The Detection of Influential Subsets in Linear Regression by using an Influence Matrix

Daniel Pena; Victor J. Yohai

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), 145-156.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281995%2957%3A1%3C145%3ATDOISI%3E2.0.CO%3B2-D

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 145–156

The Detection of Influential Subsets in Linear Regression by using an Influence Matrix

By DANIEL PEÑA† and VICTOR J. YOHAI

Universidad Carlos III de Madrid, Spain

Universidad de San Andreés and Universidad de Buenos Aires, Argentina

[Received July 1992. Final revision February 1994]

SUMMARY

This paper presents a new method to identify influential subsets in linear regression problems. The procedure uses the eigenstructure of an influence matrix which is defined as the matrix of uncentred covariances of the effect on the whole data set of deleting each observation, normalized to include the univariate Cook statistics on the diagonal. It is shown that the eigenstructure of the influence matrix is useful to identify influential subsets and a procedure for detecting influential sets is proposed. The method is illustrated with two examples.

Keywords: EIGENVECTORS; MASKING; MULTIVARIATE INFLUENCE; OUTLIERS

1. INTRODUCTION

Many procedures are available to identify a single outlier or an isolated influential point in linear regression. Beckman and Cook (1983) and Chatterjee and Hadi (1986) have surveyed some of these procedures. The detection of influential subsets or multiple outliers is more difficult, owing to masking and swamping problems. Masking occurs when one outlier is not detected because of the presence of others, swamping when a non-outlier is wrongly identified owing to the effect of some hidden outliers. Several procedures have been proposed for dealing with multiple outliers. Marasinghe (1985) and Kianifard and Swallow (1990) have suggested a sequential testing strategy to identify a set of k points, where the maximum number of outliers in the sample, k, must be fixed in advance. Atkinson (1986), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) have suggested the use of robust estimates with high breakdown for the regression parameters to overcome the masking problem. These estimates are computed by using a resampling scheme. Hawkins *et al.* (1984) have proposed a diagnostic procedure which is also based on a resampling scheme. Gray and Ling (1984) proposed the use of cluster analysis over a modified hat matrix to identify influential sets, and Hocking (1984) has suggested that the eigenstructure of the matrix $(\mathbf{X}:\mathbf{y})'(\mathbf{X}:\mathbf{y})$ should be computed, where $\mathbf{y}$ is the vector of responses and the matrix $\mathbf{X}$ contains the explanatory variables.

In this paper we present a new method to identify influential subsets by looking at the eigenvalues of an 'influence matrix'. This matrix is defined as the uncentred covariance of a set of vectors which represent the effect on the fit of the deletion of each data point. This matrix is normalized to have the univariate Cook (1979)

†*Address for correspondence*: Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, calle Madrid 126, 28903 Getafe, Madrid, Spain.
E-mail: dpena@eco.uc3m.es

 0035–9246/95/57145

statistics on the diagonal. The method seems to work very well in all the data sets in which it has been tested.

The paper is organized as follows. Section 2 defines the influence matrix. Section 3 gives a heuristic justification of why its eigenvectors linked to non-null eigenvalues can be used to identify influential subsets. Section 4 presents a procedure to identify the set of influential points by using the eigenvectors of the influence matrix. Section 5 applies the procedure to two examples.

2. INFLUENCE MATRIX

Consider a linear regression model between an independent variable Y and p carriers $X_1, \ldots, X_p$, and suppose that there are n data points $(y_i, x_{i1}, \ldots, x_{ip})$, $1 \leqslant i \leqslant n$. The following notation will be used in the rest of the paper: $\mathbf{y} = (y_1, \ldots, y_n)'$; $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$; $\mathbf{X}$ is the $n \times p$ matrix with rows $\mathbf{x}_1', \ldots, \mathbf{x}_n'$. Then, according to the standard linear model assumptions, $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$, where $\mathbf{b} = (b_1, \ldots, b_p)'$ is the vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ the vector of regression errors, and the ϵ_i are independent random variables with distribution $N(0, \sigma^2)$. Let $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ be the least squares estimate (LSE) and let $\hat{\mathbf{b}}_{(i)}$ be the LSE when the ith data point is deleted. Then, the effect of this point on $\hat{\mathbf{b}}$ is given by (see Cook and Weisberg (1982), p. 110)

$$\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(i)} = \frac{e_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1 - h_{ii}}, \tag{1}$$

where h_{ij} is the ijth element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Consequently, if we denote by $\hat{y}_{j(i)}$ the new fitted value for observation j, we obtain

$$\hat{y}_j - \hat{y}_{j(i)} = \frac{h_{ij}e_i}{1 - h_{ii}}. \tag{2}$$

Put $\hat{\mathbf{y}}_{(i)} = (\hat{y}_{1(i)}, \ldots, \hat{y}_{n(i)})'$; then the vector $\mathbf{t}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$ summarizes the effect on the fit of deleting the observation i and is given by $\mathbf{t}_i = \{e_i/(1 - h_{ii})\}\mathbf{h}_i$, where $\mathbf{h}_i$ is the ith column of the $\mathbf{H}$-matrix. The individual deletion statistics identify influential points as those with large values of $\mathbf{t}_i$ in some suitable norm. For instance, Cook's statistic is given by $\mathbf{t}_i'\mathbf{t}_i/p\sigma^2$. However, when masking is present, the $\mathbf{t}_i$-values corresponding to outliers tend to be small, and therefore they are not detected.

One of the most important types of masking situations occurs when several observations have similar effects on the least squares fit. This is formalized by the following definition: we shall say that two observations i and j have similar effects on the least squares fit when $\mathbf{t}_i \approx \lambda\mathbf{t}_j$ for some scalar $\lambda > 0$ and opposed effects when $\lambda < 0$. Of course it is possible to find sets of influential data points where this condition does not hold, and it should be stressed that not all types of masking imply proportional effects: further there are different types of proportional subset that do not produce masking. However, we think that this situation is particularly interesting because the standard procedures based on individual deletion will not work in this case. Then, to detect possible sets of influential observations having similar or opposite effects on the fit, it seems plausible to look at the uncentred covariance matrix of the $\mathbf{t}_i$. Let us call $\mathbf{T}$ the $n \times n$ matrix $\mathbf{T} = (\mathbf{t}_1 \ldots \mathbf{t}_n)$ whose

columns are the vectors $\mathbf{t}_i$. Then we define the $n \times n$ influence matrix $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{ps^2}\mathbf{T}'\mathbf{T}, \tag{3}$$

where $s^2 = \Sigma_{i=1}^{n} e_i^2/(n-p)$. Using equation (2) and the fact that $\mathbf{H}$ is idempotent it immediately follows that $\mathbf{M}$ is given by

$$\mathbf{M} = \frac{1}{ps^2}\mathbf{EDHDE} \tag{4}$$

where $\mathbf{E}$ is the diagonal matrix that has the residuals on the main diagonal and $\mathbf{D}$ is a diagonal matrix with elements $(1-h_{ii})^{-1}$. Therefore, the ijth element of $\mathbf{M}$ is

$$m_{ij} = \frac{e_i e_j h_{ij}}{(1-h_{ii})(1-h_{jj})ps^2}. \tag{5}$$

Assuming that all the residuals are different from 0, from equation (4) the rank of $\mathbf{M}$ is equal to p, the rank of $\mathbf{H}$. Observe that the diagonal elements of $\mathbf{M}$ are the Cook (1979) statistics.

3. INTERPRETING EIGENVECTORS OF INFLUENCE MATRIX

Let I be an index set corresponding to a subset of data points. Cook and Weisberg (1982) proposed to measure the joint influence of the data points with index in I by

$$D_I = (e_I'(\mathbf{I}-\mathbf{H}_I)^{-1}\mathbf{H}_I(\mathbf{I}-\mathbf{H}_I)^{-1}e_I)/ps^2,$$

where the components of e_I are the least squares residuals and $\mathbf{H}_I$ the submatrix of $\mathbf{H}$ corresponding to the set I. A large value of D_I may be due to a single influential observation included in the set I and can also be due to the sum of small individual effects of a set of observations that are masking each other. However, in the first case this single observation will be easily identified. Also, a subset of individually highly influential points whose effect is to cancel each other out will lead to a small value of D_I, but, in this case, again the individual effects will be easy to identify, as shown in example 1, case (b). Therefore, we shall concentrate here on the most interesting case in which masking is due to points that can only be identified by looking at them jointly.

Let r_{ij} be the uncentred correlation coefficient between $\mathbf{t}_i$ and $\mathbf{t}_j$, that, as defined in Section 2, measures the effects on the least square fit of the ith and jth points. Then,

$$r_{ij} = \frac{m_{ij}}{m_{ii}^{1/2}m_{jj}^{1/2}}. \tag{6}$$

Suppose that there are k groups of influential observations $I_1, \ldots, I_k$, such that

(a) if $i, j \in I_h$, then $|r_{ij}| = 1$ (this means that the effects on the least squares fit produced by the deletion of two points in the same set I_h have correlation 1 or -1),

(b) if $i \in I_j$ and $l \in I_h$ with $j \neq h$, then $r_{il} = 0$ (this means that the effects produced on the least squares fit by observations i and j belonging to different sets are uncorrelated) and

(c) if i does not belong to any I_h, then $m_{ij}=0$ for all j (this means that data points outside these groups have no influence on the fit).

Now, according to (a) we can split each set I_h in I_h^1 and I_h^2 such that

(i) if $i, j \in I_h^q$, then $r_{ij}=1$ and
(ii) if $i \in I_h^1$ and $j \in I_h^2$, then $r_{ij}=-1$.

Let $\mathbf{v}_1=(v_{11}, \ldots, v_{1n})', \ldots, \mathbf{v}_k=(v_{k1}, \ldots, v_{kn})'$ be defined by $v_{hj}=m_{jj}^{1/2}$ if $j \in I_h^1$, $v_{hj}=m_{jj}^{1/2}$ if $j \in I_h^1$, $v_{hj}=-m_{jj}^{1/2}$ if $j \in I_h^2$ and $v_{hj}=0$ if $j \notin I_h$. Then, if (a)–(c) hold, by equation (6) the matrix $\mathbf{M}$ is

$$\mathbf{M}=\sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i',$$

and since the $\mathbf{v}_i$ are orthogonal the eigenvectors of $\mathbf{M}$ are $\mathbf{v}_1, \ldots, \mathbf{v}_k$, and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$ are given by

$$\lambda_h=\sum_{i \in I_h} m_{ii}. \tag{7}$$

It is clear that when the matrix $\mathbf{M}$ satisfies (a)–(c) the only sets I with large C_I are I_h^q, $1 \leqslant h \leqslant k$, $q=1, 2$, and these sets may be found by looking at the eigenvectors associated with non-null eigenvalues of $\mathbf{M}$. Equation (6) can also be written as

$$r_{ij}=\frac{\operatorname{sign}(e_i)\operatorname{sign}(e_j)h_{ij}}{(h_{ii}h_{jj})^{1/2}} \tag{8}$$

which means that, in the extreme case that we have presented, the $\mathbf{H}$-matrix and the signs of the residuals can, by themselves, identify the set of points that are associated with masking. For real data sets, conditions (a)–(c) do not hold exactly. However, the masking effect is typically due to the presence in the sample of blocks of influential observations having similar or opposite effects. These blocks are likely to produce a matrix $\mathbf{M}$ with a structure close to that described by (a)–(c). In fact, two influential observations i and j producing similar effects should have r_{ij} close to 1, and close to -1 when they have opposite effects. Influential observations with non-correlated effects have $|r_{ij}|$ close to 0. The same will happen with non-influential observations. Therefore, the eigenvectors will have approximately the structure described above, and the null components will be replaced by small values.

This suggests the following procedure to identify influential sets.

(a) Find the eigenvectors corresponding to the p non-null eigenvalues of the influence matrix $\mathbf{M}$.
(b) Consider the eigenvectors corresponding to large eigenvalues, and define the sets I_j^1 and I_j^2 by those components with large positive and negative weights respectively.

Cook (1986) proposed a procedure for assessing the influence on a vector of parameters θ of minor perturbations of a statistical model. He suggested the introduction of an $n \times p$ vector $\mathbf{w}$ of case weights and used the likelihood displacement $L(\hat{\theta})-L(\hat{\theta}_{\mathbf{w}})$, where $\hat{\theta}$ is the maximum likelihood (ML) estimator of $\hat{\theta}$ and $\hat{\theta}_{\mathbf{w}}$ the ML when the case weight $\mathbf{w}$ is introduced. Then he showed that the directions of

greatest local change in the likelihood displacement for the linear regression model are given by the eigenvectors linked to the largest eigenvalues of the curvature matrix $\mathbf{L}=\mathbf{EHE}$. The influence matrix $\mathbf{M}$, given by equation (4), may be considered a generalization of Cook's local influence matrix $\mathbf{L}$. It replaces the matrix of residuals $\mathbf{E}$ by the matrix of standardized residuals $\mathbf{ED}$. If there are no high leverage observations and the h_{ii} are similar for all points, both matrices will also be similar and will have similar eigenvectors. However, when the observations have very different leverage, the directions corresponding to the eigenvectors of the matrix $\mathbf{M}$ give more weight to the influence of the high leverage observations, which are precisely those more likely to produce masking effects.

4. PROCEDURE FOR DETECTING INFLUENTIAL SETS

The previous considerations have shown the need to look at the eigenvectors corresponding to the largest non-zero eigenvalues of the influence matrix. However, different influential subsets are expected to appear in different eigenvectors, as explained in Section 3, and, also, in some extreme cases, the only influential subset will be indicated by the eigenvector linked to a small eigenvalue. For instance, suppose that we add, to a sample of n points, $(\mathbf{y}_0, \mathbf{X}_0)$ k identical points $(\mathbf{y}_a, \mathbf{x}_a')$. Let h_a be the leverage of any of these points with respect to the initial sample; then, it is proved in Appendix A that the residual for these points in the entire regression using the $n+k$ points is

$$\mathbf{e}_a = \frac{\mathbf{y}_a - \mathbf{x}_a'\hat{\mathbf{b}}_0}{1+kh_a} \tag{9}$$

where $\hat{\mathbf{b}}_0 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{y}_0$. Therefore, as the leverage h_a is unbounded, the residual at these added points may be close to 0, and the diagonal elements of the hat matrix for these new points when h_a is large will tend to $1/k$. In this case, the terms m_{ij} of the influence matrix corresponding to these points will tend to 0, and the eigenvalue linked to this set, given by equation (7), will be very small.

In summary, it is useful to develop a strategy to look at all the eigenvectors linked to non-zero eigenvalues to find influential sets. In each eigenvector we must search for sets of co-ordinates with relatively large weight and the same sign. When the set of influential points has many components, and the eigenvectors are standardized to norm 1, the individual weights cannot be very large. Therefore, we must compare the relative value of the components to identify the elements of the set. The method suggested is to look at the ratios between the components in decreasing order, searching for a clear cut-off point, to form a set of candidate outliers, and then to test the points in this set to identify the outliers.

4.1. *Step 1: Identifying Sets of Outlier Candidates*

A set of candidate outliers is obtained by analysing the eigenvectors corresponding to the non-null eigenvalues of the influence matrix $\mathbf{M}$, and by searching in each eigenvector for a set of co-ordinates with relatively large weight and the same sign. The search can be done as follows.

(a) Order the co-ordinates of the eigenvector $\mathbf{v}_i$, obtaining $v_{i(1)} \leqslant v_{i(2)} \leqslant, \ldots,$

$\leqslant v_{i(n)}$, and let us call $i_{(1)}, \ldots, i_{(n)}$ the indices of the ordered co-ordinates of the eigenvector.

(b) Compute the ratios $a_j = v_{i(j)}/v_{i(j-1)}$ for $j = n, \ldots, n - c_1$ and $b_j = v_{i(j)}/v_{i(j+1)}$ for $j = 1, \ldots, c_2$. The constants c_1 and c_2 are smaller than $n/2$ and will be discussed below.

(c) Look for the first j_0 such that $|a_j| > k$ and the first i_0 such that $|b_j| > k$.

(d) If $i_0 > 1$ and/or $j_0 > 1$, consider the sets $J_0 = \{i_{(n)}, i_{(n-1)}, \ldots, i_{(n-i_0+1)}\}$ and/or $I_0 = \{i_{(i)}, i_{(2)}, \ldots, i_{(j_0-1)}\}$ as candidate outlier.

The choice of c_1 and c_2 is related to the desired breakdown point of the procedure that will be smaller than $\min(c_1/n, c_2/n)$. In practice we suggest c_1 and c_2 close to $n/4$. This number seems to be sufficiently small to avoid numerical instability due to denominators in the ratios close to 0. In any case the ratios should be computed for $|v_{i(j-1)}| > d$ where d is a small value but different from 0 to avoid numerical instability. The power of the procedure for the detection of outliers depends on the choice of k. However, since the candidate outlier will be further scrutinized, as explained in step 2, taking k too large can have more serious consequences than taking it too small. This explains why we do not recommend a choice of k according to a significance level. Our experience with real and simulated data leads us to recommend a value of $k = 2.5$.

4.2. *Step 2: Checking for Outliers*

(a) Remove all candidate outliers.

(b) Use the standard F- and t-statistics to test for groups or individual outliers. Reject sets or individual points with F- or t-statistic larger than some constant c. For the F-statistic the c-value corresponds to the distribution of the maximum F over all sets of the same size, and this distribution is unknown (see Beckman and Cook (1983)). Therefore, it is better to use the t-statistic and to choose the c-value by the Bonferroni inequality or, better, by simulating the procedure with normal errors.

(c) If the number of candidate outliers is larger than $n/2$, the previous procedure can be applied separately to the points identified in each eigenvector.

In most regression applications the sample size n is much larger than p, the rank of $\mathbf{X}$. Then, since we are only interested in the eigenvectors corresponding to the p non-null eigenvalues, the direct computation of the eigenvalues and eigenvectors of $\mathbf{M}$ will be very inefficient. A better procedure is to compute first $\mathbf{A} = \mathbf{B}\Lambda^{1/2}$, where the columns of $\mathbf{B}$ are the eigenvectors of $(\mathbf{X}'\mathbf{X})^{-1}$ and Λ is the diagonal matrix containing the corresponding eigenvalues. Then $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{A}\mathbf{A}'$, and therefore $\mathbf{M} = \mathbf{P}\mathbf{P}'$, where $\mathbf{P}$ is the $n \times p$ matrix given by

$$\mathbf{P} = \frac{1}{(ps^2)^{1/2}}\mathbf{EDXA}.$$

Therefore, as $\mathbf{PP}'$ and $\mathbf{P}'\mathbf{P}$ have the same non-null eigenvalues, we need only to compute the eigenvalues and eigenvectors of $\mathbf{P}'\mathbf{P}$. The eigenvectors of $\mathbf{M}$ linked to non-zero eigenvalues are obtained as $\mathbf{P}\mathbf{v}_i$ where $\mathbf{v}_i$ are the eigenvectors of $\mathbf{P}'\mathbf{P}$.

The power of the suggested procedure was investigated by using the model $y_i = b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 + \epsilon_i$, $1 \leqslant i \leqslant 40$, where for $1 \leqslant i \leqslant 40 - n_0$ the values y_i,

x_{i1} x_{i2} and x_{i3} are independent random samples from an $N(0, 1)$ distribution, and, therefore, correspond to the case $b_1 = b_2 = b_3 = b_4 = 0$. The last n_0 observations, i.e. for $40 - n_0 + 1 \leqslant i \leqslant 40$, are identical and correspond to $x_{i1} = x_0$, $x_{i2} = x_{i3} = 0$, $y_0 = mx_0$. This outlier design does not suppose any loss of generality due to the sphericity of the distribution of the x-variables. The values for x_0 were chosen to be 5 and 10, and the contaminating slope m was fixed at 1, 2, 3 and 4. The number n_0 of outliers was taken as 2, 6 or 8, corresponding to 5%, 15% and 20% contamination. The value of 5% was chosen to represent a small amount of contamination, whereas 15% and 20% represent a highly contaminated sample. For each contamination design 500 simulations were made, and the automatic procedure to identify and check for the set of outliers was applied with $c = 10$ and $k = 2.5$. The results are presented in Table 1.

It can be seen that the procedure is very powerful with 5% contamination and avoids the masking problem when the outliers have large leverage (case $x_0 = 10$). The size of the procedure has also been investigated by using for the significance level for checking for outliers $0.05/n$ in step 2. The final size obtained was 0.08.

5. EXAMPLES

5.1. *Example 1*

This first example is designed to show the interpretation of the eigenvectors of the influential matrix in three simple masking schemes (see Table 2 and Fig. 1). In the three cases we have eight good points generated by $y = 1 + x + u$ where u is a normal random variable with mean 0 and standard deviation 0.1 and two high leverage points. In case (a) we have the standard masking scheme in which both outliers produce the same effect and one is masked by the other; in case (b) the two outliers produce opposite effects; in case (c) we have swamping, i.e. the ninth point appears as an outlier because of the effect of the 10th point.

TABLE 1
Percentage of outlier detection in the Monte Carlo study

Contamination (%)	*% outliers for* $x_0 = 5$				*% outliers for* $x_0 = 10$			
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 1$	$m = 2$	$m = 3$	$m = 4$
5	99	100	100	99	99	100	100	99
15	47	64	74	75	92	90	90	87
20	50	56	58	59	86	98	98	97

TABLE 2
Data for example 1

Case	1	2	3	4	5	6	7	8	9(a)	10(a)	9(b)	10(b)	9(c)	10(c)
x	1	2	3	4	5	6	7	8	12	12	12	12	12	12
y	2.0	2.9	3.9	5.1	6.2	6.9	7.8	9.1	19	20	19	7	13	7

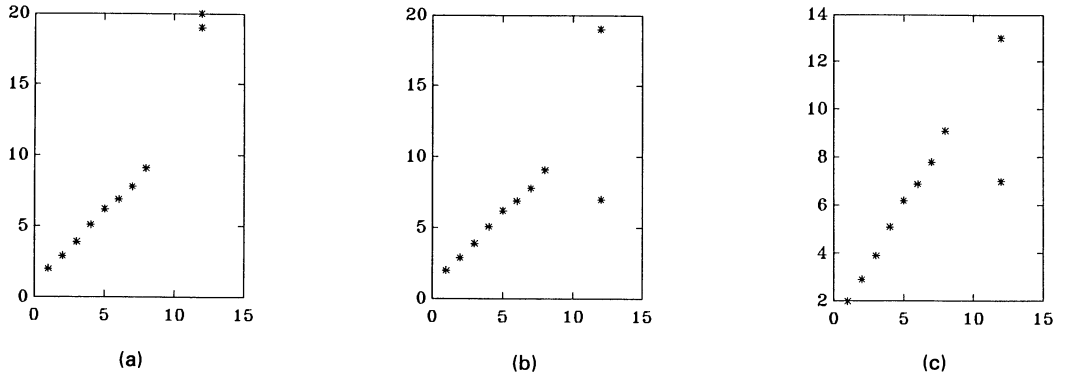

Fig. 1. Data for example 1 (the values are given in Table 2)

Table 3 presents the largest eigenvalue of the influence matrix and the corresponding eigenvector in these three cases. In case (a) the largest eigenvalue is roughly three times the next and gives the largest weight to the two outliers. Also the two outliers have positive weight, whereas all the good points have a small and negative weight. Therefore, the analysis shows two different sets of points.

The automatic procedure suggested in Section 4 produces the same results. The largest eigenvector has $a_1 = 1.9$, which corresponds to $i_{(1)} = 10$, and $a_2 = 9.974$ with $i_{(2)} = 9$. Therefore it has a clear cut-off point at the set $\{9, 10\}$. All the b_j are small. Therefore, this eigenvalue separates the set $\{9, 10\}$ from the rest. Table 4 presents the t-statistics for these points when they are removed from the least squares fit.

TABLE 3
Largest eigenvalue, ratio of largest eigenvalue to next eigenvalue and eigenvectors for example 1

Case	λ_1	λ_1/λ_2	*Eigenvalues for the following points:*									
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
(a)	1.27	2.87	−0.17	−0.06	−0.00	−0.00	−0.02	−0.10	−0.22	−0.33	0.42	0.79
(b)	3.78	3.783	0.00	−0.00	−0.00	−0.00	−0.00	0.00	−0.00	−0.00	−0.71	0.71
(c)	3.25	32	−0.05	−0.02	−0.00	−0.00	−0.01	−0.02	−0.04	−0.10	−0.50	0.85

TABLE 4
t-statistics for example 1

Case	*t-statistics for the following points:*	
	9	*10*
(a)	27.69	32.28
(b)	31.94	−32.09
(c)	−0.07	−32.09

In case (b) the two outliers are again clearly identified: they appear in the eigenvector corresponding to the largest eigenvalue with large values and opposite sign, whereas the rest of the points are given zero weight. The largest eigenvector has cut-off points at $a_1 = 153.9$, corresponding to observation 9, and $b_1 = 113.8$, corresponding to observation 10. Thus, again the set $\{9, 10\}$ is identified. The statistics for these points are given in Table 4. Finally, in case (c) the outlier is given a large and positive weight, whereas all the good points have negative weight, with the greatest value at the good high leverage point. The largest eigenvalue has two cut-off points. In the a_j this occurs for $a_1 = 4.8$, which corresponds to observation 9. In the b_j this occurs for $b_1 = 1226.6$ and corresponds to observation 10. Therefore the set $\{9, 10\}$ is identified as worthy of further analysis. Table 4 shows that the t-statistic identifies clearly point 10 as an outlier. In summary, the components of the eigenvector corresponding to the largest eigenvalue show in all cases the relevant structure of the data set, and, in all cases, the relevant set is automatically selected by the procedure suggested in Section 4.

5.2. *Example 2*

We use here the artificial data generated by Hawkins *et al.* (1984). The model contains 75 data points in four dimensions (one response and three explanatory variables). The first 10 data points are high leverage outliers, and the next four points are good observations with high leverage. The rest of the observations are good points with low leverage.

The eigenvalues of $\mathbf{M}$ are $\lambda_1 = 2.36$, $\lambda_2 = 1.63$, $\lambda_3 = 0.11$ and $\lambda_4 = 0.04$. The coefficients of the eigenvectors corresponding to λ_1 and λ_2 are shown in Table 5

TABLE 5
Two largest eigenvalues and their eigenvectors and univariate Cook D-statistic for the Hawkins et al. (1984) data

Case	*Eigenvalue coefficients*		*D*
	$\lambda_1 = 2.36$	$\lambda_2 = 1.63$	
1	−0.046	−0.100	0.040
2	−0.076	−0.108	0.053
3	−0.016	−0.118	0.046
4	−0.036	−0.090	0.031
5	−0.040	−0.105	0.039
6	−0.053	−0.103	0.052
7	−0.092	−0.121	0.079
8	−0.044	−0.121	0.052
9	−0.030	−0.098	0.034
10	−0.020	−0.115	0.047
11	0.15	0.297	0.035
12	−0.01	0.520	0.851
13	0.24	0.149	0.254
14	0.87	−0.138	2.11
Rest	$\lvert v_i \rvert < 0.032$	$\lvert v_i \rvert < 0.022$	$D_i < 0.10$

and Fig. 2. The first eigenvector gives high positive weight to observations in the set {11, 13, 14}, especially to observation 14. All these points are good high leverage points. Two sets of large coefficients may be distinguished in the second eigenvalue: the set {1, . . ., 10, 14} with negative coefficients and the set {11, 12, 13} with positive coefficients. Thus, the first set includes all the outliers and one good leverage point, and the second set three of the good high leverage points. It may be observed in Table 5 that the only large values of the univariate Cook D-statistic correspond to good leverage points, and therefore they do not detect any outlier points.

Table 6 summarizes the results when the automatic procedure is applied to this data set. We have chosen $c_1 = c_2 = 18$ and $k = 2.5$. It is seen in Table 6 that the first eigenvalue only identifies point {14} as a candidate outlier. However, the second eigenvector clearly shows the sets {11, 12, 13} and {1, . . ., 10, 14} as relevant. The other two eigenvalues have no clear cut-off points, and all a_j and b_j are smaller than 1.5.

Table 7 shows the t-statistics for this set. All the outliers are clearly identified.

All the computations for these examples were done on a PC386 at 33 MHz with a program written in MATLAB. The program is very simple because it only requires the matrix manipulations described in Section 4. The maximum time involved was for the last example, in which the computations were completed in 3 s.

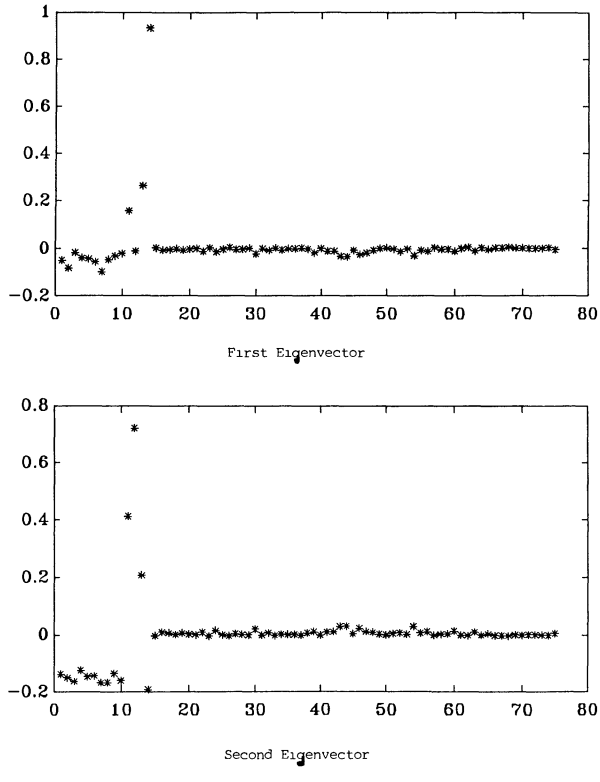

Fig. 2. Components of first and second eigenvectors for the influence matrix of Hawkins *et al.* (1984)

TABLE 6
Eigenvalue analysis for the Hawkins et al. (1984) data

Eigenvalue	a_j	*Set* a_j	b_j	*Set* b_j
2.36	–	–	3.55	{14}
1.63	6.80	{11, 12, 13}	33.15	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 14}
0.10	–	–	–	–
0.04	–	–	–	–

TABLE 7
t-statistics for the Hawkins et al. (1984) data

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14
t-statistic	6.10	6.20	6.06	5.57	5.86	6.05	6.43	6.37	5.74	6.05	1.08	1.03	1.36	0.99

ACKNOWLEDGEMENTS

We are very grateful to the Editor and a referee for help in revising the paper, and to Dennis Cook for useful suggestions and discussions. The authors acknowledge support from the Dirección General de Investigación Científica y Técnica, Spain, under grant PB92-0232.

APPENDIX A

We prove equation (9). Calling $\mathbf{X}$ the design matrix and $\mathbf{H}$ the hat matrix for the entire regression with the $n+k$ points, i.e. $\mathbf{X}' = (\mathbf{X}_0', \mathbf{x}_a\mathbf{1}')$ where $\mathbf{1}' = (1, \ldots, 1)$ and $\mathbf{Y}' = (\mathbf{y}_0', \mathbf{1}'y_a)$, $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\mathbf{b}}_0 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{y}_0$, then, using the standard Woodbury-Sherman-Morrison formula (see for instance Cook and Weisberg (1982), p. 136), we can write

$$\hat{\mathbf{b}}_0 = \hat{\mathbf{b}} - (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{x}_a\mathbf{1}'(\mathbf{I} - \mathbf{H}_k)^{-1}\mathbf{e}_a. \tag{10}$$

Here $\mathbf{e}_a = \mathbf{1}(y_a - \mathbf{x}_a'\hat{\mathbf{b}})$ and $\mathbf{H}_k$ is the square block of $\mathbf{H}$ corresponding to the k identical points $\mathbf{x}_a'$ that is given by

$$\mathbf{H}_k = \mathbf{1}\mathbf{x}_a'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_a\mathbf{1}' = \mathbf{1}\mathbf{1}'\frac{h_a}{1 + kh_a} \tag{11}$$

where $h_a = \mathbf{x}_a'(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{x}_a$. Then, from equation (10),

$$\mathbf{1}(y_a - \mathbf{x}_a'\hat{\mathbf{b}}_0) = (\mathbf{I} + \mathbf{H}_k(\mathbf{I} - \mathbf{H}_k)^{-1})\mathbf{e}_a, \tag{12}$$

as $(\mathbf{I} - \mathbf{H}_k)^{-1} = \mathbf{I} + \mathbf{H}_k(\mathbf{I} - \mathbf{H}_k)^{-1}$, using equation (11) we finally obtain

$$\mathbf{e}_a = \mathbf{1}\left(\frac{y_a - \mathbf{x}_a'\hat{\mathbf{b}}_0}{1 + kh_a}\right)$$

as claimed.

REFERENCES

Atkinson, A. C. (1986) Masking unmasked. *Biometrika*, **73**, 533–541.
Beckman, R. J. and Cook, R. D. (1983) Outlier. . .s. *Technometrics*, **25**, 119–163.
Chatterjee, S. and Hadi, A. S. (1986) Influential observations, high leverage points, and outliers in lineal regression. *Statist. Sci.*, **1**, 379–416.
Cook, R. D. (1979) Influential observations in linear regression. *J. Am. Statist. Ass.*, **74**, 169–174.
——— (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.
Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. New York: Chapman and Hall.
Daniel, C. and Wood, F. S. (1980) *Fitting Equations to Data*. New York: Wiley.
Gray, J. B. and Ling, R. F. (1984) *K*-clustering as a detection tool for influential subsets in regression. *Technometrics*, **26**, 305–330.
Hawkins, D. M., Bradu, D. and Kass, G. V. (1984) Location of several outliers in multiple regression data using elemental sets. *Technometrics*, **26**, 197–208.
Hocking, R. R. (1984) Discussion of Gray and Ling paper. *Technometrics*, **26**, 321–323.
Kianifard, F. and Swallow, W. (1990) A Monte Carlo comparison of five procedures for identifying outliers in lineal regression. *Communs Statist. Theory Meth.*, **19**, 1913–1938.
Marasinghe, M. G. (1985) A multistage procedure for detecting several outliers in linear regression. *Technometrics*, **27**, 395–399.
Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *J. Am. Statist. Ass.*, **85**, 633–651.