Samples	#Attribute
Heart	2	270 (150, 120)	13
Liver-Disorder	2	345 (145, 200)	6
Pima-Indians-Diabetes	2	768 (500, 268)	8
WDBC-Cancer	2	569 (357, 212)	30
Iris	3	150 (50, 50, 50)	4
Wine	3	178 (59, 71, 48)	13
Vehicle	4	846 (218, 212, 217, 199)	18

the above three criteria are known as the Fisher's (quadratic) discriminant analysis (FDA). It should be noted that when $\beta = 0$, the above criteria (32), (33) and (34) with the β-CMVE are expected to be very similar to the FDA since solving Problem (CMVE(β)) with $\beta = 0$ is equivalent to the maximization of the normal likelihood as shown in Proposition 2.2. Further, the optimal $\boldsymbol{Q}_k$ and $\boldsymbol{\gamma}_k$ of (9) imply that the ellipsoids employed in the Fisher's analysis and the β-CMVE are parallel to each other, and their difference diminishes as the number of samples for learning increases.

In order to examine the potential of the proposed multiclass classification model, ten-fold cross-validation is carried out for several famous data sets which are obtained from the UCI repository of databases [3] and summarized in Table 2.

Table 3 shows the total testing (out-sample) error rates of the β-CMVE approaches, the FDAs and the one-against-one ν-SVM approach provided by LIBSVM [5]. The three different criteria: (a) the Bayes decision rule (32), (b) the normal likelihood rule (33) and (c) the modified Mahalanobis distance rule (34), are adopted for the β-CMVE and FDAs as mentioned above. The rates by each CMVE approach are the best results among all combinations of the β-CMVEs with (possibly different) eleven βs, $\beta = 0.0, 0.05, 0.15, 0.25, ..., 0.95$. On the other hand, for the ν-SVMs, linear and RBF (Radius Basis Function) kernels are adopted, and common eleven νs are applied to all classes, that is, $\nu = 0.0, 0.05, 0.15, 0.25, ..., 0.95$. For an additional parameter γ of ν-SVM with RBF kernel, different ten values $2^{-25}, 2^{-21}, \ldots, 2^{-7}$ are used.

From Table 3, we see that in many data sets the β-CMVE attains lower testing error rate than the FDAs by choosing adequate β. For example, for the Wine data, the three CMVE criteria found the testing error of zero, while the FDAs attain the positive error rates. In addition, the proposed approaches outperform the ν-SVM for some data sets.

These results can be explored in a more detailed manner by making the subdivision of the β finer. Tables 4 (i) to (iv) show the total learning and testing error rates via the Mahalanobis criterion when the WDBC cancer data is applied with only the three attributes (no. 2, 24 and 25) mentioned above. One of the authors applied an extended quadratic model of their linear model, and sees it difficult to outperform their model [11]. From Table 4 (iii)-(iv), we

Table 3: Testing Error Rates of 10-Fold Cross-Validation [%]

Testing (Out-sample) Error Rate

	CMVE			FDA			ν-SVM	
Data Set	(a)	(b)	(c)	(a)	(b)	(c)	linear	RBF
Heart	16.67	16.67	16.30	17.04	17.04	17.04	15.56	14.44
Liver-Disorder	33.04	32.75	29.86	42.03	42.61	29.86	38.50	24.92
Pima-Diabetes	23.70	23.44	22.92	26.30	27.08	32.90	23.96	22.66
WDBC-Cancer	4.04	4.22	11.25	4.04	4.39	12.83	5.98	4.40
WDBC-Cancer $\langle 3 \rangle$	3.69	3.51	2.99	3.87	4.22	8.61	7.04	6.86
Iris	2.00	2.00	2.00	2.00	2.00	2.67	1.33	1.33
Wine	0.00	0.00	0.00	0.56	0.56	1.69	7.81	7.81
Vehicle	13.95	13.83	14.07	14.89	14.89	15.37	28.59	17.86

(a) via Bayes (32); (b) via Likelihood (33); (c) via Mahalanobis (34)
Results by the β-CMVE are the minimum for all the combination of the eleven βs.

see that nicely small error rates comparable to those of Mangasarian et al. [13] are achieved via the modified Mahalanobis distance criterion when $(\beta_1, \beta_2) = (0.93, 0.76)$ and $(0.93, 0.77)$. Also, we see from these tables that learning and testing error rates are almost same with the same parameter setting β_1 and β_2.

5 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 [15]. The formulation yields a generalized notion of both the minimum volume ellipsoid covering all the data points and the ellipsoid characterized by the maximum likelihood estimation of the normal distribution. Computation of the generalized ellipsoid is accomplished through a convex optimization, referred to as Problem (CMVE(β)), and a modified version of an interior point algorithm developed by Sun and Freund [19] can solve it in a fairly efficient manner. Besides, when the parametric computation is needed for various βs (e.g., $\beta_1 = 0.0 < \beta_2 < ... < \beta_N < 1$), computational shortcut can be employed by using the facts: i) for $\beta = 0$, the explicit solution is available, ii) an optimal solution of Problem (CMVE(β_{h-1})) can be adopted as an initial feasible solution of Problem (CMVE(β_h)), and iii) only the active sample points (and the corresponding constraints) at optimality of Problem (CMVE(β_{h-1})) are expected to be a good superset of those of Problem (CMVE(β_h)). Numerical experiments show that exploiting these facts can reduce the total computation time for computing ten β-CMVEs. If some heuristics are incorporated as developed in Sun and Freund [19], much more large instances can be solved in a fairly efficient manner.

Table 4: Learning and Testing Error Rates of 10-Fold Cross-Validation for WDBC Cancer Data by Using the Three Attributes No.2, 24 and 25 [%]

via the Modified Mahalanobis' Distance (34)

(i) Learning Error

β_1 \ β_2	0.00	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.00	8.46	9.45	11.13	12.87	14.37	15.54	16.99	19.00	23.14	27.28	27.10
0.05	7.11	8.51	10.08	11.89	13.47	14.88	15.93	18.08	21.44	25.56	25.60
0.15	5.64	6.35	8.18	9.78	11.50	13.26	14.80	16.19	19.12	22.16	22.57
0.25	4.26	4.92	6.29	7.85	9.51	11.09	12.89	15.04	17.22	19.61	20.19
0.35	3.93	4.18	4.73	5.99	7.36	8.92	10.29	12.69	15.50	17.97	17.87
0.45	3.55	3.75	4.10	4.57	5.84	7.05	8.49	10.29	13.83	16.17	15.60
0.55	3.09	3.30	3.53	3.83	4.22	5.35	6.35	8.16	10.76	14.02	13.42
0.65	3.10	3.03	3.05	3.28	3.51	3.83	4.59	5.98	8.47	10.94	12.09
0.75	3.71	3.51	3.24	3.14	3.03	3.48	3.55	4.37	5.98	8.44	9.88
0.85	4.41	4.24	4.00	3.83	3.38	3.14	2.97	2.95	3.91	5.17	6.91
0.95	6.42	6.05	5.21	4.75	4.37	4.08	3.91	3.85	3.01	3.03	4.43

(ii) Testing Error

β_1 \ β_2	0.00	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.00	8.44	10.02	11.07	12.65	14.76	16.34	17.22	19.16	23.73	27.59	27.24
0.05	7.03	8.79	10.37	12.30	13.53	15.11	16.34	18.63	21.97	25.83	25.83
0.15	5.80	6.33	8.44	10.19	11.95	13.53	15.11	16.70	19.16	23.02	23.02
0.25	4.39	5.80	6.15	8.26	9.84	11.60	13.36	15.47	17.75	19.68	20.39
0.35	3.87	4.22	5.10	6.15	7.56	9.14	10.54	13.18	15.82	17.57	18.45
0.45	3.51	3.69	4.04	5.10	6.15	7.21	8.61	10.54	14.06	16.34	16.70
0.55	3.69	3.16	3.34	3.87	4.39	5.62	5.98	7.91	10.72	13.88	14.06
0.65	3.51	3.34	3.87	3.16	3.51	4.22	4.75	5.98	8.79	11.60	11.95
0.75	4.22	3.87	3.51	3.69	3.34	3.16	4.22	4.57	6.15	8.79	10.37
0.85	4.92	4.75	4.22	4.04	3.69	3.69	3.34	3.51	4.22	5.10	6.85
0.95	6.68	6.50	5.27	4.92	4.75	4.39	4.57	4.04	3.51	2.99	5.27

(iii) 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

(iv) 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

Motivated by such computational accessibility and the fact that the ellipsoidal construction approximately generalizes the Fisher's discriminant methods through a parameterization with β, we adopted this ellipsoid construction in a multiclass discrimination problem. From computational experiments, we see that this generalization improves the predictive accuracy than the classical Fisher's discrimination approaches. Also, for some data set, the proposed methods can achieve better predictive accuracy than the one-against-one ν-SVM approach. Moreover, the computation can be carried out in an efficient manner since the number of times for solving the optimization problem is proportional to the number of classes and the number of subdivisions on β. Analysis on statistical properties of the proposed methods and other applications of this method will be the future research.

Acknowledgment 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 Theorem 2.1

At first, we show that Problem (4) has an optimal solution. Since the equivalence between (4) and (5) is obvious, it suffices to show that (5) has an optimal solution. The Lagrangian dual of (5) is given as

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

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

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

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

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

with corresponding primal solution

$$\boldsymbol{Q} = (\boldsymbol{X}\tilde{\boldsymbol{\Lambda}}\boldsymbol{X}^\top - \boldsymbol{X}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\lambda}}^\top\boldsymbol{X}^\top)^{-1/2};\ \ \boldsymbol{\gamma} = \sum_{i\in I}\tilde{\lambda}_i\boldsymbol{Q}\boldsymbol{x}^i. \tag{36}$$

We observe that the dual (35) 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, (36) is a solution of Problem (5).

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 (4), 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 $\beta > 0$, $\alpha^* \ge 0$ follows and hence, $U^* > 0$ is shown. Indeed, assuming on the contrary that $\alpha^* < 0$, the constraint $F_\beta(\boldsymbol{Q}^*,\boldsymbol{\gamma}^*,\alpha^*) \le n$ of (4) is expressed as

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

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 (4) 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.2 Proof of Proposition 2.2

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 [15] 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.3 Proof of Proposition 2.3

The Lagrangian dual of (12) becomes

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

and the corresponding primal solution is given by

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

which is in common with the Lagrangian dual (35) of (5) as shown in the proof of Theorem 2.1. □

A.4 Proof of Proposition 2.4

Let us prove the proposition by contradiction. Suppose that Problem (13) has a solution $(\boldsymbol{Q}',\boldsymbol{\gamma}',\boldsymbol{\xi}')$ better than $(\frac{1}{\sqrt{\alpha^*}}\boldsymbol{Q}^*, \frac{1}{\sqrt{\alpha^*}}\boldsymbol{\gamma}^*, \frac{1}{\alpha^*}\boldsymbol{\xi}^*)$, i.e., $-\ln\det[\boldsymbol{Q}'] + P\boldsymbol{e}^\top\boldsymbol{\xi}' < -\ln\det[\frac{1}{\sqrt{\alpha^*}}\boldsymbol{Q}^*] + P\boldsymbol{e}^\top(\frac{1}{\alpha^*}\boldsymbol{\xi}^*)$, which implies that $-\ln\det[\sqrt{\alpha^*}\boldsymbol{Q}'] + \frac{1}{2}\alpha^* + \frac{1}{2(1-\beta)m}\boldsymbol{e}^\top(\alpha^*\boldsymbol{\xi}') < -\ln\det[\boldsymbol{Q}^*] + \frac{1}{2}\alpha^* + \frac{1}{2(1-\beta)m}\boldsymbol{e}^\top\boldsymbol{\xi}^*$. Since $(\sqrt{\alpha^*}\boldsymbol{Q}', \sqrt{\alpha^*}\boldsymbol{\gamma}', \alpha^*, \alpha^*\boldsymbol{\xi}')$ is feasible to Problem (12), this contradicts the optimality of $(\boldsymbol{Q}^*,\boldsymbol{\gamma}^*,\alpha^*,\boldsymbol{\xi}^*)$ to Problem (12). □

A.5 Proof of Proposition 2.5

Note that dual problem (35) of (CMVE(β)) requires that an optimal solution $\boldsymbol{\lambda}^*$ satisfy $\boldsymbol{e}^\top\boldsymbol{\lambda} = 1$ and $\boldsymbol{0} \leq \boldsymbol{\lambda} \leq \frac{1}{(1-\beta)m}\boldsymbol{e}$. Moreover, as complementarity conditions for optimality, $(\frac{1}{(1-\beta)m} - \lambda_i^*)z_i^* = 0$ holds for all $i \in I$. Noting that the complementarity conditions imply $\lambda_i^* = \frac{1}{(1-\beta)m}$ for all $i \in ERR$, we have

$$\begin{aligned} 1 = \boldsymbol{e}^\top\boldsymbol{\lambda}^* = \textstyle\sum_{i\in SV}\lambda_i^* \leq \sum_{i\in SV}\frac{1}{(1-\beta)m} = \frac{|SV|}{(1-\beta)m}, \\ 1 = \boldsymbol{e}^\top\boldsymbol{\lambda}^* \geq \textstyle\sum_{i\in ERR}\frac{1}{(1-\beta)m} = \frac{|ERR|}{(1-\beta)m}, \end{aligned}$$

which lead to $\frac{|ERR|}{m} \leq 1-\beta \leq \frac{|SV|}{m}$.

References

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

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

[3] C.L. Blake and C.J. Merz, 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, 1998.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[5] C.C. Chang and C.J. Lin, LIBSVM : A Library for Support Vector Machines [http://www.csie.ntu.edu.tw/~cjlin/libsvm], 2001.

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

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

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

[9] C.J. Huberty, *Applied Discriminant Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, USA, 1994.

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

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

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

[13] O.L. Mangasarian, W.N. Street, and W.H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, 43, pp.570–577, 1995.

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

[15] T.R. Rockafellar and S. Uryasev, "Conditional Value-at-Risk for General Loss Distributions," *Journal of Banking and Finance*, 26, pp.1443–1471, 2002.

[16] P.J. Rousseeuw, "Multivariate Estimation with High Breakdown Point," *Proc. of the 4th Pannonian Symp. on Math. Stat.*, Bad Tatzmannsdorf, Austria, 1983, pp.283–297.

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

[18] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, New support vector algorithms, *Neural Comput.* 12, pp.1207–1245, 2000.

[19] P. Sun and R.M. Freund, "Computation of Minimum Volume Covering Ellipsoids," *Operations Research* 52, no.5, pp.690–706, 2004.

[20] D.M. Titterington, "Optimal Design: Some Geometrical Aspects of D-Optimality," *Biometrika*, 62, pp.313–320, 1975.

[21] K. Toh, M. Todd, and R. Tütüncü, "Sdpt3 – a matlab software package for semidefinite programming," *Optimization Methods and Software*, 11, pp.545–581, 1999.

[22] E. Welzl, "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, pp.359–370, Springer-Verlag, Berlin, 1991.

[23] D.L. Woodruff and D.M. Rocke, "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, 2, pp.69–95, 1993.

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