British Journal of Mathematical and Statistical Psychology (2000), 53, 255–273 *Printed in Great Britain*

A local influence approach to identifying multiple multivariate outliers

Wai-Yin Poon* and **Shing-Fong Lew**

The Chinese University of Hong Kong, Hong Kong

Yat Sun Poon

University of California at Riverside, USA

We make use of Cook's local influence approach and its recent modification by Poon and Poon to develop measures for detecting multivariate outliers. The motivation and the foundation of the theory are geometrical and are different from classical approaches; however, whilst the proposed measure exhibits a form similar to those in the literature, it still has a considerable advantage in having transformed the classical measures to the unit interval. The new approach unifies outlier identification measures using geometrical concepts. It involves no distributional assumption or large-sample properties, and allows the flexibility of identifying outliers with respect to different metrics. The approach therefore provides a valid reason for using the various measures in complicated situations, such as in non-normal cases and in small-sample problems.

1. Introduction

This paper concerns the problem of identifying multiple outliers in multivariate data. Let $\mathbf{x}_j$, $j = 1, \ldots, n$, be p-dimensional data points. Let $\boldsymbol{\mu}$ be the local parameter and $\mathbf{V}$ be a chosen metric; the distance between the jth observation $\mathbf{x}_j$ and $\boldsymbol{\mu}$ relative to $\mathbf{V}$ is given by

$$D_j(\boldsymbol{\mu}, \mathbf{V}) = \sqrt{(\mathbf{x}_j - \boldsymbol{\mu})'\mathbf{V}(\mathbf{x}_j - \boldsymbol{\mu})}, \tag{1}$$

where $\mathbf{V}$ is a positive definite symmetric matrix. If $\{\mathbf{x}_j\}$ is a random sample with known variance covariance matrix $\boldsymbol{\Sigma}$, a common practice is to put $\mathbf{V} = \boldsymbol{\Sigma}^{-1}$. Under the multivariate normal assumption, $\boldsymbol{\mu}$ is usually replaced by the sample mean $\overline{\mathbf{X}}$ and $\boldsymbol{\Sigma}$ by the sample covariance matrix $\mathbf{S}$, and (1) reduces to the Mahalanobis distance

$$MD_j = D_j(\overline{\mathbf{X}}, \mathbf{S}^{-1}) = \sqrt{(\mathbf{x}_j - \overline{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{x}_j - \overline{\mathbf{X}})}. \tag{2}$$

Various methods based on the Mahalanobis distance have been proposed in the literature for outlier identification.

Rousseeuw & van Zomeren (1990) suggest using a robust distance RD_j which replaces $\overline{\mathbf{X}}$

* Requests for reprints should be addressed to Dr Wai-Yin Poon, Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: wypoon@hp735.sta.cuhk.edu.hk).

and $\mathbf{S}$ in (2) by the minimum volume ellipsoid estimators introduced by Rousseeuw (1985). Hadi (1992) uses a robust estimator based on the median to provide an initial ordering of the data points, and develops a forward-stepping algorithm. He divides the data into two sets: a basic subset of good observations and a non-basic subset containing the remaining observations. The basic subset is updated in such a way as to exclude outliers. The final non-basic subset of observations is declared as outlying. A modification which increases the power of this method is given by Hadi (1994). Fung (1993), on the other hand, advocates the importance of confirmatory analysis. Let I be the index set of m outliers detected by RD_j; each observation in I is added back to the reduced sample of size $n - m$, and diagnostic measures are computed. Observations are deleted from the reduced sample if they are outliers and are restored if they are not. Atkinson & Mulira (1993) consider a forward procedure. They compute Mahalanobis distances based on the mean and covariance of a small subset of m observations, intended to be outlier-free; the $m + 1$ observations to be used for the next set of distances are those with the $m + 1$ smallest distances. A stalactite plot is constructed to indicate cases with large distances in different subsets. Those cases that continue to have large distances when the pattern of the stalactite plot is stabilized are considered as outliers. Atkinson (1994) suggested the following refinements to the procedure: use several starting points to increase the probability of obtaining initial distances based on an outlier-free set; correct the large number of outliers for small m by a simulation-based normalization; and use robust distances to replace Mahalanobis distances. Recently, Rocke & Woodruff (1996) studied the nature of multivariate outliers. They concluded that outliers with the same shape as the main data are the hardest to find, and that shift outliers provide a reasonable test bed for multivariate outlier detection. They proposed a hybrid algorithm to detect outliers.

The basic building block of these outlier identification procedures is the Mahalanobis distance or its robust version. These distances are used in the generic sense of a quadratic form distance, and are usually compared to the chi-squared or the F distribution. In this paper, with a completely different motivation, we propose a method to develop measures for detecting multivariate outlyingness. Using Cook's (1986) local influence approach and its recent modification by Poon & Poon (1999), the conformal normal curvature (Poon & Poon, 1999) of an appropriately selected influence graph is used to construct an outlying measure. The measure therefore has a good geometrical basis. It will be seen that the proposed measure has a close relationship with various classical outlier measures, hence provides a unifying concept for outlier measures. The proposed measure also inherits all of the nice features of the conformal normal curvature. In particular, it assumes values in the unit interval and hence is easier to judge its magnitude. If objectivity is desired, a geometrically orientated method can be used to determine a reference constant to judge outlyingness. This new approach is based on geometrical reasoning rather than distributional properties and large-sample theories. As a result, it allows various classical measures to be applied in non-classical situations, such as in non-normal cases and for small-sample problems.

In the next section, we summarize the local influence approach and its recent development. In particular, we discuss why the classical local influence approach is difficult to apply to the present problem and how the recent development of Poon & Poon (1999) improves the situation. Section 3 develops the measure for detection of multivariate outliers, provides a reference constant to judge outlyingness, discusses the choices of $\mathbf{V}$, and addresses the relation between the proposed measure and the classical measures. In Section 4, we summarize the results of the analyses on one constructed data set with large dimension

($p = 40$) and several data sets reported in the literature. In Section 5, the paper is concluded with a discussion.

2. The local influence approach and its recent development

Statistical models usually involve some degree of approximation. It is therefore important to study the influences on key results of an analysis under a minor modification of such approximate description. The local influence approach proposed by Cook (1986) is a unified approach for assessing the influence of minor perturbations of a statistical model. Such perturbations can apply to model assumptions, to data values and to case weights. This geometrically orientated approach employs certain elementary ideas from differential geometry to develop influence measures and the approach can be used to handle a variety of problems by using appropriately chosen perturbations. Typical examples are its applications to the diagnostics and influence analysis in mixed-model analysis of variance (Beckman, Nachtsheim & Cook, 1987), in regression transformation (Lawrance, 1988; Tsau & Wu, 1992), in the generalized linear model (Thomas & Cook, 1989, 1990), in nonlinear regression (St Laurent & Cook, 1993), in structural equation models (Lee & Wang, 1996; Poon, Wang & Lee, 1999) and in principal components analysis (Shi, 1997).

Let $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta} \mid \boldsymbol{\omega})$ be the log-likelihoods for a postulated model and a perturbed model, where $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ is an $n \times 1$ vector in $\Omega \in \mathbb{R}^n$, and Ω represents the set of relevant perturbation. It is assumed that there is an $\boldsymbol{\omega}_0$ such that $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \boldsymbol{\omega}_0)$ for all $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ be the maximum likelihood estimator under $L(\boldsymbol{\theta} \mid \boldsymbol{\omega}_0)$ and $L(\boldsymbol{\theta} \mid \boldsymbol{\omega})$ respectively. Cook (1986) defines the likelihood displacement

$$f(\boldsymbol{\omega}) = 2(L(\hat{\boldsymbol{\theta}}|\boldsymbol{\omega}_0) - L(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_0)). \tag{3}$$

A straight line in Ω passing through $\boldsymbol{\omega}_0$ is defined by $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_0 + a\mathbf{l}$, where a is a scalar, and $\boldsymbol{\omega}_0$ and $\mathbf{l}$ are fixed column vectors in $\mathbb{R}^n$. Cook suggests using the normal curvature $C_{\mathbf{l}}$ of the graph of the likelihood displacement function along a direction $\mathbf{l}$ at the optimal point $\boldsymbol{\omega}_0$ to study characteristics of the influence graph. For the computation of $C_{\mathbf{l}}$, Cook (1986, eq. (11)) further deduces that

$$C_{\mathbf{l}} = -2(\mathbf{l}'\ddot{\mathbf{F}}\mathbf{l})_{|\boldsymbol{\omega}=\boldsymbol{\omega}_0}, \tag{4}$$

where $\ddot{\mathbf{F}}$ is the $n \times n$ matrix with elements $\partial^2 L(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})/\partial\omega_i\partial\omega_j$. Let $\boldsymbol{\Delta}$ be the $p \times n$ matrix with elements

$$\boldsymbol{\Delta}_{ij} = \frac{\partial^2 L(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\theta_i\partial\omega_j}, \tag{5}$$

and $\ddot{\mathbf{L}}$ be the $p \times p$ matrix with elements

$$\ddot{L}_{ij} = \frac{\partial^2 L(\theta)}{\partial\theta_i\partial\theta_j}, \tag{6}$$

both evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$; then (4) reduces to (Cook, 1986 eq. (16))

$$C_{\mathbf{l}} = -2(\mathbf{l}'\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\mathbf{l})_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}. \tag{7}$$

Large values of $C_{\mathbf{l}}$ indicate strong local influence, that is, the likelihood displacement has substantial local change along the direction $\mathbf{l}$; and the group of perturbation parameters with

large coefficients in $\mathbf{l}$ exhibits strong joint influence, that is, the perturbation is sensitive to these perturbation parameters. Cook also proposes paying special attention to $C_{\max} = \max_{\mathbf{l}} C_{\mathbf{l}}$ and the corresponding direction $\mathbf{l}_{\max}$. These are respectively the maximum eigenvalue and the corresponding eigenvector of the matrix $-\ddot{\mathbf{F}} = -\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}$. Note that it is difficult to obtain the explicit expression of $\mathbf{l}_{\max}$ in general situations; and hence the eigenvector $\mathbf{l}_{\max}$ must be computed numerically. In problems with large n, the computation of $\mathbf{l}_{\max}$ becomes intractable. Furthermore, the normal curvature may take on any value and it is not invariant under a uniform change of scale. These issues remain unresolved and hinder the applicability of the approach. Therefore, although the local influence approach has been applied successfully in various statistical analyses, its use in the detection of multivariate outliers has not been studied.

Poon & Poon (1999) introduced the conformal normal curvature which resolves some of the above-mentioned issues. The conformal normal curvature is a one–one function of the normal curvature and assumes values in the unit interval. In a direction $\mathbf{l}$ at a critical point $\boldsymbol{\omega}_0$, it is given by (Poon & Poon, 1999, eq. (2.11))

$$B_{\mathbf{l}} = -\frac{\mathbf{l}'\ddot{\mathbf{F}}\mathbf{l}}{\sqrt{\operatorname{tr}\ddot{\mathbf{F}}^2}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\boldsymbol{\omega}=\boldsymbol{\omega}_0}} = -\frac{(\mathbf{l}'\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\mathbf{l})}{\sqrt{\operatorname{tr}(\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta})^2}}|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \boldsymbol{\omega} = \boldsymbol{\omega}_0. \tag{8}$$

Let $\mathbf{E}_j$, $j = 1, \ldots, n$, be $n \times 1$ vectors of the standard basis; $\mathbf{E}_j$ is called the jth basic perturbation vector of the perturbation space. For simplicity of presentation, the normal curvature $C_{\mathbf{E}_j}$ and the conformal normal curvature $B_{\mathbf{E}_j}$ along the j th basic perturbation vector $\mathbf{E}_j$ are denoted by C_j and B_j respectively. Poon & Poon (1999, Theorem 4) developed a relation between B_j, $j = 1, \ldots, n$, and the influential eigenvector directions at $\boldsymbol{\omega}_0$; they established as a special case the relation between $\mathbf{l}_{\max}$ and B_j. As a result, some of the information carried in $\mathbf{l}_{\max}$ can be revealed by the B_j or C_j. More specifically, if $C_{\max}$ is large and the magnitude of the j th coefficient in $\mathbf{l}_{\max}$ is large, C_j will also be large. Note that C_j is the j th diagonal element of the matrix $-\ddot{\mathbf{F}}$, hence both C_j and B_j can be computed very easily when $-\ddot{\mathbf{F}}$ is available. In some situations, such as in the development to appear in Section 3, diagonal elements of $-\ddot{\mathbf{F}}$ can be obtained explicitly. Using the concept of mean curvature, Poon & Poon (1999) also note that if the contribution of all B_j is uniform, then each is equal to

$$b = -\frac{\operatorname{tr}(\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta})}{n\sqrt{\operatorname{tr}(\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta})^2}}. \tag{9}$$

Therefore, they recommend referring to b when judging the largeness of B_j. For example, one can compare B_j to $2b$.

Neither the work of Cook (1986) nor that of Poon & Poon (1999) is restricted to the likelihood displacement. The above results are valid for other objective functions $f(\boldsymbol{\omega})$ if the gradient vector of $f(\boldsymbol{\omega})$ vanishes at $\boldsymbol{\omega}_0$; and if the Hessian matrix $\mathbf{H}_f = (\partial^2 f/\partial\omega_i\partial\omega_j) = -2\ddot{\mathbf{F}}$ is positive semi-definite at $\boldsymbol{\omega}_0$.

3. Measure of Outlyingness

We apply the local influence approach to develop an outlier measure. In the light of the work of Rocke & Woodruff (1996), we adopt the shift outlier as a reasonable target. We assume an

appropriate metric defined by $\mathbf{V}$ is known, or a good estimate of it is available. The issue of the choice of $\mathbf{V}$ will be further addressed in Section 3.4.

3.1. Location estimate, case-weights perturbation and displacement

Let $\mathbf{V}$ be a known metric; the location estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ relative to the metric $\mathbf{V}$ is obtained by maximizing the function

$$L(\boldsymbol{\mu}) = -\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$= -\sum_{i=1}^{n}\langle(\mathbf{x}_i - \boldsymbol{\mu}), \mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})\rangle, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ represents the inner product. Consider the case-weights perturbation given by

$$L(\boldsymbol{\mu}|\boldsymbol{\omega}) = -\sum_{i=1}^{n}\omega_i(\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$= -\sum_{i=1}^{n}\omega_i\langle\mathbf{x}_i - \boldsymbol{\mu}, \mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})\rangle, \tag{11}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$. When $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1} = (1, \ldots, 1)'$, $L(\boldsymbol{\mu}|\boldsymbol{\omega})$ reduces to $L(\boldsymbol{\mu})$; if $\omega_j = 0$, the perturbation scheme becomes the deletion of case j. Let $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ be the quantities that maximize (10) and (11), respectively. It can be shown (Appendix A) that

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}} = \frac{\sum_{i=1}^{n}\omega_i\mathbf{x}_i}{\sum_{i=1}^{n}\omega_i}. \tag{12}$$

When $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$, we have

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i. \tag{13}$$

If an observation is outlying in location, its influence on $\hat{\boldsymbol{\mu}}$ is large. To assess the influence of an individual case on $\hat{\boldsymbol{\mu}}$, we consider the displacement defined by

$$f(\boldsymbol{\omega}) = 2(L(\hat{\boldsymbol{\mu}}|\boldsymbol{\omega}_0) - L(\boldsymbol{\mu}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_0)). \tag{14}$$

Note that $f(\boldsymbol{\omega})$ achieves its minimum 0 at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, and the Hessian matrix $\mathbf{H}_f = (\partial^2 f/\partial\omega_i\partial\omega_j)$ at $\boldsymbol{\omega}_0$ is positive semi-definite. For a given $\boldsymbol{\omega}$, if $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ differs substantially from $\hat{\boldsymbol{\mu}}$, the magnitude of $f(\boldsymbol{\omega})$ is large. In particular, if case j is a shift outlier, then slightly perturbing $\boldsymbol{\omega}_j$ in $\boldsymbol{\omega}$ from $\boldsymbol{\omega}_0 = \mathbf{1}$ will induce a substantial change from $\hat{\boldsymbol{\mu}}$ to $\boldsymbol{\mu}_{\boldsymbol{\omega}}$, which in turn will lead to a substantial change in $f(\boldsymbol{\omega})$. Therefore, a graph of $f(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contains essential information on the influence of the observations. More specifically, from Section 2, we know that the conformal normal curvature B_j, or its equivalent, the normal curvature C_j, for the displacement function in (14) can be used to assess the outlyingness of case j. If B_j is large, the jth observation is influential in the estimation of the location, and hence is a location outlier. If several cases possess large value in their B_j, they simultaneously have substantial influences on the estimation of the location, and hence are identified as a set of multiple outliers.

3.2. The conformal normal curvature as a measure of outlyingness

To compute B_j, we need to find the matrices $\boldsymbol{\Delta}$ and $\ddot{\mathbf{L}}$ in (5) and (6), respectively. Define the $p \times n$ matrix $\mathbf{Y} = (\mathbf{x}_1 - \hat{\boldsymbol{\mu}}, \ldots, \mathbf{x}_n - \hat{\boldsymbol{\mu}})$. We show in Appendices B and C that

$$\ddot{\mathbf{L}} = -2n\mathbf{V}, \quad \boldsymbol{\Delta} = 2\mathbf{V}\mathbf{Y}. \tag{15}$$

Therefore, $\ddot{\mathbf{F}}$ in (4) is given by

$$\begin{aligned} \ddot{\mathbf{F}} &= \boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta} \\ &= \frac{-2}{n}\mathbf{Y}'\mathbf{V}\mathbf{Y} \end{aligned} \tag{16}$$

(see also (7)). Also from (4), we obtain

$$C_{\mathbf{E}_j} = C_j = \frac{4}{n}(\mathbf{x}_j - \hat{\boldsymbol{\mu}})'\mathbf{V}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}). \tag{17}$$

From (17), we see that the usual distance measure given in the form of (1) has a nice geometrical interpretation. Its square can be considered as a normal curvature.

For easy judgment of the magnitude, we further transform the normal curvature onto the conformal normal curvature, which is a one–one transformation. From (16), we have

$$\sqrt{\operatorname{tr}(\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta})^2} = \sqrt{\operatorname{tr}(\ddot{\mathbf{F}}^2)} = \sqrt{\sum_k \sum_l \ddot{F}_{kl}^2} = \frac{2}{n}\sqrt{\sum_k \sum_l ((\mathbf{x}_k - \hat{\boldsymbol{\mu}})'\mathbf{V}(\mathbf{x}_l - \hat{\boldsymbol{\mu}}))^2}; \tag{18}$$

together with (8) and (17), we obtain

$$B_{\mathbf{E}_j} = B_j = \frac{(\mathbf{x}_j - \hat{\boldsymbol{\mu}})'\mathbf{V}(\mathbf{x}_j - \hat{\boldsymbol{\mu}})}{\sqrt{\Sigma_k \Sigma_l ((\mathbf{x}_k - \hat{\boldsymbol{\mu}})'\mathbf{V}(\mathbf{x}_l - \hat{\boldsymbol{\mu}}))^2}}. \tag{19}$$

A very nice feature of B_j is that it assumes value in the unit interval. It is easier to interpret its magnitude. We consider case j as an outlier if the value of B_j or C_j is large. The next question is how large should be considered large.

3.3. A reference constant to judge largeness

From (19), we see that B_j is a normalized version of classical outlying measure. Therefore, we may follow common practice and compare B_j, or functions of B_j, with the quantiles of known distributions, such as the chi-squared or the F distribution. However, since normal curvature is a geometrical concept, we prefer to use the geometrically orientated method proposed by Poon & Poon (1999) to identify abnormal cases. They point out that if the contribution of all B_j is uniform, then each is equal to the constant

$$\begin{aligned} b &= \frac{\operatorname{tr}(-\ddot{\mathbf{F}})}{n\sqrt{\operatorname{tr}(\ddot{\mathbf{F}}^2)}} \\ &= \frac{\Sigma_i(\mathbf{x}_i - \boldsymbol{\mu})'\mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})}{n\sqrt{\Sigma_k \Sigma_l ((\mathbf{x}_k - \boldsymbol{\mu})'\mathbf{V}(\mathbf{x}_l - \boldsymbol{\mu}))^2}}. \end{aligned} \tag{20}$$

Therefore, cases with B_j values much greater than b may be considered as outliers. For example, one can use $2b$ as a handy benchmark. Furthermore, according to Cook (1986),

Table 1. An artificial data set

Case	1	2	3	4	5	6	7	8	9	10
$\mathbf{X}_1$	1.00	1.01	1.00	1.00	1.01	1.01	1.00	1.00	1.03	1.01
$\mathbf{X}_2$	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	5.00	10.00

Poon & Poon (1999) and the discussion in Section 3.1, we conclude that the set of cases with large B_j is a set of multiple outliers.

The availability of a benchmark is nice because it facilitates automation for large problems. However, we have no intention of recommending a mechanical use of the benchmark. In problems of manageable size, an index plot can always provide helpful information.

3.4. The choice of $\mathbf{V}$

Although the chosen metric $\mathbf{V}$ affects the objective function given in (10), the location estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is independent of $\mathbf{V}$. However, $\mathbf{V}$ is involved in (19) and plays an important role in identifying outliers. Different cases would be regarded as outliers with different choices of $\mathbf{V}$, and the choice depends on the objective of a particular study.

To illustrate, consider the artificial data set given in Table 1. The $\mathbf{X}_1$ observation of case 9 and the $\mathbf{X}_2$ observation of case 10 deviate from those of the other cases. When the units of $\mathbf{X}_1$ and $\mathbf{X}_2$ are essentially the same, for example when the cases are spots on a map, one will consider case 10 as an outlier. And an appropriate choice of $\mathbf{V}$ is $\mathbf{V} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix of dimension p (= 2 in this example). The values of B_j with $\mathbf{V} = \mathbf{I}_p$ are given in the second row of Table 2. Compared to $2b = 0.2$, cases 10 and 1 are identified as outliers. B_9 is very small, and is definitely an inlier under this metric.

On the other hand, when the observations are assumed to come from a bivariate normal population, we may wish to measure distance with the effect of dispersions being taken into account. That is, for the data set, we may wish to flag case 9 as an outlier. To achieve this objective, a natural choice is $\mathbf{V} = \mathbf{S}^{-1}$. This is the metric that defines the Mahalanobis distance. The values of B_j with $\mathbf{V} = \mathbf{S}^{-1}$ are presented in the third row of Table 2. B_9 is larger than $2b = 0.283$ and case 9 is identified as an outlier. Under this metric, the outlying nature of case 10 and case 1 is not as noticeable as case 9.

Although we set $\mathbf{V} = \mathbf{S}^{-1}$ in the example, we have no intention of recommending the use of $\mathbf{S}^{-1}$ in general. It is well known that $\mathbf{S}$ as an estimate of $\boldsymbol{\Sigma}$ is highly affected by the outlying observations that are supposed to be identified, and a sensible alternative is to replace $\mathbf{S}$ by $\mathbf{R}$, where $\mathbf{R}$ is a robust estimate of the covariance matrix $\boldsymbol{\Sigma}$. Various methods are available for finding $\mathbf{R}$ (see Rousseeuw, 1985; Rousseeuw & van Zomeren, 1990; Hadi, 1992; Hawkins &

Table 2. B_j for the artificial data set

$\mathbf{V} \backslash j$	1	2	3	4	5	6	7	8	9	10	$2b$
$\mathbf{I}_p$	0.244	0.139	0.064	0.018	0.000	0.012	0.052	0.122	0.000	0.348	0.200
$\mathbf{S}^{-1}$	0.200	0.113	0.080	0.051	0.008	0.015	0.088	0.014	0.468	0.248	0.283
$\mathbf{R}^{-1}$	0.183	0.085	0.085	0.060	0.009	0.016	0.079	0.117	0.546	0.198	0.275

Simonoff, 1993; and the nice review given by Rocke & Woodruff, 1996). For the numerical examples given in this paper, $\mathbf{R}$ is obtained by the method proposed by Hadi (1992, Appendix A). The result of $\mathbf{V} = \mathbf{R}^{-1}$ for the data set in Table 1 is presented in Table 2; the outlying nature of case 9 is prominent.

This example demonstrates the effectiveness and flexibility of the proposed procedure. We see that the proposed measure works nicely in a small sample. Moreover, when we have different objectives in mind, we consider different metrics and use different $\mathbf{V}$s; the proposed method is valid for any $\mathbf{V}$ which is positive definite and can identify the outliers effectively with respect to the chosen metric.

3.5. A unifying concept

The proposed geometrical approach provides a unifying concept for constructing measures of outlyingness. If we let $\mathbf{V} = \mathbf{S}^{-1}$ in our development, the B_j become the Mahalanobis distances. If we let $\mathbf{V} = \mathbf{R}^{-1}$, we obtain a distance which combines non-robust location and robust scale estimates. In the literature, there are many outlier identification procedures which are based on distance measures with robust location and robust scale estimates. The proposed procedure can be modified slightly to obtain such measures: We can choose $\mathbf{V}$ equal to the robust scale estimate and choose the null perturbation, that is, the weights in $\boldsymbol{\omega}_0$, appropriately to obtain a robust location estimate.

The entire procedure is based on geometrical reasoning; no distributional assumption or large-sample properties are involved. Therefore, unlike other classical statistical procedures with various assumptions, it is difficult to provide the statistical properties of the proposed measure in general. However, if the distributions of the variables are known or if the sample size is large, for example when B_j is equivalent to the Mahalanobis distance, it is possible to establish the statistical properties of the proposed measure by making use of the equivalence. Nevertheless, the advantage of the proposed procedure is that without knowing the statistical properties of the measure which are only available with certain imposed assumptions, the measure and its benchmark can still be applied on geometrical grounds. In other words, this procedure can be used in situations where assumptions required by classical statistical approaches are not satisfied, including non-normal cases or small-sample problems.

4. Examples

4.1. Example 1: High-dimensional artificial data set

The first example is based on a constructed data set with $p = 40$ and $n = 200$. The sample size is small relative to the large dimension ($n/p = 5$). The purpose of this example is to demonstrate that the proposed method works nicely with large dimension and a relatively small sample. Observations were generated from a multivariate standard normal distribution and the last 10 cases (cases 191–200) were constructed to be outliers by adding a value $c = 2.7092$ to each component. Note that $c = 2\sqrt{\chi^2_{40,0.999}/40}$, where $\chi^2_{40,0.999}$ is the 0.1% probability point of the chi-squared distribution with 40 degrees of freedom. This method of constructing outliers was used by Rocke & Woodruff (1996). They regard the outliers created as 'close shift' outliers. For $\mathbf{V} = \mathbf{I}_p$, $\mathbf{V} = \mathbf{R}^{-1}$, and $\mathbf{V} = \mathbf{S}^{-1}$, B_j and $2b$ were computed. Figure 1 presents the index plots of the B_j. The outlying observations are nicely identified and

there is no misclassification when $\mathbf{V} = \mathbf{I}_p$ and $\mathbf{V} = \mathbf{R}^{-1}$. However, no outlying observations can be identifed when $\mathbf{V} = \mathbf{S}^{-1}$.

4.2. Example 2: Hawkins et al.*'s data set*

The second data set was constructed by Hawkins, Bradu & Kass (1984) and has been analysed many times in the literature (see Rousseuw & Leroy, 1987; Rousseeuw & van Zomeren, 1990; Atkinson & Mulira, 1993; Fung, 1993; Hadi & Simonoff, 1993; Rocke & Woodruff, 1996). The data set contains three explanatory variables, one dependent variable, and 75 observations. The first 14 observations are leverage points and the effects of cases 11–14 are stronger. We use the three explanatory variables only. Figure 2 gives the index plots of B_j for three different choices of $\mathbf{V}$: $\mathbf{V} = \mathbf{I}_p$, $\mathbf{V} = \mathbf{R}^{-1}$ and $\mathbf{V} = \mathbf{S}^{-1}$. When $\mathbf{V} = \mathbf{I}_p$ and $\mathbf{V} = \mathbf{R}^{-1}$, the values of $B_j, j = 1, \ldots, 14$, are greater than $2b$ and the cases are considered as outliers. The values of B_j fall into three groups. Cases 11–14 form the most extreme group, cases 1–10 form the second extreme group, and the remaining observations form another group. The structure of the data has been nicely revealed. When $\mathbf{V} = \mathbf{S}^{-1}$, we see from Fig. 2(c) that the pattern in the index plot is not as clear as that for $\mathbf{V} = \mathbf{I}_p$ or $\mathbf{V} = \mathbf{R}^{-1}$.

4.3. Example 3: Brain and body weight data set

The third data set is the brain and body weight data set (in logarithms to base 10) (Rousseeuw & Leroy, 1987, p. 58) which has been analysed, for example, by Rousseeuw & van Zomeren (1990) and Atkinson & Mulira (1993). Using a robust version of the Mahalanobis distance, Rousseeuw & van Zomeren (1990) identify cases 25, 6, 16, 14 and 17 in this order as outlying observations, where the effect of case 17 is marginal. The analyses of Atkinson & Mulira (1993) provide a similar conclusion. Their stalactite plot clearly indicates cases 25, 6 and 16 as the most extreme outliers, and there is evidence that cases 14 and 17 may not agree with the bulk of the data. We reanalysed the data set by the proposed method with $\mathbf{V} = \mathbf{I}_p$, $\mathbf{V} = \mathbf{R}^{-1}$ and $\mathbf{V} = \mathbf{S}^{-1}$, respectively. The index plots of B_j are given in Fig. 3. When $\mathbf{V} = \mathbf{R}^{-1}$ (Fig. 3(b)), we identify cases 25, 6 and 16 as outliers. Cases 14, 20 and 17 are the next largest in influence but they are not larger than $2b$. When $\mathbf{V} = \mathbf{S}^{-1}$ (Fig. 3(c)), cases 25, 6, 16 and 20 are identified; however, the influences of cases 14 and 17 are less prominent. The outliers identified by the metric $\mathbf{V} = \mathbf{I}_p$ (Fig. 3(a)) are different from those identified by the metric $\mathbf{V} = \mathbf{R}^{-1}$. Cases 20, 25, 19, 27, 15 and 26 are considered as outliers. Note that in this data set the majority of data points form the shape of an ellipse. Therefore, the results obtained from $\mathbf{V} = \mathbf{R}^{-1}$, which takes the shape into account, are quite different from those obtained from $\mathbf{V} = \mathbf{I}_p$, in which the shape is irrelevant.

4.4. Example 4: Open/closed book data set

The fourth data set is the open/closed book data set (Mardia, Kent & Bibby, 1979) which has been analysed in the literature using factor analysis models (Tanaka & Odaka, 1989; Tanaka, Watadani & Moon, 1991; Lee & Wang, 1996). The data set consists of five variables and 88 observations among which case 81 was identified by all analyses as the most influential observation. Moreover, cases 3, 28, and 56 and cases 1, 2, 33 and 87 were also identified by Tanaka *et al.* (1991) and Lee & Wang (1996) respectively as cases meriting special attention.

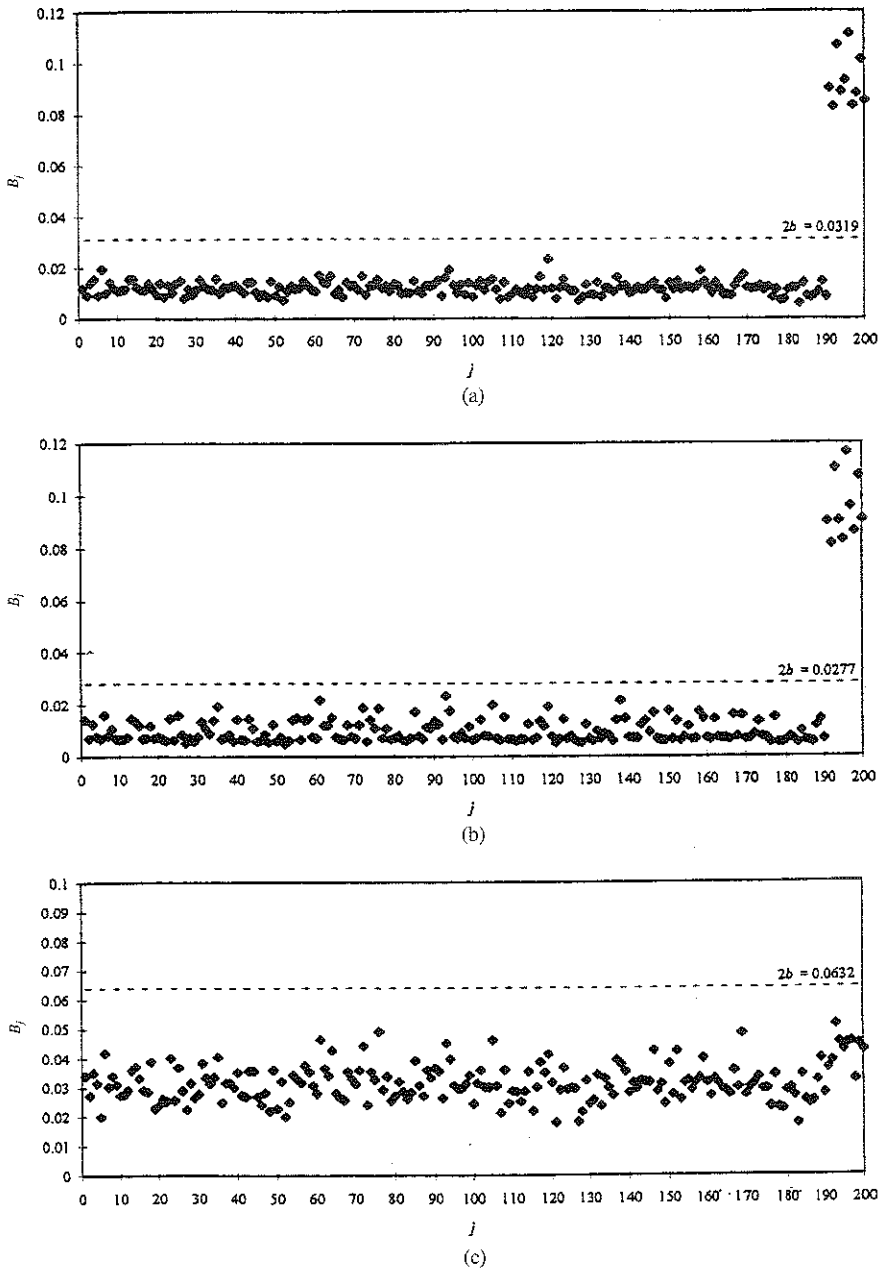

Figure 1. Index plot of B_j for the high-dimensional data set: (a) $\mathbf{V} = \mathbf{I}_p$; (b) $\mathbf{V} = \mathbf{R}^{-1}$; (c) $\mathbf{V} = \mathbf{S}^{-1}$.

The index plots of B_j for the three different choices of $\mathbf{V}$ are presented in Fig. 4. When $\mathbf{V} = \mathbf{I}_p$ (Fig. 4(a)), it is found that cases 88, 1, 87, 2 and 3 are clear outliers while cases 81, 82 and 85 are less extreme outliers. The results of $\mathbf{V} = \mathbf{R}^{-1}$ and $\mathbf{V} = \mathbf{S}^{-1}$ basically agree: cases 28, 54, 56, 61, 81, 87 and 88 are identified as outliers and case 82 is considered as a marginal case.

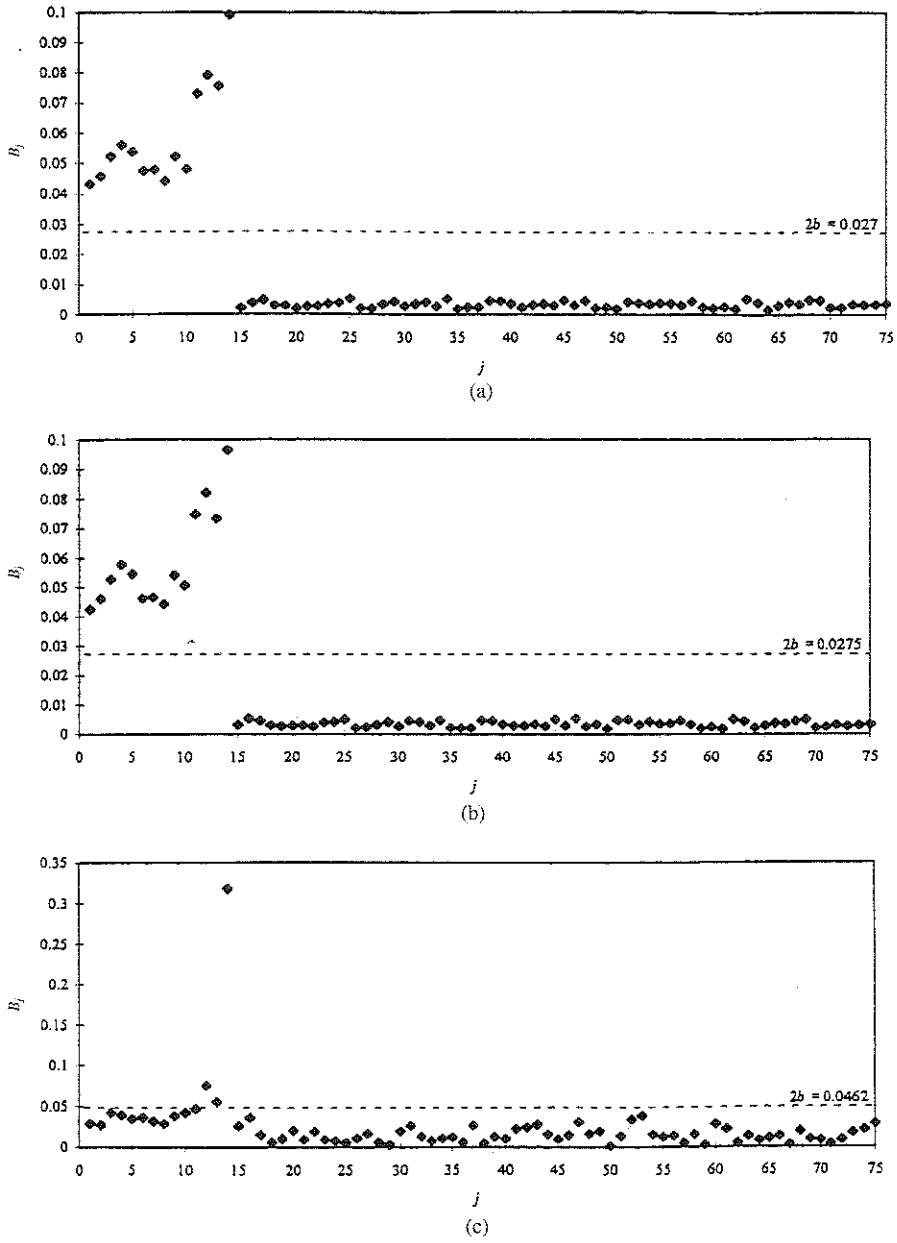

Figure 2. Index plot of B_j for the Hawkins *et al.* data set: (a) $\mathbf{V} = \mathbf{I}_p$; (b) $\mathbf{V} = \mathbf{R}^{-1}$; (c) $\mathbf{V} = \mathbf{S}^{-1}$.

Note that outliers identified by the current procedure are influential as far as location estimates are concerned, while those identified in the literature are influential in the analysis of a factor model; therefore, the set of outlying observations is not the same as those given in the literature.

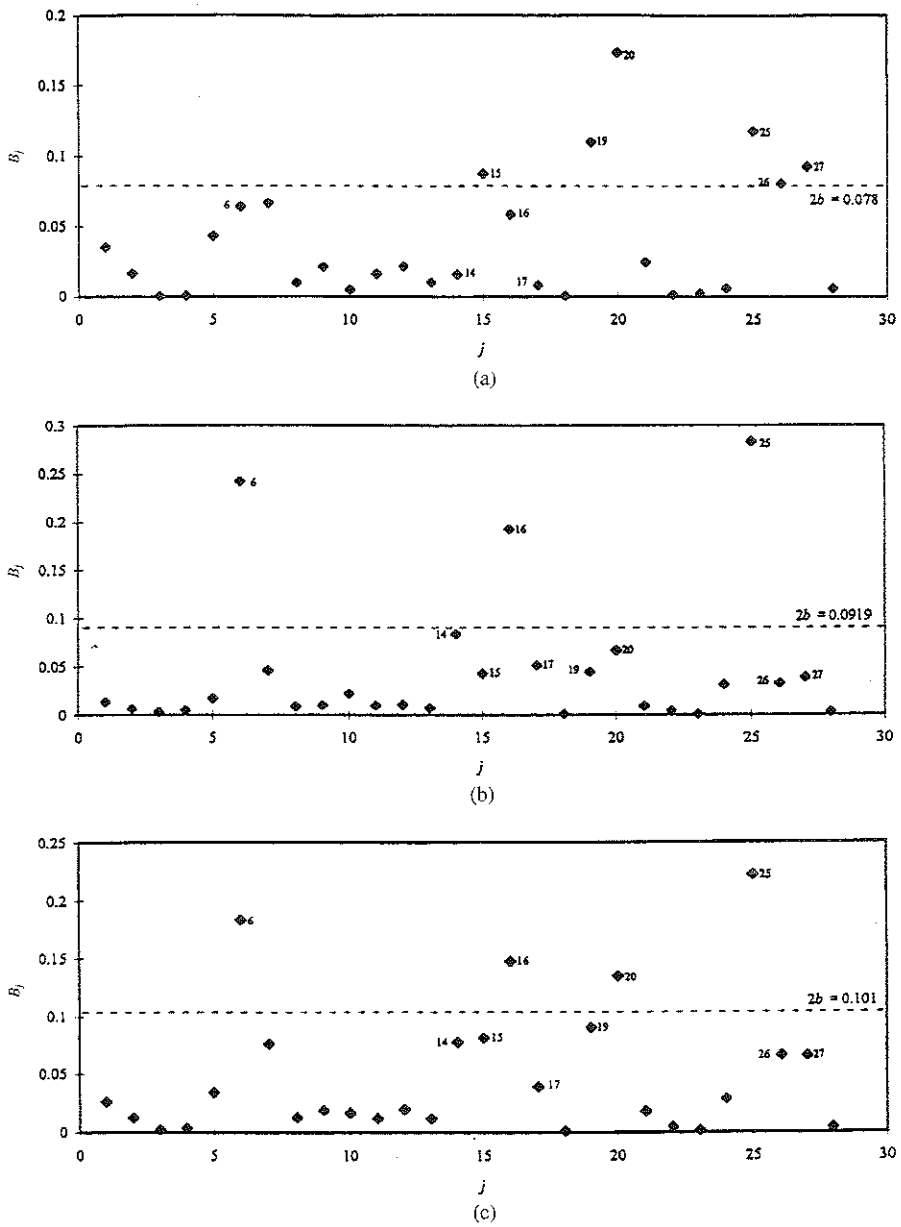

Figure 3. Index plot of B_j for the brain and body weight data set: (a) $\mathbf{V} = \mathbf{I}_p$; (b) $\mathbf{V} = \mathbf{R}^{-1}$; (c) $\mathbf{V} = \mathbf{S}^{-1}$.

4.5. Example 5: Attitudes data set

The fifth data set is taken from the 'Attitudes of Morality and Equality' example given in LISREL (Jöreskog & Sörbom, 1988, p. 193). Swedish school children in grade 9 were asked questions about their attitudes regarding social issues in family, school and society. The data

set given in LISREL consists of responses on eight items: for me, questions about (1) human rights, (2) equal conditions for all people, (3) racial problems, (4) equal value of all people, (5) euthanasia, (6) crime and punishment, (7) conscientious objectors, and (8) guilt and bad conscience are (1) unimportant, (2) not important, (3) important, or (4) very important. Responses to the eight questions were scored 1, 2, 3 and 4, where 4 = very important. Since the observed variables are ordinal categorical, statistical methods designated for analysing ordinal categorical variables are more appropriate for studying this data set. However, there are still many practical researchers who prefer to use the simple method of treating the ordinal measures as if they were interval scales. In this case, the proposed procedure can be used to detect location outliers since the entire development does not rely on any distributional assumption or scale type of the variable. We therefore analysed the data set of size 200 using five different choices of $\mathbf{V}$: $\mathbf{V} = \mathbf{I}_p$, $\mathbf{V} = \mathbf{R}^{-1}$, $\mathbf{V} = \mathbf{S}^{-1}$, $\mathbf{V} = \mathbf{K}^{-1}$ and $\mathbf{V} = \mathbf{P}^{-1}$, where $\mathbf{K}$ and $\mathbf{P}$ are correlation matrices storing Kendall's tau and the polychoric correlations, respectively. The index plots of the resulting B_j are presented in Fig. 5. Since variables are ordinal categorical with many observations taking the same value, the effect of influential observations is not as prominent as those in other examples where the variables are interval. In order to obtain a comprehensive picture of the data set, we present in Table 3 a plot analogous to the stalactite plot (Atkinson & Mulira, 1993) providing a cogent summary of suspected outliers, where a case identified as an outlier is marked. Note that among the 200 observations, there are five cases which contain missing values scored by 0. They are cases 5, 33, 52, 98 and 159. Apart from case 98 which is marginal when $\mathbf{V} = \mathbf{R}^{-1}$, all observations have been identified as influential observations under various choices of $\mathbf{V}$. Moreover, it is observed that while the sets of cases identified by $\mathbf{V} = \mathbf{R}^{-1}$, $\mathbf{S}^{-1}$, $\mathbf{K}^{-1}$ and $\mathbf{P}^{-1}$ are similar, the cases identified by $\mathbf{V} = \mathbf{I}_p$ are quite different from the others. Nevertheless, it is evident that cases 67, 109, 113, 117 and 157 are location outliers because they are identified as influential under all chosen metrics.

5. Discussion

Using the local influence approach, we have proposed a method to develop measures for outlier identification. The method has a clear geometrical basis and does not depend on large-sample or distributional properties, hence a change of the chosen metric or its estimate does not affect the validity of the procedure. With different choices of the null perturbation at ω_0 and different choices of the metric $\mathbf{V}$, the proposed procedure can produce measures which are equivalent to those in the literature. Therefore, the method provides a unifying approach for outlier measures. Furthermore, since the proposed measure is in fact a conformal normal curvature, it inherits many desirable features of the conformal normal curvature. For example, the measure assumes values in the unit interval; and if the contribution of all conformal normal curvatures along the directions in the standard basis is uniform, each is equal to a known constant. As a result, it becomes easier to judge the magnitudes.

Nevertheless, when an observation is considered as an outlier with a specific metric in mind, whether the chosen $\mathbf{V}$ nicely matches the specific metric determines the performance of B_j as an outlying measure. In this paper, for the purpose of easy comparison with the results in the literature, we consider $\mathbf{V} = \mathbf{\Sigma}^{-1}$ as one of the true metrics when we work on the examples given in Section 4, and estimate $\mathbf{\Sigma}$ by $\mathbf{S}$. However, since $\mathbf{S}$ as an estimate of $\mathbf{\Sigma}$ is highly affected by outliers, the performance of the proposed measure is not satisfactory. A robust estimate $\mathbf{R}$ of $\mathbf{\Sigma}$ is a better alternative. In the examples, we use the method proposed by Hadi

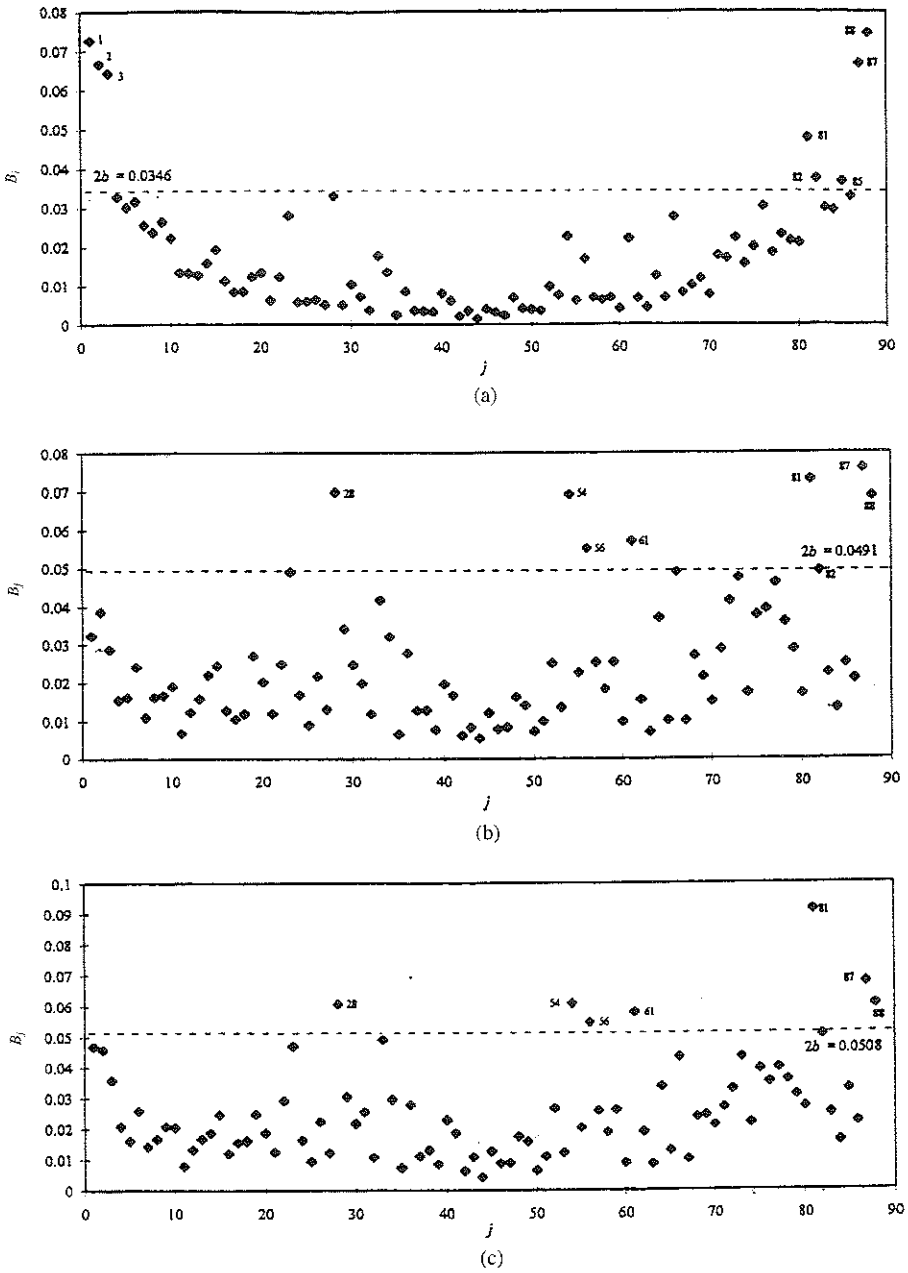

Figure 4. Index plot of B_j for the open/closed book data set: (a) $\mathbf{V} = \mathbf{I}_p$; (b) $\mathbf{V} = \mathbf{R}^{-1}$; (c) $\mathbf{V} = \mathbf{S}^{-1}$.

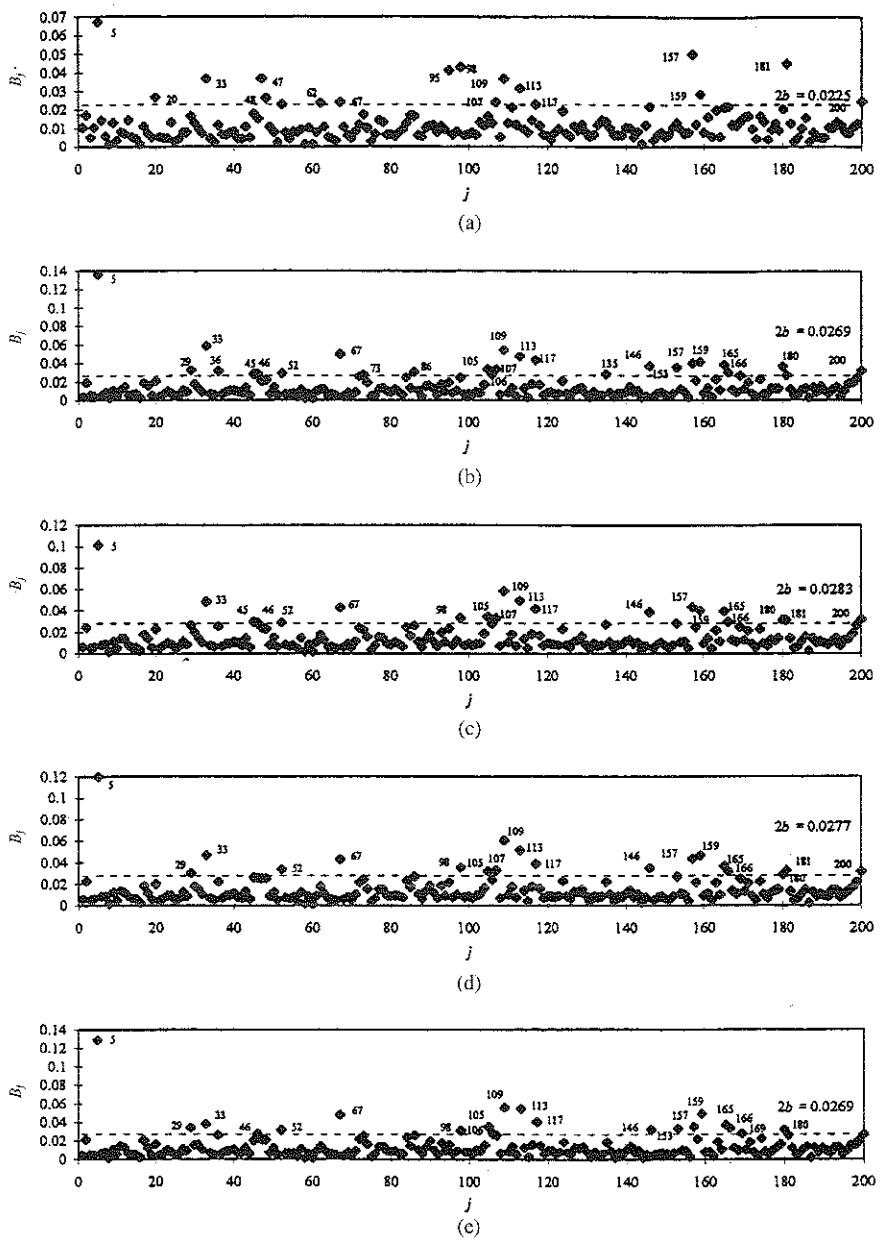

Figure 5. Index plot of B_j for the attitude data set: (a) $\mathbf{V} = \mathbf{I}_p$; (b) $\mathbf{V} = \mathbf{R}^{-1}$; (c) $\mathbf{V} = \mathbf{S}^{-1}$; (d) $\mathbf{V} = \mathbf{K}^{-1}$; (e) $\mathbf{V} = \mathbf{P}^{-1}$.

(1992, Appendix A) to compute $\mathbf{R}$. Hadi's method was employed simply because the computation of $\mathbf{R}$ is easier compared to other robust estimation methods, and the results produced are good enough for illustrative purpose. Other robust estimators can be considered, and the validity of the procedure is not affected. For example, one may consider using

Table 3. Suspected outliers for the attitude data set

Case \ V	$\mathbf{I}_p$	$\mathbf{R}^{-1}$	$\mathbf{S}^{-1}$	$\mathbf{K}^{-1}$	$\mathbf{P}^{-1}$
5[a]	*	*	*	*	*
20	*				
29		*		*	*
33[a]	*	*	*	*	*
36		*			
45		*	*		
46		*	*		*
47	*				
48	*				
52[a]		*	*	*	*
62	*				
67	*	*	*	*	*
73		*			
86		*			
95	*				
98[a]	*		*	*	*
105		*	*	*	*
106		*			*
107	*	*	*	*	
109	*	*	*	*	*
113	*	*	*	*	*
117	*	*	*	*	*
135		*			
146		*	*	*	*
153		*			*
157	*	*	*	*	*
159[a]	*	*	*	*	*
165		*	*	*	*
166		*	*	*	*
169					*
180		*	*	*	*
181	*		*	*	
200	*	*	*	*	

[a] Cases with missing value(s).

estimators that have a high breakdown point (see Rousseeuw, 1985), but the computational burden will increase.

One prominent advantage of the proposed procedure is its generality. Since the basic theory is geometrical and no distributional or large-sample properties are involved, the potential exists to apply the measure in complicated situations, such as in non-normal cases and in problems with small sample sizes. For example, since the estimate of $\boldsymbol{\mu}$ obtained by minimizing (10) does not rely on any distributional assumption or scale type of the $\mathbf{x}$ variables, it is possible to generalize the idea of considering an observation that affects the estimate of the location parameter $\boldsymbol{\mu}$ substantially as an influential observation in the presence of ordinal categorical variables. We demonstrate how this can be done in Section 4.5. However, it is worth knowing that when variables are observed in the form of ordered categories and when the dimension is of manageable size, it is possible to arrange the observations into frequencies of a contingency table and use a multinomial model to perform

analysis. In these circumstances, case-weights perturbation may not be appropriate because it is no longer the influence of individual observation which is of interest but rather the influence of the observed frequency in a specific cell of the contingency table (see Anderson, 1992; Poon *et al.*, 1999).

Finally, due to the relationship between the proposed measure and the measures available in the literature, the procedure throws new light on the possibility of applying available outlier identification procedures in complicated situations.

Acknowledgement

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC ref. no. CUHK4186/98P). Y. S. Poon thanks the Department of Statistics at the Chinese University of Hong Kong for their hospitality, and all the authors thank the referees and the Editor for useful comments.

References

Anderson, E. B. (1992) Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society B, 54*, 781–791.

Atkinson, A. C. (1994) Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association, 89*, 1329–1339.

Atkinson, A. C., & Mulira, H. M. (1993) The stalactite plot for the detection of multivariate outliers. *Statistics and Computing, 3*, 27–35.

Beckman, R. J., Nachtsheim, C. J., & Cook, R. D. (1987) Diagnostics for mixed-model analysis of variance. *Technometrics, 19*, 413–426.

Cook, R. D. (1986) Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B, 48*, 133–169.

Fung, W. K. (1993) Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*, 515–519.

Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B, 54*, 761–771.

Hadi, A. S. (1994) A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society B, 56*, 393–396.

Hadi, A. S., & Simonoff, J. S. (1993) Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88*, 1264–1272.

Hawkins, D. M., Bradu, D., & Kass, G. V. (1984) Location of several outliers in multiple regression data using elemental sets. *Technometrics, 26*, 197–208.

Hawkins, D. M., & Simonoff, J. S. (1993) High breakdown regression and multivariate estimation. *Applied Statistics, 42*, 423–441.

Jöreskog, K. G., & Sörbom, D. (1998) *LISREL 7: A guide to the program and applications*. Chicago: SPSS.

Lawrance, A. J. (1988) Regression transformation diagnostics using local influence. *Journal of the American Statistical Association, 83*, 1067–1072.

Lee, S.Y., & Wang, S.J. (1996) Sensitivity analysis of structural equation models. *Psychometrika, 61*, 93–108.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979) *Multivariate analysis*. New York: Academic Press.

Poon, W. Y., & Poon, Y. S. (1999) Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society B, 61*, 51–61.

Poon, W. Y., Wang, S. J., & Lee, S. Y. (1999) Influence analysis of structural equation models with polytomous variables. *Psychometrika, 64*, 461–473.

Rocke, D. M., & Woodruff, D. L. (1996) Identification of outliers in multivariate data. *Journal of the American Statistical Association, 91*, 1047–1061.

Rousseeuw, P. J. (1985) Multivariate estimation with high breakdown point. In W. Grossmann, G.Pflug,

I. Vincze, & W. Wertz (eds.), *Mathematical statistics and applications* (Vol. B, pp. 283–297). Dordrecht: Reidel.

Rousseeuw, P. J., & Leroy, A. (1987) *Robust regression and outlier detection.* New York: John Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633–639.

Shi, L. (1997) Local influence in principal component analysis. *Biometrika, 84*, 175–186.

St Laurent, R. Y., & Cook, R. D. (1993) Leverage, local influence and curvature in nonlinear regression. *Biometrika, 80*, 99–106.

Tanaka, Y., & Odaka, Y. (1989). Influential observations in principal factor analysis. *Psychometrika, 54*, 475–485.

Tanaka, Y., Watadani, S., & Moon, S. H. (1991). Influence in covariance structure analysis: With an application to confirmatory factor analysis. *Communications in Statistics A, 20*, 3805–3821.

Thomas, W., & Cook, R. D. (1989) Assessing influence on regression coefficients in generalized linear models. *Biometrika, 76*, 741–749.

Thomas, W., & Cook, R. D. (1990) Assessing influence on predictions in generalized linear models. *Technometrics, 32*, 59–65.

Tsau, C. L., & Wu, X. (1992) Transformation-model diagnostics. *Technometrics, 34*, 197–202.

Received 10 May 1999; revised version received 26 January 2000

Appendix: Derivations

A.1. Derivation of $\boldsymbol{\mu}_{\boldsymbol{\omega}}$

Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ be the standard basis for the p-dimensional vector space; then for α such that $1 \leq \alpha \leq p$, we obtain from (11) that

$$\frac{\partial L(\boldsymbol{\mu}|\boldsymbol{\omega})}{\partial \mu_\alpha} = -2\sum_i \omega_i \langle -\mathbf{e}_\alpha, \mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu})\rangle$$

$$= 2\sum_i \omega_i \langle \mathbf{V}\mathbf{e}_\alpha, \mathbf{x}_i - \boldsymbol{\mu}\rangle \tag{A1}$$

$$= 2\left\langle \mathbf{V}\mathbf{e}_\alpha, \sum_i \omega_i \mathbf{x}_i - \left(\sum_i \omega_i\right)\boldsymbol{\mu}\right\rangle. \tag{A2}$$

The critical point is given when the above derivative is equal to zero for every α. When $\mathbf{V}$ is non-degenerate, then $\{\mathbf{V}\mathbf{e}_1, \ldots, \mathbf{V}\mathbf{e}_p\}$ is a basis. Therefore,

$$\sum_i \omega_i \mathbf{x}_i - \left(\sum_i \omega_i\right)\boldsymbol{\mu} = 0 \tag{A3}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\omega}} = \frac{\sum_{i=1}^n \omega_i \mathbf{x}_i}{\sum_{i=1}^n \omega_i}. \tag{A4}$$

In particular, when $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$, then

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i. \tag{A5}$$

A.2. Derivation of the matrix $\ddot{\mathbf{L}}$

From (6), $\ddot{\mathbf{L}}$ is a $p \times p$ matrix with (α, β)th entry given by $\ddot{\mathbf{L}}_{\alpha\beta} = \frac{\partial^2 L(\boldsymbol{\mu})}{\partial \mu_\alpha \partial \mu_\beta} | \hat{\boldsymbol{\mu}}$. Now, from (A1), we have

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\mu})}{\partial \mu_\alpha} &= 2 \sum_i \langle \mathbf{V}\mathbf{e}_\alpha, \mathbf{x}_i - \boldsymbol{\mu} \rangle \\
&= 2 \sum_i \mathbf{e}_\alpha' \mathbf{V}(\mathbf{x}_i - \boldsymbol{\mu}) \\
&= 2 \sum_i \sum_k \sum_l e_{\alpha k} V_{kl}(x_{il} - \mu_l)) \qquad \text{(with obvious notation)} \\
&\downarrow \\
&= 2 \sum_i \left(\sum_l V_{\alpha l}(x_{il} - \mu_l) \right) \qquad (e_{\alpha k} = 1 \text{ for only } k = \alpha).
\end{aligned}$$

As a result, the (α, β)th element of $\ddot{\mathbf{L}}$ is given by:

$$\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\mu})}{\partial \mu_\alpha \partial \mu_\beta} &= 2 \sum_i \frac{\partial \Sigma_l V_{\alpha l}(x_{il} - \mu_l)}{\partial \mu_\beta} \\
&= 2 \sum_i V_{\alpha\beta}(-1) \\
&= -2nV_{\alpha\beta}.
\end{aligned}$$

Therefore, we have $\ddot{\mathbf{L}} = -2n\mathbf{V}$.

A.3 Derivation of the matrix $\boldsymbol{\Delta}$

From (A2), we have

$$\frac{\partial L(\boldsymbol{\mu}|\boldsymbol{\omega})}{\partial \mu_\alpha} = 2 \left\langle \mathbf{V}\mathbf{e}_\alpha, \sum_i \omega_i \mathbf{x}_i - \left(\sum_i \omega_i \right) \boldsymbol{\mu} \right\rangle. \tag{A6}$$

Therefore, the (α, j)th entry of the $p \times n$ matrix $\boldsymbol{\Delta}$ is given by

$$\begin{aligned}
\Delta_{\alpha j} = \frac{\partial^2 L(\boldsymbol{\mu}|\boldsymbol{\omega})}{\partial \mu_\alpha \omega_j} &= 2 \left\langle 0, \sum_i \omega_i \mathbf{x}_i - \left(\sum_i \omega_i \right) \boldsymbol{\mu} \right\rangle + 2 \langle \mathbf{V}\mathbf{e}_\alpha, \mathbf{x}_j - \boldsymbol{\mu} \rangle) \\
&= 2\mathbf{e}_\alpha' \mathbf{V}(\mathbf{x}_j - \boldsymbol{\mu}) \\
&= 2\mathbf{V}_\alpha'(\mathbf{x}_j - \boldsymbol{\mu}),
\end{aligned} \tag{A7}$$

where $\boldsymbol{\mu}$ is evaluated at $\hat{\boldsymbol{\mu}}$, and $\mathbf{V}_\alpha$ is an $p \times 1$ vector that stores the αth row of the matrix $\mathbf{V}$. Define the $p \times n$ matrix $\mathbf{Y} = (\mathbf{x}_1 - \hat{\boldsymbol{\mu}}, \ldots, \mathbf{x}_n - \hat{\boldsymbol{\mu}})$; then the matrix $\boldsymbol{\Delta}$ is given by

$$\boldsymbol{\Delta} = 2\mathbf{V}\mathbf{Y}. \tag{A8}$$