©Freund Publishing House Ltd. International Journal of Nonlinear Sciences and Numerical Simulation 5(2), 157-170, 2004

A Probabilistic Method for Detecting Multivariate Extreme Outliers

Shafiu Jibrin

Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, Arizona 86011, USA, E-mail: Shafiu.Jibrin@nau.edu

Irwin S. Pressman

School of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6, Canada, E-mail: irwin_pressman@math.carleton.ca

Matias Salibian-Barrera

School of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6, Canada, E-mail: matias@math.carleton.ca

Abstract

Given a data set arising from a series of observations, an *outlier* is a value that deviates substantially from the natural variability of the data set as to arouse suspicions that it was generated by a different mechanism. We call an observation an *extreme outlier* if it lies at an abnormal distance from the "center" of the data set. We introduce the Monte Carlo *SCD* algorithm for detecting extreme outliers. The algorithm finds extreme outliers in terms of a subset of the data set called the *outer shell*. Each iteration of the algorithm is polynomial. This could be reduced by preprocessing the data to reduce its size.

This approach has an interesting new feature. It estimates a relative *measure* of the degree to which a data point on the outer shell is an outlier (its "outlierness"). This measure has potential for serendipitous discoveries in data mining where unusual or special behavior is of interest. Other applications include spatial filtering and smoothing in digital image processing. We apply this method to baseball data and identify the ten most exceptional pitchers of the 1998 American League. To illustrate another useful application, we also show that the *SCD* can be used to reduce the solution time of the D-optimal experimental design problem.

Keywords: Extreme outliers, outlierness, semidefinite programming, Monte Carlo, redundancy, D-optimal design

1. Introduction

Let $D = \{a^j\}_{j=1}^n$ be a set of multivariate observations in IR^p. We assume throughout that the number of points $n > p$. An *outlier* is a member of the data set D that is abnormal in comparison to the main body of observations. An outlier relatively deviates from the natural variability of the data set as to arouse suspicions that it was generated by a different mechanism. An *extreme outlier* is an observation that lies at an abnormal distance from the "center" of the data set, and we shall give a precise definition of this below.

Outlier detection has become an area of great interest in data mining. For example [15] discusses applications to detect various forms of transactional fraud. Data miners, in their analysis of large data sets, may regard outliers as indicators of interesting events.

A common estimator of the "center" of a dataset is the sample mean. Sample means can be distorted nonetheless by the presence of just one outlier [18]. A commonly used measure of robustness is the *asymptotic breakdown point*.

Intuitively, it reflects the smallest proportion of contamination of the data set that can cause the estimator to break down when the sample size $n \to \infty$. The mean has an asymptotic breakdown point of 0%.

One procedure for detecting outlier uses the Mahalanobis distance d_j given by

$$d_j = \sqrt{(a^j - \bar{x})^T \mathbf{Cov}^{-1} (a^j - \bar{x})}, \qquad (1.1)$$

where $\bar{x}$ is the mean of D and $\mathbf{Cov} = 1/(n-1)\sum_{i=1}^{n} (a^j - \bar{x})(a^j - \bar{x})^T$ denotes its covariance matrix. Points with large d_j value may be deleted from D. Since the distances d_j are based on the sample mean, it is not surprising that this approach does not always work well when more than one outlier is present ([4], [18]).

References for statistical methods for outlier detection include [4] and [20]. Many of the techniques in [4] require prior knowledge about the underlying distribution. These methods identify outliers by examining the deviations of individual data points relative to their distribution (e.g., Normal, Poisson, etc.) [4]. Unfortunately, this approach is not applicable in most practical cases when the underlying distribution is unknown.

Another approach uses Mahalanobis distance as in (1.1) above but replacing $\bar{x}$ and $\mathbf{Cov}$ by highly robust multivariate estimators of location and scatter [20]. Rocke and Woodruff [17] discuss some challenges encountered in detecting multivariate outliers and explain why the level of difficulty increases with the dimension of the data. Other outlier detection methods are distance-based ([13], [20]). These methods distinguish outliers by the number of data points in the neighborhood of an individual data point. No assumptions on the distribution of the data set are made. Density-based methods [7] are based on the local density of an observation's neighborhood. These methods measure how strongly an observation can be viewed as an outlier. Layer-based methods rely on the computation of different layers or hulls that partition the data set into shells. Outliers are observations in the outer layers of these shells ([4], [13]). An example is the ellipsoidal peeling method [21]. More recently, depth-based [23] multivariate estimates have been proposed in the statistical literature ([9], [16], [27]). Projection estimates [1] provide another interesting approach.

We introduce a new procedure here to detect outliers, that we call the *SCD* or *(semidefinite coordinate direction)* method. It uses the L_1 estimator of "center" of D, known as the Fermat-Weber point x^*, given by

$$x^* = \arg\min_{x \in \mathbb{R}^p} \sum_{j=1}^{n} \| a^j - x \|. \qquad (1.2)$$

Although the Fermat-Weber point is not affine equivariant in general, it is equivariant with respect to all affine transformations that preserve Euclidean distances. It generalizes the univariate median and it has a high breakdown point of 50%.

We use semidefinite programming techniques to determine the points of the data that are on the "outer shell" of the sample (see Definition 1 below). A matrix $\mathbf{A}$ is called *positive definite* ($\mathbf{A} \succ 0$) (respectively, *positive semidefinite* ($\mathbf{A} \succeq 0$)) if and only if all of its eigenvalues are strictly positive *(respectively, non-negative)*.

A linear matrix inequality (*LMI*) is a constraint of the form $\mathbf{A}_0 + \sum_{i=1}^{p} x_i \mathbf{A}_i \succeq 0$, where $\mathbf{A}_i$ are $m \times m$ symmetric matrices. Constraints of this type generalize linear and quadratic constraints. They arise directly in Control Theory, Statistics and Combinatorial Optimization ([2], [24]).

Consider a system of *LMI*'s

$$\mathbf{A}^{(j)}(x) := \mathbf{A}_0^{(j)} + \sum_{i=1}^{p} x_i \mathbf{A}_i^{(j)} \succeq 0, \qquad (1.3)$$

$(1 \le j \le n)$,

where $x \in \mathrm{IR}^p$ and $\mathbf{A}_i^{(j)}$ are $m_j \times m_j$ symmetric matrices. An *LMI* constraint is called *necessary* (in this system) if its removal changes the feasible region of the system, otherwise it is called *redundant*. An *LMI* can be both necessary and redundant, for example, in the case of *duplicate LMI*'s.

The *semidefinite coordinate direction algorithm*, *SCD*, presented in [12] is a Monte

Carlo method for detecting necessary linear matrix inequalities. *SCD* applies to convex bounded regions constrained by a finite set of *LMI*'s. It works as follows: let $\hat{x}$ be a random interior point in the feasible region called the *standing point*. Select a random search vector $s \in \mathrm{IR}^p$ along a coordinate direction to determine a ray $\{\hat{x}+\sigma s : \sigma \geq 0\}$. Let

$$\sigma_1 = \max_{1\leq j<n}\{\sigma : \mathbf{A}^{(j)}(\hat{x}+\sigma s) \succeq 0\}.$$

The k^{th} constraint is necessary if

$$\sigma_1 = \max\{\sigma : \mathbf{A}^{(k)}(\hat{x}+\sigma s) \succeq 0\}. \qquad (1.4)$$

The above procedure is repeated until a termination criterion is satisfied.

Extreme outliers will be defined to correspond to the necessary *LMI*'s within a certain system. The use of coordinate directions greatly reduces the number of calls to the random number generator and provides for easy updating of the inverse of a key matrix.

As an application of this method, consider the linear model $z = \hat{\mathbf{A}}\beta + \omega$, where ω is an "n-vector" of errors and $\hat{\mathbf{A}}$ is called the design matrix. Let x_j be the number of times the *jth* row of $\hat{\mathbf{A}}$ is selected. The "Fisher information matrix"

$$\mathbf{F}(x) := \sum_{j=1}^{n} x_j \mathbf{A}^{j\bullet}(\mathbf{A}^{j\bullet})^T$$

arises in the formula for the least-squares estimator of β, based on the design x. Experimental Design theory aims at finding the experimental conditions that provide the best regression estimates by minimizing a measure of the overall variability of the regression estimator, $\sigma^2\mathbf{F}^{-1}(x)$. In D-optimal design, the determinant of the inverse of the Fisher matrix is minimized [25]. Finding this optimum can be computationally expensive. In Section 4, we show how *SCD* facilitates the efficient solution of the D-optimal design problem. This simplification is an important application of the method introduced in this paper.

2. Outer shells and Redundancy

In this section, we develop the *SCD* method for detecting points on the outer shell of D. We show the connection between the identification of the outer shell and the problem of detecting necessary constraints in semidefinite programming. The practicality of the ellipsoidal approach to describing extreme outliers will be confirmed by the simplicity of the calculation that arises

2.1. Affine Independence

The set $D = \{a^j\}_{j=1}^n \subset \mathrm{IR}^p$ is an affinely independent set if no point a^k is contained in the affine space spanned by p or fewer points of $D - \{a^k\}$. In a randomly generated set, the probability that $p+1$ points are an affinely dependent set is zero. We make the following independence assumption throughout:

Assumption: *Every subset of D of size* $(p+1)$ *is affinely independent.* (2.1)

Recall that the Fermat-Weber point x^* of D given in (1.2) is a robust estimator of "center". It can be computed by solving the following optimization problem:

$$\min \sum_{j=1}^{n} t_j$$

$$s.t. \ \| a^j - x \| \leq t_j, j = 1,2,\cdots$$

where t_j are auxiliary variables in IR. This can be cast as a semidefinite programming program *SDP* [24]:

$$\min \sum_{j=1}^{n} t_j \qquad (2.2)$$

$$s.t. \begin{bmatrix} t\mathbf{I} & a^j - x \\ (a^j - x)^T & t \end{bmatrix} \succeq 0, \qquad (2.3)$$

$$(1 \leq j \leq n).$$

Note that the constraints (2.3) can be expressed in the form (1.3), and so are *LMI* constraints. The above *SDP* can be computed efficiently using the SDPSOL [26] software. The Fermat-Weber point is commonly used for facility location to find the point that minimizes the total travel distance to each of a population

of customers. We consider a data point that relatively deviates substantially from the Fermat-Weber point to be an extreme outlier.

Definition 2.1 A point $a^k \in D$ belongs to the *outer shell* of D denoted by $O(D)$, if there is an ellipsoid centered at the Fermat-*Weber point* x^* that contains $D-\{a^k\}$ and whose boundary passes through a^k. We call all such points *extreme outliers.*

Note that if $a^k \in D$ belongs to the outer shell $O(D)$, then it is not representable as a linear combination of p or fewer other points. The independence assumption on D implies that p of the vectors $a^j - a^j$ are linearly independent and form a basis of IR^p. In practice any finite set can be made affinely independent by slight perturbations of the data. Not every point of D that is on its convex hull is an extreme outlier.

The *SCD* algorithm that we are presenting is designed to find points in the outer shell $O(D)$ of D. The set of all ellipsoids centered at x^* that contain D is given by the solutions to the system

$$\|\mathbf{E}(a^j - x^*)\|^2 \le 1, \quad 1 \le j \le n \qquad (2.4)$$

$$\mathbf{E} \succeq 0, \qquad (2.5)$$

where $\mathbf{E} \in S^p$ denotes the set of all $p \times p$ symmetric matrices. Note that (2.4) can be expressed equivalently in the form (1.3), so, the above system is equivalent to:

$$A^{(j)}(\mathbf{E}) := \begin{bmatrix} \mathbf{I} & \mathbf{E}(a^j - x^*) \\ (\mathbf{E}(a^j - x^*))^T & 1 \end{bmatrix} \succeq 0, \qquad (2.6)$$

$(1 \le j \le n)$

$$\mathbf{E} \succeq 0. \qquad (2.7)$$

The ambient space of the constraints is IR^N, where $N = p(p+1)/2$.

Let $R(D) \subset \mathrm{IR}^N$ be the feasible region defined by the system (2.6) and (2.7). The set $R(D)$ corresponds to the space of ellipsoids centered at the Fermat-Weber x^* that contain D. It is well known [24] that the *LMI* constraints (1.3) are convex constraints in IR^{iv}. Since SCD applies to bounded feasible regions we show the connection between affine independence of D and the boundedness of $R(D)$.

Theorem 2.1 The feasible region $R(D)$ given by (2.6) and (2.7) is bounded if a subset of $p+1$ of the data points $\{a^j\}_{j=1}^n$ is an affinely independent set.

Proof: Suppose $R(D)$ is unbounded. With $A^{(j)}(\mathbf{E})$ given by (2.6), a necessary condition [12] for the unboundedness of $R(D)$ is the existence of $\mathbf{S} \in S^p$, not zero, such that

$$\begin{bmatrix} 0 & \mathbf{S}a^j \\ (\mathbf{S}a^j)^T & 0 \end{bmatrix} \succeq 0, \; 1 \le j \le n \qquad (2.8)$$

$$\mathbf{S} \succeq 0. \qquad (2.9)$$

The n equations (2.8) imply that

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}a^j \\ (\mathbf{S}a^j)^T & 0 \end{bmatrix} \succeq 0, \; 1 \le j \le n$$

By the Schur Lemma [11], we obtain $-(\mathbf{S}a^j)^T(\mathbf{S}a^j) \ge 0, 1 \le j \le n$. Hence,

$$\mathbf{S}a^j = 0, \; 1 \le j \le n \qquad (2.10)$$

$$\mathbf{S}(a^i - a^j) = 0, \; 1 \le i, j \le n \qquad (2.11)$$

By the affine independence of $\{a^j\}_{j=1}^n$, p of the vectors $a^i - a^j$ are linearly independent and form a basis of IR^p. The equation (2.11) yields $\mathbf{S}=0$. This is a contradiction of the assumption that $\mathbf{S}$ is not zero. Thus, the region $R(D)$ is bounded.

2.2. Topology of $R(D)$

We adapt some ideas from Topology to the case of a finite set. Denote the topological boundary of a closed set $T \subseteq \mathrm{IR}^p$ by $\partial(T)$.

Definition 2.2 The *convex hull*, $CH(D)$, of D is the intersection of all (closed) half planes in IR^p containing D. Its boundary, $\partial(CH(D))$, is a polytope.

A convex linear combination of x and $y, \lambda x + (1-\lambda)y, 0 \le \lambda \le 1,$ is *strict* if

$0 < \lambda < 1$.

Definition 2.3 A point of D that is not a strict convex linear combination of two other points of D is called an *extreme point*.

The ellipsoidal peeling method described in [21], *trims the most extreme observations*, namely those lying on the boundary of the minimum volume ellipsoid containing the data set. The number of points trimmed from D is almost certain to be at most $p(p+3)/2$ [21].

For any $\mathbf{E} \succ 0$, there is a corresponding *ellipsoid* $e(\mathbf{E}) \subset \mathrm{IR}^p$ with center x^*:

$$e(\mathbf{E}) = \{z \in \mathrm{IR}^p : \| \mathbf{E}(z - x^*) \|^2 \le 1\}. \quad (2.12)$$

If $\mathbf{E} \succeq 0$ and $\det(\mathbf{E}) = 0$, then $e(\mathbf{E})$ describes an ellipsoidal cylinder. The boundary of the ellipsoid *(or ellipsoidal cylinder)* determined by $\mathbf{E}$ is denoted by

$$\partial(e(\mathbf{E})) = \{z : \| \mathbf{E}(z - x^*) \|^2 = 1\}. \quad (2.13)$$

Definition 2.4 $e(\mathbf{E})$ *properly* contains the set D if for all $a^j \in D$, $\| \mathbf{E}(a^j - x^*) \| < 1$.

We next show that the interior of $R(D)$ is non-empty by giving a necessary and sufficient condition for a point to be in the interior. Ellipsoids in IR^p that properly contain D correspond to points in the interior of $R(D)$. We will show later that all ellipsoidal cylinders that properly contain D are on the boundary of $R(D)$.

Proposition 2.1 $e(\mathbf{E})$ properly contains $D \iff$ there is an open neighborhood $N(\mathbf{E}) \subseteq R(D)$ for which $\| \mathbf{X}(a^j - x^*) \| < 1$ for every $\mathbf{X} \in N(\mathbf{E})$ and for each $a^j \in D$.

Proof: Suppose that $\| \mathbf{E}(a^j - x^*) \| < 1$, $1 \le j \le n$, so that we can define a strictly positive number $\zeta = 1 - max_{1 \le j \le n} \| \mathbf{E}(a^j - x^*) \| > 0$. Use the set D to define an open neighborhood U of the origin in IR^N,

$$U = \left\{ \mathbf{W} : \| \mathbf{W}(a^j - x^*) \| < .9\zeta, 1 < j \le n \right\}.$$

The required open neighborhood $N(\mathbf{E})$ is the Euclidean translation of this open set, $U + \mathbf{E}$. Take z in the convex hull of D, i.e., $z = \sum_{j=1}^n \mu_j a^j$, $\mu_j \ge 0$, $\sum_{j=1}^n \mu_j = 1$. We have:

$$\| \mathbf{E}(z - x^*) \| = \| \mathbf{E}(\sum_{j=1}^n \mu_j a^j - \sum_{j=1}^n \mu_j x^*) \|$$

$$= \| \sum_{j=1}^n \mu_j \mathbf{E}(a^j - x^*) \| \le \sum_{j=1}^n \mu_j \times (1 - \zeta)$$

$$= (1 - \zeta).$$

Similarly, we have $\| \mathbf{W}(z - x^*) \| < .9\zeta$. Thus, for all $(\mathbf{W} + \mathbf{E}) \in U + \mathbf{E}$,

$$\| (\mathbf{W} + \mathbf{E})(z - x^*) \| \le \| \mathbf{W}(z - x^*) \| +$$

$$\| \mathbf{E}(z - x^*) \| < .9\zeta + (1 - \zeta) < 1.$$

Conversely, if there is an open neighborhood $N(\mathbf{E})$ surrounding $\mathbf{E}$ for which $\| \mathbf{X}(a^j - x^*) \| < 1$ for every j, $1 \le j \le n$, and for all $\mathbf{X} \in N(\mathbf{E})$ then in particular $\| \mathbf{E}(a^j - x^*) \| < 1$ holds for every $a^j \in D$.

Since there are many ellipsoids surrounding the finite set D, the proposition above proves that the interior of $R(D)$ must be nonempty. The volume of an ellipsoid $e(\mathbf{E})$ is given by:

$$V = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d+2}{2}) \times \det(\mathbf{E})},$$

and $e(t\mathbf{E})$ has volume V / t^d. As the parameter $t \to \infty$ the ellipsoid shrinks. Along the ray from $\mathbf{E}$ moving directly away from the origin in IR^N, the family of concentric ellipsoids $e(t\mathbf{E})$ contract around D until there is t_1 and an $a^k \in D$ for which $\| t_1 \mathbf{E}(a^k - x^*) \| = 1$. This gives an extreme point (2.13) of $R(D)$. If $\mathbf{E} \in R(D)$, the ellipsoid contains all of D, and this containment is proper if $\mathbf{E}$ is in the interior of $R(D)$.

2.3. The boundary of $R(D)$

The role of ellipsoidal cylinders is noteworthy

here. We can take an $\mathbf{X} \succ 0$ in the interior of $R(D)$ and drive its smallest eigenvalue to zero, without forcing the ellipsoid to pass through a point of D. For instance, consider the subset of IR^2

$D_4 = \{(3,4),(-3,4),(-3,-4),(3,-4)\}$. The circle of radius 5.1 centered at the origin containing D_4 can be described by the matrix $\mathbf{E} = 1/5.1\mathbf{I}_2$. The circle is converted into the *cylinder* $|x_1| \leq 5.1$ in IR^2 as the second of the diagonal elements (eigenvalues) becomes very small.

The following proposition shows that on the boundary of the feasible region $R(D)$ one also finds the ellipsoids containing D that have a point $a^k \in D$ on the boundary, plus all ellipsoidal cylinders containing D.

Proposition 2.2 Suppose $\mathbf{X} \succeq 0$. Then $\mathbf{X} \in \partial(R(D)) \Longleftrightarrow \| \mathbf{X}a^j \| \leq 1$ for all $1 \leq j \leq n$ and either $\mathbf{X} \succ 0$ and $\| \mathbf{X}a^k \| = 1$ for at least one value of k, or $\det(\mathbf{X}) = 0$.

Proof: For $\mathbf{X} \in \partial(R(D))$, consider the case $\det(\mathbf{X}) > 0$. If $\| \mathbf{X}a^k \| < 1$ for all $1 \leq k \leq n$, then by Proposition 2.1 $\mathbf{X}$ is in the interior so $\mathbf{X} \notin \partial(R(D))$. Hence $\| \mathbf{X}a^k \| = 1$ for at least one value of k, $1 \leq k \leq n$. Hence, $e(\mathbf{X})$ passes through at least one a^k. Otherwise $\det(\mathbf{X}) = 0$. Conversely, suppose $\| \mathbf{X}a^j \| \leq 1$ for all j, $1 \leq j \leq n$. First, suppose $\det(\mathbf{X}) = 0$. Let P denote the cone of positive semidefinite matrices contained in S^p. Since $\mathbf{X}$ is a positive semidefinite matrix, it is on the boundary of the cone P. Geometrically this is an ellipsoidal cylinder $e(\mathbf{X})$ that contains D. Hence $\mathbf{X} \in \partial(R(D))$. Now suppose $\mathbf{X} \succ 0$ and $\| \mathbf{X}a^k \| = 1$. By Proposition 2.1 X cannot be in the interior of $R(D)$ so it must be on the boundary.

That is, given $\mathbf{E}_0 \in S^p$. such that $\| \mathbf{E}_0(a^j - x^*) \| \leq 1$ for all $a^j \in D$, and for some k $\| \mathbf{E}_0(a^k - x^*) \| = 1$, then $\mathbf{E}_0$ is a boundary point.

Theorem 2.2 A point $a^k \in D$ belongs to the outer shell $O(D)$ if and only if the k^{th} constraint in the system (2.6) and (2.7) is necessary.

Proof: Let $a^k \in D$. If $a^k \in O(D)$, then there is an ellipsoid centered at x^*

$$e(\mathbf{E}_0) = \{x : \| \mathbf{E}_0(x - x^*) \|^2 \leq 1\}$$

that contains $D - \{a^k\}$ and whose boundary passes through a^k. Hence, $\| \mathbf{E}_0(a^k - x^*) \|^2 = 1$ and $\| \mathbf{E}_0(a^j - x^*) \|^2 \leq 1$ for all $a^j \in D - \{a^k\}$. By Proposition 2.2, $\mathbf{E}_0$ is on the boundary of $R(D)$. Choose a point $\mathbf{E}^*$ in the interior of the convex region $R(D)$ and define the ray $\mathbf{S}$ from $\mathbf{E}^*$ to $\mathbf{E}_0$, $\mathbf{S} = \mathbf{E}_0 - \mathbf{E}^*$. Since $\mathbf{E}_0 \in R(D)$, $\mathbf{E}_0 = \mathbf{E}^* + 1\mathbf{S}$,

$$1 = \max\{\sigma : \mathbf{A}^{(k)}(\mathbf{E}^* + \sigma\mathbf{S}) \succeq 0\} =$$
$$\max\{\sigma : \mathbf{A}^{(j)}(\mathbf{E}^* + \sigma\mathbf{S}) \succeq 0, 1 \leq j \leq n\}.$$

It follows that the k^{th} constraint is necessary. Conversely, suppose the k^{th} constraint is necessary (1.4). Choose a point $\mathbf{E}^*$ in the interior of $R(D)$. Then, there is a search vector S for which we have:

$$\sigma_1 = \max\{\sigma : \mathbf{A}^{(k)}(\mathbf{E}^* + \sigma\mathbf{S}) \succeq 0\} =$$
$$\max\{\sigma : \mathbf{A}^{(j)}(\mathbf{E}^* + \sigma\mathbf{S}) \succeq 0, 1 \leq j \leq n\}.$$

Let $\mathbf{E}_0 = \mathbf{E}^* + \sigma_1\mathbf{S}$. It follows that $\| \mathbf{E}_0(a^k - x^*) \|^2 = 1$ and $\| \mathbf{E}_0(a^j - x^*) \|^2 \leq 1$ for all $j = 1, \ldots, k-1, k+1, \ldots$. Hence $\mathbf{E}_0$ is an ellipsoid that contains $D - \{a^k\}$ and whose boundary passes through a^k, so $a^k \in O(D)$.

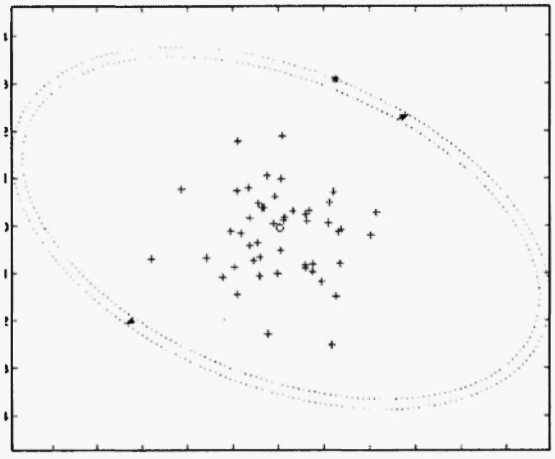

Figure 1 The figure shows how separation of point * from other points by an ellipsoid extends to separation of * by an ellipsoid whose boundary passes through *.

Proposition 2.3 Let $e(\mathbf{E})$ be an ellipsoid centered at x^*. If $e(\mathbf{E})$ separates a^k from $D-\{a^k\}$, then $a^k \in O(D)$.

Proof: As t decreases from 1 towards 0, $e(t\mathbf{E})$ continuously expands to an ellipsoid centered at x^* that separates a^k from $D-\{a^k\}$ and whose boundary passes through a^k, (*as indicated in Figure 1*). Hence, $a^k \in O(D)$.

Denote the minimum volume ellipsoid centered at x^*, containing D by e_Ω. The ellipsoid e_Ω can be found by solving the optimization problem:

$$\max \ \log\det(\mathbf{E})$$

$$s.t.\begin{bmatrix} \mathbf{I} & \mathbf{E}(a^j - x^*) \\ (\mathbf{E}(a^j - x^*))^T & 1 \end{bmatrix} \succeq 0,$$

$$(1 \le j \le n)$$

$$\mathbf{E} \succ 0$$

This problem can be solved using SDPSOL [26] software. In the Proposition below we note that, by the definition of $O(D)$, all of the points on the boundary of e_Ω belong to the outer shell.

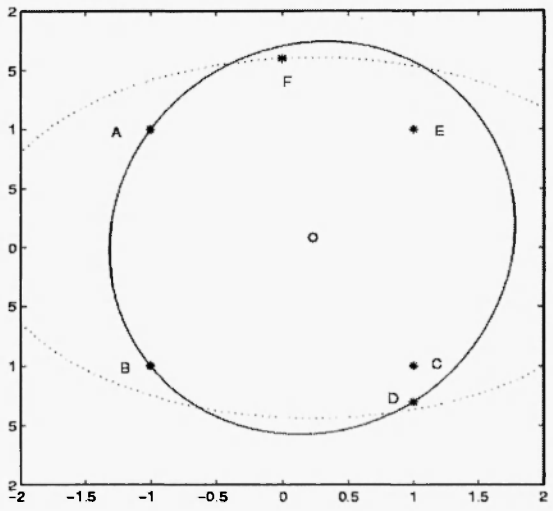

Figure 2. The solid ellipsoid is e_Ω. The point F is in the outer shell $O(D)$, but not on e_Ω. Note that F is separated from $D-\{F\}$ by the dotted ellipsoid centered at the Fermat-Weber of D and passing through F. The Fermat-Weber point is $(0.2325, 0.0842)$.

Proposition 2.3 If $a^k \in D \cap \partial(e_\Omega)$, then $a^k \in O(D)$.

Figure 2 shows that a point in $O(D)$ is not necessarily a boundary point of e_Ω, where $D = \{A, B, C, D, E, F\} =$ $\{(-1,1)^T, (-1,-1)^T, (-1,1)^T, (1,-1.3)^T,$ $(1,1)^T, (0,1.6)^T\}$

3. The *SCD* Algorithm

We introduce the *SCD* algorithm to find points on the outer shell of D by examining the system (2.6) and (2.7) to find the corresponding outer shell $O(D)$.

We adapt SCD [12] here to apply to the system (2.6) and (2.7) and describe the *SCD* algorithm. The SCD algorithm is an extension of the CD algorithm [5] for detecting necessary linear constraints.

Let $\mathbf{E}_0$ be the standing point, and $\mathbf{S} \in S^p$ a search direction. The hitting step involves the computation of the distances $\sigma_1^{(j)}$ and $\sigma_2^{(j)}$ to the boundary of each constraint (2.6) along the line determined by point $\mathbf{E}_0$ and direction $\mathbf{S}$. The following proposition shows how to compute the distances $\sigma_1^{(j)}$ and $\sigma_2^{(j)}$. For $a^j \in D$, define

$$z^j = \mathbf{E}_0(a^j - x^*) \tag{3.1}$$

$$w^j = \mathbf{S}a^j \tag{3.2}$$

Proposition 3.1 The two distances $\sigma_1^{(j)}$ and $\sigma_2^{(j)}$ to the boundary of the j^{th} constraint (2.6) along the line determined by the point $\mathbf{E}_0$ and the direction $\mathbf{S}$ are given by

$$\sigma_+^{(j)}(z^j, w^j) = \frac{-(z^j)^T w^j + \sqrt{((z^j)^T w^j)^2 + \| w^j \|_2^2 (1 - \| z^j \|_2^2)}}{\| w^j \|_2^2} \tag{3.3}$$

$$\sigma_-^{(j)}(z^j, w^j) = \frac{-(z^j)^T w^j - \sqrt{((z^j)^T w^j)^2 + \| w^j \|_2^2 (1 - \| z^j \|_2^2)}}{\| w^j \|_2^2} \tag{3.4}$$

Moreover, the two distances to the boundary of the $\mathbf{E} \succeq 0$ constraint along the line through the standing point $\mathbf{E}_0$ in direction $\mathbf{S}$ are given by

$$\sigma_+^{(n+1)} = \frac{1}{\lambda_{max}(-\mathbf{S}\mathbf{E}_0^{-1})}, \sigma_-^{(n+1)} = \frac{1}{\lambda_{min}(-\mathbf{S}\mathbf{E}_0^{-1})}$$

where σ_+ and σ_- respectively denote the maximum positive eigenvalue and the minimum negative eigenvalue.

Proof: Clearly $\| z^j \| < 1$ by (2.12). Note that it is possible for $w^j = 0$; in such cases we will have $\sigma_+^{(j)}(z^j, w^j) = \infty$, or $\sigma_-^{(j)}(z^j, w^j) = -\infty$. By (2.4), the distances $\sigma_1^{(j)}$ and $\sigma_2^{(j)}$ are given by the two solutions of the system

$$[\mathbf{E}_0(a^j - x^*) + \sigma(\mathbf{S}a^j)]^T [\mathbf{E}_0(a^j - x^*) + \sigma(\mathbf{S}a^j)] = 1,$$

where $\sigma \in \mathrm{IR}$. Solve the system for σ and then use the substitutions (3.1) and (3.2) to get (3.3) and (3.4). The rest follows from Theorem 2.1 in [12].

The *SCD* procedure effects a rank one perturbation $\mathbf{E}_0 + \sigma\mathbf{S}$ of $\mathbf{E}_0$, where $\sigma > 0$. In the algorithm, $\mathbf{S}$ is chosen from the set of N coordinate direction vectors, so there are only two possible cases in which $\mathbf{S} = [\mathbf{S}_{ij}] \neq 0$:

1. for one $i, \mathbf{S}_{ii} = 1$, $\mathbf{S}_{jk} = 0$ otherwise;
2. for one pair (k, j), $k \neq j$, $\mathbf{S}_{jk} = \mathbf{S}_{kj} = 1$, $\mathbf{S}_{jk} = 0$ otherwise.

Denote the perturbed matrix $\mathbf{E}_0 + \sigma\mathbf{S}$ by $\mathbf{E}_0$. Proposition 3.1 shows the two bounds on σ, i.e., $\sigma_+^{(q+1)}$ and $\sigma_-^{(q+1)}$, that ensure that $\mathbf{E}_0$ remains positive definite. Note that $\mathbf{E}_0$ changes by at most two entries, after each iteration. The inverse $\mathbf{E}_0^{-1}$ requires $4p^2$ multiplications to update using the following modification of a procedure given in [6]. The eigenvalues of $\mathbf{S}\mathbf{E}_0^{-1}$ are found by examining a 1×1 or 2×2 matrix, so at worst this is equivalent to solving a quadratic.

Procedure Φ**:** to find $\hat{\mathbf{E}}_0^{-1}$ from $\mathbf{E}_0^{-1}$

Let $\mathbf{S}_{kj} = \mathbf{S}_{jk} \neq 0$ and $\mathbf{S}_{it} = 0$, otherwise. Let $\mathbf{B} := \hat{\mathbf{E}}_0^{-1}$ and $\mathbf{A} := \mathbf{E}_0^{-1}$. Denote the k^{th} row (respectively i^{th} column) of matrix $\mathbf{M}$ by $\mathbf{M}^{k\bullet}(\mathbf{M}^{\bullet i})$. Set $\mathbf{B}^{\bullet k} := 1/(\hat{\mathbf{E}}_0^{k\bullet} * \mathbf{A}^{\bullet k}) * \mathbf{A}^{\bullet k}$.

```
for i=1:p
  if i ≠ k
    B^•i = E_0^•i − (Ê_0^k• * A^•i) * B^•k
  end
end
if k ≠ j
  Update A := B
  Update B^•j := 1/(Ê_0^j• * A^•j) * A^•j
  for i=1:p
    if i ≠ j
      B^•i = E_0^•i − (E_0^j• * A^•i) * B^•j
```

end
end
end

Corollary 3.1 $\sigma_+^{(q+1)}, \sigma_-^{(q+1)}$ can be computed with $O(p^2)$ floating point operations for each *SCD* search.

Proof: By the structure of $\mathbf{S}$, the matrix $\mathbf{S}\mathbf{E}_0^{-1}$ in Proposition 3.1 has at most two nonzero rows and therefore requires $O(p^2)$ flops to find its eigenvalues. The inverse $\mathbf{E}_0^{-1}$ is found from Procedure Φ so the computation requires at most $4p^2$ multiplications.

An interior point $\mathbf{E}_0$ of $R(D)$ can be found using the following procedure on the given set of data points D to use as the standing point of the algorithm.

1. Find $\rho = \max\{\| a^j - x^* \|, 1 \le j \le n\}$.
2. Choose the sphere $\mathbf{E}_0 = 1/(2\rho)\mathbf{I}$ of radius 2ρ, centered at x^*, as an interior point.

Each iteration of *SCD* gives a new standing point $\mathbf{E}_0$. The standing points generated by the *CD* algorithm are asymptotically uniform [5] in linear programming. We conjecture that the standing points $\mathbf{E}_0$ generated by *SCD* are asymptotically uniform in $R(D)$ [8].

SCD Outlier Detection Algorithm

Initialization: Set $O(D) = \square$. Choose an interior point $\mathbf{E}_0$ of $R(D)$ and call it the standing point. Choose *MAX*, a maximum number of iterations of this algorithm. Calculate $z^j = \mathbf{E}_0(a^j - x^*), 1 \le j \le n$ Set $f_j = 0, \ 1 \le j \le n+1$. Set $N = p(p+1)/2$. Set $iter = 0$.

Repeat

Search Direction: Choose a random coordinate search direction $\mathbf{S} \in S^p$

Hitting Step: For each $j, 1 \le j \le n+1$, compute $\sigma_+^{(j)}$ and $\sigma_-^{(j)}$ as in Prop. 3.1

$\sigma_+ = \min\{\sigma_+^{(j)} : 1 \le j \le n+1\}$

$\sigma_- = \min\{\sigma_-^{(j)} : 1 \le j \le n+1\}$.

For $1 \le k \le n$,

if $\sigma_+^{(k)} = \sigma_+$ or $\sigma_-^{(k)} = \sigma_-$

$f_k \leftarrow f_k + 1$

if $k \notin O(D)$

set $O(D) \leftarrow O(D) \cup \{k\}$

end

end

end

Move: Take $u \in U(0,1)$ and set

$\mathbf{E}_0 := \mathbf{E}_0 + (u\sigma_+ + (1-u)\sigma_-)\mathbf{S}$.

Set $iter \leftarrow iter + 1$

Until either *iter*=*MAX*.or $|O(D))| = 0.25 * q$.

It is usually the case that the percentage of contamination due to outliers is less than 25% [19]. We use this fact to give an upper bound on the size of $O(D)$. At the end of the *SCD* algorithm we obtain the set $O(D)$. The algorithm counts the number of times observation k is detected and records it as f_k. The *relative frequency of detection of detection* for k is defined by $f_k / iter$, where *iter* is the total number of iterations. The higher the value of *MAX* used, the greater the chance of detecting less extreme outliers. This measure is a *major contribution* of this paper.

The main computational effort of the algorithm is in the hitting step. Computing w^j in (3.2) requires *no* multiplications because of the special form of $\mathbf{S}$! Computing the two distances $\sigma_+^{(j)}$ and $\sigma_-^{(j)}$ requires p^2 multiplications to compute z^j (3.1), another p^2 to compute its norm, 2 divisions and 1 square root. Hence it takes $O(p^2)$ flops to compute $\sigma_+^{(n+1)}$ and $\sigma_-^{(n+1)}$. There are $4p^2$ flops needed to update $\mathbf{E}_0$. Hence, each *SCD* iteration is at most $O(np^2)$.

We now show give four examples of the method covering a variety of possible objectives in the detection of points on the outer shell.

These range from data cleanup before doing some other statistical analysis to simply identifying interesting points. The latter might be more typical in a data mining application. In any real problem variables are expected to be correlated, hence no attempt is made to orthogonalize. Similarly variability will be different across variables, hence no standardization is done.

In the figures, the points marked with an asterisk, *, belong to the outer shell and correspond to the detected necessary constraints of (2.6) and (2.7). The relative frequency of detection is recorded. The analytic center of the system (2.6) and (2.7) was used as the standing point. We used $MAX = 1000 * p$.

Example 1: This has 50 data points in IR^4, randomly generated from $N(0,1)$. The outer shell consists of the points $\{47,37,46,30\}$ and $\{0.4043,0.3176,0.2515,0.0266\}$ are the respective relative detection frequencies. The extreme outlier 47 has the largest relative detection frequency, and corresponds to the point $(1.2490,3.0868)^T$ in Figure 4.

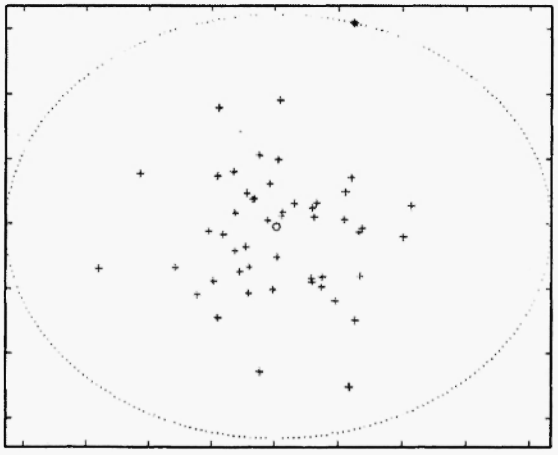

Figure 3 The figure shows that point marked * belongs to the outer shell since * can be separated from the other points by an ellipsoid passing through it and centered at the Fermat-Weber point $(0.0300,-0.0454)$

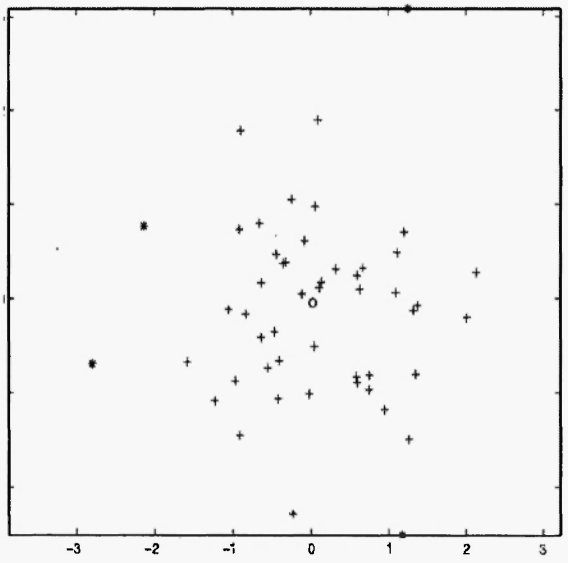

Figure 4 The points marked with a * belong to the outer shell. These are the extreme outliers.

Example 2: We consider the "so-called" Hawkins-Bradu-Kass data set constructed in [10] data with 75 points in IR^4. It is considered a benchmark in outlier detection methods; observations $1-14$ are outliers. Classical techniques only discover that 12 and 14 as outliers due to a masking effect caused by these points [3]. *SCD* found ten of the fourteen outliers in terms of the outer shell $\{12,14,7,3,10,6,2,4,13,5\}$, and the corresponding relative detection frequencies are $\{0.3131,0.3108,0.2098,0.0628,0.0505,$ $0.0354,0.0105,0.0044,0.0025,0.0002\}$ respectively. Observation 12 has the highest detection frequency.

The next two examples are instances of data mining. We make several incidental observations using the relative detection frequencies as our guide.

Example 3: Table 1 and Table 2 is taken from boxes of 20 brands of Kellogg's cereals [14].

SCD gives the outer shell $\{1,12,9,20,5\}$ with the relative detection frequencies $\{0.4194,0.2258,0.1613,0.1613,0.0323\}$ respectively. All Bran has the highest relative frequency of detection; its high fiber content protein and potassium (K) discriminates it.

Mueslix was second and Frosted Mini Wheats third (0 Sodium discriminates the latter.)

No.	Cereal
1	AllBran
2	AppleJacks
3	CornFlakes
4	CornPops
5	CracklinOatBran
6	Crispix
7	FrootLoops
8	FrostedFlakes
9	FrostedMiniWheats
10	FruitfulBran
11	JustRightCrunchyNuggets
12	MueslixCrispyBlend
13	NutNHoneyCrunch
14	NutriGrainAlmondRaisin
15	NutriGrainWheat
16	Product19
17	RaisinBran
18	RiceKrispies
19	Smacks
20	SpecialK

Table 1. A selected subset of Kellogg's cereals

No.	Cal	Protein	Fat	**Na**	Fiber	Carbs	Sugar	**K**
1	70	4	1	260	9	7	5	320
2	110	2	0	125	1	11	14	30
3	100	2	0	290	1	21	2	35
4	110	1	0	90	1	13	12	20
5	110	3	3	140	4	10	7	160
6	110	2	0	220	1	21	3	30
7	110	2	1	125	1	11	13	30
8	110	1	0	200	1	14	11	25
9	100	3	0	0	3	14	7	100
10	120	3	0	240	5	14	12	190
11	110	2	1	170	1	17	6	60
12	160	3	2	150	3	17	13	160
13	120	2	1	190	0	15	9	40
14	140	3	2	220	3	21	7	130
15	90	3	0	170	3	18	2	90
16	100	3	0	320	1	20	3	45
17	120	3	1	210	5	14	12	240
18	110	2	0	290	0	22	3	35
19	110	2	1	70	1	9	15	40
20	110	6	0	230	1	16	3	55

Table 2. Nutritional data of Kellogg's cereals in Table 1

Example 4: We consider the 1998 American League Pitching Final Statistics data from the Baseball Archive *http://baseball1.com/statistics/98alpit.html*. The data has 285 pitchers and 14 entries for each pitcher. *SCD* determined that 61 of the 285 pitchers are outliers. This means the contamination is less than 25% . The most interesting result comes from the relative detection frequencies that distinguish these 61 pitchers.

Table 3. 1998 American League pitchers with highest relative frequency of detection and a subset of their statistics.

Table 3 gives the ten players with the highest relative detection frequencies in the American League in 1998. Clemens and Wells all pitched in the All Star game. Scott Erickson (16 wins - 13 losses) led the league with the highest starts (36), innings pitched (251.1) and complete games (11). Roger Clemens (20-6) had a great year. Clemens had the best ERA (2.65), won the Cy Young Award and was the Triple Crown winner. David Wells (18-4) had the most shutouts (5) and the highest winning percentage (0.818). Not all the ten players were outstanding, e.g., John Burkett (9-13) gave up the most earned runs (123) and gave up 230 hits.

No.	Player	Team	Rel Freq	W	L	S	ERA
1	CLEMENS, ROGER	TOR	0.103	20	6	0	2.65
2	JOHNSON, RANDY	SEA	0.081	9	10	0	4.33
3	WELLS, DAVID	NY	0.060	18	4	0	3.49
4	ERICKSON, SCOTT	BAL	0.059	16	13	0	4.01
5	SAUNDERS, TONY	TB	0.054	6	15	0	4.12
6	BURKETT, JOHN	TEX	0.047	9	13	0	5.68
7	FINKLEY, CHUCK	ANA	0.046	11	9	0	3.99
8	ROGERS, KENNY	OAK	0.045	16	8	0	3.17
9	SELE, AARON	TEX	0.042	19	11	0	4.23
10	NAVARRO, JAIME	CHI	0.039	8	16	1	6.36

These examples show the advantage of *SCD* method as a tool for the detection of extreme outliers and for data mining. The algorithm can efficiently identify the most extreme outliers, especially those that are isolated. The method can be applied to any set of quantitative data. It has the disadvantage that it might miss some outliers and might identify valid observations that are extremal in some sense but not statistical outliers. This is true of any such tool. However, it has the advantage that more potential extreme outliers can be found than by using one step of ellipsoid peeling. It also seems to be more robust than ellipsoidal peeling since *SCD* is based on ellipsoids centered at the Fermat-Weber point.

Furthermore, our algorithm has the advantage that it gives as a by-product a measure of outlierness for each outlier found. This is meaningful, especially in data mining. Our measure of outlierness seems easier to compute than the local outlier factor (LOF) described in [7]. In contrast to *SCD*, the LOF based method of [7] does not explicitly say what data points are outliers.

Remark: The relative detection frequencies correspond to the fraction of the total solid angle in IR^N subtended at the standing point $\mathbf{E}_0$ by the portion of the surface of $\partial(R(D))$ due to the constraint generated by a point $a^k \in O(D)$. A point that is *more isolated* gives a larger relative detection frequency. It is a challenge to develop a complete analytical description of this boundary surface.

4. Application to D-optimal experimental designs

We consider the application of the *SCD* algorithm to the D- optimal experimental design problem. Here, the set D is a set of test vectors in IR^p and the point x^* is fixed at the origin in IR^p.

The goal of the D-optimal design problem is to furnish information on the linear model $z = \hat{\mathbf{A}}\beta + \omega$, where $\omega_j \in N(0,1)$, $\beta \in \mathrm{IR}^p$ is a vector of variables, $\hat{\mathbf{A}}^T = \{b^1, b^2, \cdots, b^m\}$, for $b^i \in D = \{a^1, a^2, \cdots, a^n\}$. The vectors a^j, $1 \le j \le n$, correspond to n possible settings for running an experiment.

The problem is to decide which a^j, from among a very large set, to use to define the columns of $\hat{\mathbf{A}}^T$ These columns are called *optimal test vectors*. We describe how *SCD* is used to select them. Let x_j be the fraction of the columns b^i of $\hat{\mathbf{A}}^T$ that equal a^j. The test vector a^j is chosen if and only if $x_j > 0$ in the optimal solution of the following optimization problem [25]:

$$\min \quad \log\det[\sum_{j=1}^{n} x_j a^j (a^j)^T]^{-1} \tag{4.1}$$

$$s.t. \sum_{j=1}^{n} x_j = 1, \quad 1 \le j \le n$$

$$x_j > 0, \; 1 \le j \le n$$

The experiment is carried out using the optimal design settings $\hat{x}$ that are estimated from the experiment by $\hat{x} = (\hat{\mathbf{A}}\mathbf{A})^{-1}\hat{\mathbf{A}}^T z$ [25].

The dual of the D-experimental design problem is to find the minimum volume ellipsoid $e_\Omega = \{y : \| \hat{\mathbf{E}} y \|^2 \le 1\}$ centered at the origin and containing D [22]. The problem can be written as the semidefinite optimization problem:

$$\max \quad \log\det(\mathbf{E}) \tag{4.2}$$

$$s.t. \begin{bmatrix} I & \mathbf{E}a^j \\ (\mathbf{E}a^j)^T & 1 \end{bmatrix} \succeq 0, \; 1 \le j \le n$$

$$E \succ 0$$

The matrix $\hat{\mathbf{E}}$ that defines e_Ω is given by the optimal solution of the above optimization problem. We note from [25] that the optimal test vectors a^j in the design problem are those that lie on the boundary of the minimum volume ellipsoid e_Ω.

Lemma 4.1 [25] Let $\hat{x}$ be the optimal solution of (4.1) and $\hat{\mathbf{E}}$ the optimal solution of the dual problem (4.2). Then, $x_j > 0$ only if a^j lies on the boundary of the minimum volume ellipsoid

e_Ω.

To apply *SCD*, choose x^* as the origin, since e_Ω is centered at the origin. In this case, we say that a point $a \in \mathrm{IR}^p$ belongs to the outer shell $O(D)$ if there is an ellipsoid centered at the origin that contains $D-\{a\}$ and whose boundary passes through a .

Corollary 4.1 If $\hat{x}$ is the optimal solution of (4.1), then $\hat{x}_j$ is positive only if a^j lies on the outer shell $O(D)$.

Proof: By Proposition 2.3, if a^j lies on the boundary of the minimum volume ellipsoid e_Ω, then a^j lies on $O(D)$. The rest follows from Lemma 4.1.

The corollary indicates that the optimal values a^j in the design problem (4.1) can be found from $O(D)$, without having to solve the dual problem (4.2) with the entire set D. *SCD* can be used to give a reduced subset given by the outer shell to solve the design problem. This can reduce the time required for solving the design problem when the *SCD* time to find the outer shell $O(D)$ is relatively low.

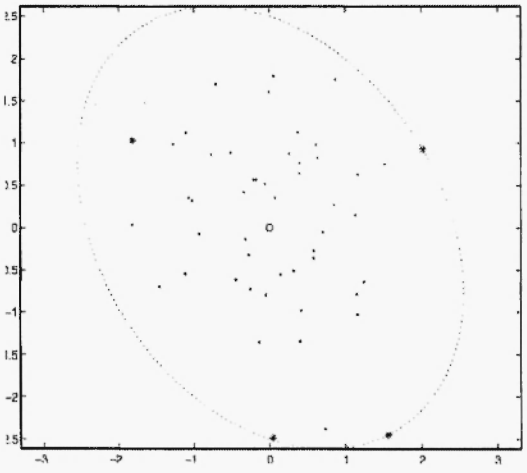

Figure 5. A D-optimal experiment design involving 50 test vectors in IR^2. The points in the outer shell are marked as *.

As an example, consider the D-optimal experiment design problem with the 50 test vectors in IR^2 given in [25]. The application of SDPSOL [26] to solve (4.2) yields the ellipsoid indicated in Figure 5. The four vectors represented by * are vectors detected by *SCD*, and these can be used to solve the optimization problem (4.1) instead of using the entire 50 vectors. The two test vectors lying on the ellipsoid are the optimal design test vectors.

5. Conclusion

We have introduced the *SCD* Monte Carlo method for detecting extreme outliers in a multivariate data set D. The extreme outliers are defined in terms of a subset called the outer shell, and correspond to those points whose constrained in IR^N can be viewed from an interior point. The problem of finding the outer shell is shown to be equivalent to the problem of detecting necessary constraints in a semidefinite program.

Each iteration of the algorithm is polynomial of order $O(np^2)$. This method has the new feature that it gives a measure of the "outlierness". This is a correlative percentage measure of a data point being an isolated extreme outlier in terms of the relative frequencies of detection of the associated constraint in IR^N. This can be used in digital image processing to find points on the boundary of a convex image that are essential.

Furthermore, we describe how the *SCD* algorithm can be used to reduce the solution time of the D-optimal experimental design problem where D contain possible test vectors. The algorithm finds a reduced subset of D that can be used to more efficiently find the optimal test vectors.

There are several interesting open questions that remain unsolved. First, we would like to better understand the robustness properties of the *SCD*. Second, suppose that two constraints $\mathbf{A}^{(1)}(\mathbf{E}) \succeq 0$ and $\mathbf{A}^{(2)}(\mathbf{E}) \succeq 0$ are necessary;. what necessary and sufficient conditions would guarantee that they intersect on some open subset of the boundary $\partial(R(D))$? Finally, if after a given number of Monte Carlo iterations a point is found to have zero detection frequency, what is the probability that it is not an extreme outlier?

Acknowledgements: The authors would like to thank G. Bell for his many helpful suggestions on improving the early drafts of this paper.

References

1. J. Adrover and V. Yohai, "Projection estimates of multivariate location", *The Annals of Statistics*, vol. 30, 2002, pp. 1760-1781.
2. F. Alizadeh, "Interior Point Methods in Semidefinite Programming with Applications to Combinatorial Optimization", *SIAM J. Optim.*, Vol. 5, no. 1, 1995, pp. 13-51.
3. Anna Bartkowiak and Adam Szustalewicz, "The Grand Tour as a Method for Detecting Multivariate Outliers", *Machine Graphics & Vision*, Vol. 6, 1997, pp. 487-505.
4. V. Barnett and T. Lewis, "*Outliers in Statistical Data*", John Wiley & Sons, 1994
5. C. Belisle, A. Boneh and R. Caron, "Convergence properties of hit-and-run samplers", *Stochastic Models*, 14(4), 1998.
6. M. J. Best, and K. Ritter, "*Linear Programming: Active Set Analysis and Computer Programs*", Prentice-Hall, Englewood Cliffs, New Jersey, 1985.
7. M. Brueening, H. P. Kriegel, R. Ng and J. Sander, "LOF: Identifying Density-Based Local Outliers", *ACM SIGMOND Conference Proceedings*, 2000
8. R. J. Caron, Private Communication with S. Jibrin, 2003.
9. Z. Chen and D. Tyler, "The influence function and maximum bias of Tukey's median", *The Annals of Statistics*, vol. 30, 2002, pp. 1737-1759.
10. D. M. Hawkins, D. Bradu and G. V. Kass, "Location of several outliers in multiple regression using elemental subsets", *Technometrics*, vol. 26, 1984, pp. 197-208.
11. R. A. Horn and C. R. Johnson, "*Matrix Analysis* ", Cambridge University Press, Cambridge 1993.
12. S. Jibrin and I. S. Pressman, "Monte Carlo Algorithms for the Detection of Nonredundant Linear Matrix Inequality Constraints", *Int. J. Non. Sci. Num. Simulation*,2(2),2001, 139-153.
13. W. Jin, A. K. H. Tung and J. Han, "Mining top-n local outliers in large databases", *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, California , 2001, pp. 293-298.
14. R. A. Johnson and D. W. Wichern, "*Applied Multivariate Statistical Analysis*", Prentice-Hall Inc., Upper Saddle River, 2002.
15. E. M. Knorr, "*On Digital Money and Card Technologies*", Technical Report 97-02, University of British Columbia, 1997.
16. R. Liu, J. Parelius and K. Singh, "Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion)", *The Annals of Statistics*, vol. 27, 1999, pp. 783-858.
17. D. M. Rocke and D. L. Woodruff, "Identification of Outliers in Multivariate Data", *Journal of the American Statistical Association*, vol. 91, 1996, pp. 1047-1061
18. P. J. Rousseeuw and Annick M. Leroy, "*Robust Regression and Outlier Detection*", Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons. Inc., New York, 1987.
19. P. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, vol. 41, 1999, pp. 212-223.
20. P. Rousseeuw and B. van Zomeren, "Unmasking Multivariate Outliers and Leverage Points", *J. the American Statistical Association*, vol. 85, 1990, pp. 633-639.
21. D. M. Titterington, "Estimation of Correlation Coefficients by Ellipsoidal Trimming", *Appl. Statist.*, vol. 27, no. 3, 1978, pp. 227-234.
22. D. M. Titterington, "Optimal design: some geometrical aspects of D-optimality", *Biometrika*, vol. 62, 1975, pp. 313-320.
23. J. W. Tukey, "Mathematics and picturing data", *Proc. Int. Congress of Mathematicians, Vancouver*, vol 2. pp. 523-531.
24. L. Vandenberghe and S. Boyd, "Semidefinite Programming", *SIAM Review*, 38, 1996, 49-95.
25. L. Vandenberghe, S. Boyd and S. -P. Wu, "Determinant maximization with linear matrix inequality constraints", *SIAM Journal on Matrix Analysis and Applications*, vol. 19, 1998, pp. 499-533.
26. S. P. Wu and S. Boyd, "*SDPSOL - A Parser/Solver for Semidefinite Programming and Determinant Maximization Problems with Matrix Structure*", User's Guide, Version Beta, May 1996.
27. Y. Zuo and R. Serfling, "Nonparametric notions of multivariate 'scatter measure' and 'more scattered' based on statistical depth functions", *Journal of Multivariate Analysis*, vol. 75, 2000, pp. 62-78.