

# THE CASE SENSITIVITY FUNCTION APPROACH TO DIAGNOSTIC AND ROBUST COMPUTATION: A RELAXATION STRATEGY

Frank Critchley, Michael Schyns, Gentiane Haesbroeck, David Kinns, Richard A. Atkinson and Guobing Lu

*Key words:* combinatorial optimisation, convexity, diagnostics, Euclidean geometry, masking, multiple case effects, relaxation, robustness.

*COMPSTAT 2004 section:* Robustness

## 1 Overview and organisation

Central to both diagnostics and robustness are a range of optimisation problems that are combinatorial by definition and correspondingly hard to solve exactly. A variety of multiple case effects – such as masking – may be present, further complicating appropriate inference.

The present paper offers a computation-focused progress report on the case sensitivity function approach to diagnostics and robustness introduced in [4], on which we draw. A key idea here is that *relaxation brings benefits*. Specifically, the strategy outlined below shows how such high-dimensional ( $O({}^nC_m)$ ) discrete optimisation problems can be embedded in low-dimensional ( $O(n)$ ) smooth reformulations, in which both the insights of geometry and the power of analysis are available. In particular, informative plots become possible, while additional convexity and derivative information can be exploited.

Overall motivation for the case sensitivity function approach derives from considerations of (A) *unity*, (B) *insight* and (C) *innovation*, examples including – in order of appearance:

- (A1): an emphasis (throughout) on the connectivity of diagnostics and robustness,
- (A2): a single setting for a range of optimisation problems (Sections 3 and 4),
- (B1): visual displays affording insight into the nature and variety of multiple case effects (Section 5),
- (C1): new diagnostic methodologies (Section 6),
- (B2): insight into the performance of existing algorithms (Section 7), and:
- (C2): enhanced (potentially, encompassing) sets of algorithms for a class of robustness problems (Section 7).

A few preliminaries are established in Section 2.

## 2 Preliminaries

To gain focus, attention is restricted to one-sample contexts, with  $\{z_i : i \in N\}$ ,  $N := \{1, \dots, n\}$  denoting a random sample of  $n > 1$  cases from an unknown distribution  $F$  in  $\dim(z)$  dimensions. The associated empirical distribution is:

$$\hat{F} := \sum_{i \in N} n^{-1} \hat{F}_i \quad (1)$$

where  $\hat{F}_i$  denotes the distribution degenerate at  $z_i$ . Throughout, analysis is conducted conditional on the observed  $\{z_i\}$ .

Assuming, as we do, that no further information is available about the observed cases, it is desirable that any analysis of these data should be invariant under permutation of the arbitrary labels attached to them. Given  $n$ , this invariance is achieved – without loss of information – by replacing  $\{z_i : i \in N\}$  by  $\hat{F}$ . In particular, every statistic of interest here is of the form  $T[\hat{F}]$ , for some functional  $T[\cdot]$ . This may, for example, be (the observed significance level of) a test statistic, a parameter estimate, a prediction of future values of an observable, or a nonparametric density or regression function estimate. In particular,  $T[\cdot]$  may be scalar, vector or function valued.

Let  $Z := (z_i^T)$ . In multivariate contexts where all the random variables in  $\tilde{z} \sim F$  are on the same footing, we put  $\dim(z) = k$ ,  $\tilde{z} = \tilde{x}$ ,  $z_i = x_i$  and  $Z = X$ . In the usual linear model  $y = X\beta + \epsilon$ , we put  $\dim(z) = 1 + k$ ,  $\tilde{z}^T = (\tilde{y}, \tilde{x}^T)$  and  $z_i^T = (y_i, x_i^T)$ , so that  $Z = (y|X)$ , (a constant term being assumed and accommodated by supposing that the distribution of the first element of  $\tilde{x}$  is degenerate at the value 1).

## 3 A combinatorial optimisation problem

Two integers  $h > 0$  and  $m > 0$  are called *n-complementary* if  $h + m = n$ , in which case:

$$A \in \mathbb{N}_h \Leftrightarrow A^c \in \mathbb{N}_m \quad (2)$$

where, for any integer  $0 < a < n$ ,  $\mathbb{N}_a := \{\emptyset \subset A \subset N : |A| = a\}$ . In particular,  $|\mathbb{N}_h| = |\mathbb{N}_m|$  or, in the familiar combinatorial identity,  ${}^n C_h = {}^n C_m$ .

Throughout,  $\{H, M\}$  denotes a bipartition of  $N$ . That is,  $H$  and  $M$  are nonempty, complementary subsets of  $N$ . In particular,  $|H|$  and  $|M|$  are *n-complementary*. Of course, *holding onto* the cases labelled by  $H$  is the same thing as *missing out* those labelled by  $M$ . That is,

$$\hat{F}_H = \hat{F}_{-M} \quad (3)$$

where, for any  $\emptyset \subset A \subset N$ ,  $\hat{F}_A := \sum_{i \in A} |A|^{-1} \hat{F}_i$  and  $\hat{F}_{-A} := \hat{F}_{A^c}$ .

As is well-known, diagnostics and robustness meet at the influence function. The simple but general relations (2) and (3) provide a second, global connection

between these two areas of statistics, as we now discuss. For brevity, each scalar *target* functional  $t[\cdot]$  below is implicitly assumed to be defined wherever it is evaluated, and its possible dependence on  $\widehat{F}$  or  $T$  suppressed notationally.

A general problem arising in diagnostics is to identify subsets  $M$  of given size  $m$  whose omission causes maximal changes  $T[\widehat{F}] \rightarrow T[\widehat{F}_{-M}]$  in a statistic of interest, as measured by  $t[\widehat{F}_{-M}]$  for some appropriate target functional  $t[\cdot]$ . A lead example is Cook's (squared) distance in the linear model. With  $T[F] = \beta[F] := (\mathbb{E}_F(\widetilde{x}\widetilde{x}^T))^{-1}\mathbb{E}_F(\widetilde{x}\widetilde{y})$ ,  $\widehat{\beta} := \beta[\widehat{F}]$  and  $\widehat{\beta}_{-M} := \beta[\widehat{F}_{-M}]$ , we have:

$$t_{Cook}[\widehat{F}_{-M}] := (ks^2)^{-1}(\widehat{\beta}_{-M} - \widehat{\beta})^T X^T X(\widehat{\beta}_{-M} - \widehat{\beta}) \quad (4)$$

where  $s^2$  is the usual estimate of error variance. Again, a range of robust estimates are defined in terms of subsets  $H$  of given size  $h$  which optimise a specified target functional  $t[\widehat{F}_H]$ . A lead example is minimum covariance determinant (MCD) estimation in multivariate analysis based on minimisation of:

$$t_{MCD}[\widehat{F}_H] := \log(\det(\text{cov}[\widehat{F}_H])). \quad (5)$$

These two lead examples are developed below.

Summarising, a range of optimisation problems arising naturally in both diagnostics ( $\mathcal{D}$ ) and robustness ( $\mathcal{R}$ ) have *combinatorial* complexity and *entirely equivalent* ( $\mathcal{D}$ )  $\leftrightarrow$  ( $\mathcal{R}$ ) forms expressed in Problem 1, in which  $h$  and  $m$  are given  $n$ -complementary integers:

**Problem 1** (*Combinatorial optimisation problem*)

( $\mathcal{D}$ ) Optimise  $t[\widehat{F}_{-M}]$  over  $M \in \mathbb{N}_m$ .

( $\mathcal{R}$ ) Optimise  $t[\widehat{F}_H]$  over  $H \in \mathbb{N}_h$ .

We note in passing that a variety of other combinatorial problems, not necessarily linked to diagnostics and robustness, can also be formulated in this way.

This high-dimensional discrete problem can be embedded in a low-dimensional smooth one, as follows. It suffices to express such a relaxation strategy in, say, the ( $\mathcal{D}$ ) form, that in the ( $\mathcal{R}$ ) form following at once via (2) and (3).

## 4 A relaxation strategy

Throughout this section,  $h$  and  $m$  denote given  $n$ -complementary integers. Again,  $M$  denotes a general member of  $\mathbb{N}_m$ , and  $H$  its complement in  $N$ .

### 4.1 Probability vectors as labels for weighted empirical distributions

The first step in the relaxation strategy adopted here is to use probability vectors as labels for weighted empirical distributions.

For any  $p \equiv (p_i) \in \mathbb{P}^n := \{\text{all probability } n\text{-vectors}\}$ , let  $\widehat{F}(p) := \sum_{i \in N} p_i \widehat{F}_i$  denote the distribution attaching probability  $p_i$  to  $z_i$ , and  $\widehat{\mathbb{F}} := \{\widehat{F}(p) : p \in \mathbb{P}^n\}$ . For brevity, the  $\{z_i\}$  are assumed distinct (this avoids an elaboration required in the general case). Accordingly, (indeed, equivalently),

$$p \leftrightarrow \widehat{F}(p) \text{ is a bijection between } \mathbb{P}^n \text{ and } \widehat{\mathbb{F}}. \quad (6)$$

In particular, every weighted empirical distribution corresponds to one and only one probability vector, which provides a convenient label for it. For example (cf. (1)),  $p_0 := (n^{-1})$  labels  $\widehat{F}$ .

Moreover,  $p_{-M}$  labels  $\widehat{F}_{-M}$ , where the  $i^{\text{th}}$  element of  $p_{-M}$  is zero if  $i \in M$  and  $h^{-1}$  otherwise. That is, (6) specialises to:

$$p_{-M} \leftrightarrow \widehat{F}_{-M} \text{ is a bijection between } \mathbb{V}_{-m}^n \text{ and } \widehat{\mathbb{F}}_{-m},$$

where  $\mathbb{V}_{-m}^n$  comprises the  ${}^n C_m$  distinct probability vectors arising from permutation of  $h^{-1}(0_m^T, 1_h^T)^T$  and  $\widehat{\mathbb{F}}_{-m} := \{\widehat{F}_{-M} : M \in \mathbb{N}_m\}$  is the set of distributions optimised over in Problem 1.

The  $(\mathcal{R})$  form is immediate, writing  $p_{-M}$ ,  $\mathbb{V}_{-m}^n$  and  $\widehat{\mathbb{F}}_{-m}$  as  $p_H$ ,  $\mathbb{V}_h^n$  and  $\widehat{\mathbb{F}}_h$  respectively. Of course, in the limit when  $m = (n - 1)$  (equivalently,  $h = 1$ ),  $\mathbb{V}_{-m}^n$  comprises the  $n$  unit vectors in  $\mathbb{P}^n$  which label the degenerate distributions  $\{\widehat{F}_i\}$  in the obvious way.

Again, with  $0 < \lambda_a := a/n < 1$  denoting the proportion of cases in  $\emptyset \subset A \subset N$ , the identity:

$$\widehat{F} = (1 - \lambda_a)\widehat{F}_{-A} + \lambda_a\widehat{F}_A \quad (7)$$

has an exactly analogous probability vector form:

$$p_0 = (1 - \lambda_a)p_{-A} + \lambda_a p_A. \quad (8)$$

Finally, let  $T[\cdot]$  denote any statistic of interest. Following [4], perturbation is defined here as movement  $p \rightarrow p^*$  between probability  $n$ -vectors, with primary effect (corresponding to the identity functional  $T$ ) the induced change  $\widehat{F}(p) \rightarrow \widehat{F}(p^*)$  in distribution, and general effect  $T[\widehat{F}(p)] \rightarrow T[\widehat{F}(p^*)]$ .

## 4.2 Size and direction of perturbations

Again following arguments set out in [4], the second relaxation step embeds  $\mathbb{P}^n$  in  $n$ -dimensional Euclidean space  $\mathbb{E}^n$ , this choice of geometry assigning both *size* and *direction* to perturbations.

In particular, the size  $r_{-m}^{(n)} \equiv r_h^{(n)} = \sqrt{m/(nh)}$  of the perturbation  $p_0 \rightarrow p_{-M}$  (*not*, note, of its primary effect  $\widehat{F} \rightarrow \widehat{F}_{-M}$ ):

- (i) does not depend on which  $m$  cases are deleted,
- (ii) increases with  $m$  for fixed  $n$ , and
- (iii) decreases with  $n$  for fixed  $m$ ,

each of which is intuitive.

Again, for any nonzero vector  $v$  in  $\mathbb{E}^n$ , let  $d(v) := v / \|v\|$  denote its direction. Then, for any  $\emptyset \subset A \subset N$ ,  $d_A := d(p_A - p_0)$  and  $d_{-A} := d_{A^c}$  are the directions of the perturbations (from  $p_0$ ) which *hold onto* and *miss out*  $A$ , respectively. In particular, (7) and (8) can be tellingly re-expressed as

$$d_A = -d_{-A}. \quad (9)$$

In words, for any nonempty proper subset of cases, the perturbation which holds onto it is in the *opposite* direction to that which misses it out.

Finally, let  $\{M_r : r = 1, 2, 3\}$  denote a tripartition of  $N$ . Then it is easy to see that the perturbations  $\pm d_{M_1}$  (from  $p_0$ ) holding onto and missing out  $M_1$  are *orthogonal* to those,  $\pm d(p_{M_2} - p_{M_3})$ , which trade probability weight between the cases labelled  $M_2$  and  $M_3$ , exactly similar relations holding under cyclic permutation of subscripts.

### 4.3 Convexification of the feasible region

Recalling that  $\mathbb{V}_{-m}^n$  labels the distributions over which an optimum is sought, the third relaxation step is to embed  $\mathbb{V}_{-m}^n$  in its convex hull,  $\mathbb{P}_{-m}^n$  say, this larger set serving (below) as the feasible region for the smooth embedding of Problem 1.

It follows that:

$$\mathbb{P}_{-m}^n = \{p \in \mathbb{P}^n : p_i \leq h^{-1} (i \in N)\}, \quad (10)$$

a closed convex polyhedron of maximal dimension  $(n - 1)$  in  $\mathbb{E}^n$ . And, dually, that  $\mathbb{V}_{-m}^n$  is the set of all *vertices* (extreme points) of  $\mathbb{P}_{-m}^n$ . That is, all those members of  $\mathbb{P}_{-m}^n$  which cannot be written as a strict convex combination of two other members. Geometrically, all those points in  $\mathbb{P}_{-m}^n$  which do not lie in the interior of a line segment joining two others. Again, we have:

$$\{p_0\} = \{p \in \mathbb{P}^n : p_i \leq n^{-1} (i \in N)\} \subset \mathbb{P}_{-1}^n \subset \mathbb{P}_{-2}^n \subset \dots \subset \mathbb{P}_{-(n-1)}^n = \mathbb{P}^n \quad (11)$$

while, writing  $\mathbb{P}_{-m}^n$  as  $\mathbb{P}_h^n$ , the  $(\mathcal{R})$  form is immediate.

### 4.4 Examples

[FIGURE 1 ABOUT HERE]

Figure 1 illustrates the  $n = 3$  case.  $\mathbb{P}^3 = \mathbb{P}_{-2}^3$  is the outer equilateral triangle, whose vertices  $\mathbb{V}_{-2}^3$  are the unit vectors.  $\mathbb{P}_{-1}^3$  is the inverted, inner equilateral triangle, whose vertices  $\mathbb{V}_{-1}^3$  are the midpoints of the sides of  $\mathbb{P}^3$ . Both triangles are centred on  $p_0$ . All perturbations (from  $p_0$ ) which miss out a single case are the same size, and smaller than all which miss out two. Again, each perturbation (from  $p_0$ ) that holds onto a given case is in the opposite direction to that which misses it out, and orthogonal to that which trades weight between the other two.

[FIGURE 2 ABOUT HERE]

The  $n = 4$  case is illustrated in the 3-D polyhedra of Figure 2. The leftmost of these is the regular triangular pyramid  $\mathbb{P}^4 = \mathbb{P}_{-3}^4$ , whose vertices  $\mathbb{V}_{-3}^4$  (again, the unit vectors) are shown as solid circles. The four square symbols shown there are the vertices  $\mathbb{V}_{-1}^4$ , each  $p_{-\{i\}}$  being the centroid of the face of  $\mathbb{P}^n$  opposite to  $p_{\{i\}}$ , (a result that holds for any  $n > 1$ ). Again, the six oval symbols at the mid-points of the edges of  $\mathbb{P}^4$  are the vertices  $\mathbb{V}_{-2}^4$ . The convex hulls  $\mathbb{P}_{-1}^4$  and  $\mathbb{P}_{-2}^4$  of these two vertex sets comprise the other two polyhedra shown, all three being centred on  $p_0$ . The inclusions (11) are clear.

Overall, the three sides of  $\mathbb{P}^3$  are scaled copies of  $\mathbb{P}^2$ , each being the region where zero weight is attached to a given case. For the same reason, the four faces of  $\mathbb{P}^4$  are scaled copies of  $\mathbb{P}^3$ , similar results holding in general.

#### 4.5 A smooth reformulation

Now, exploiting (6), we define the *case sensitivity function*  $T(\cdot)$  for the statistic  $T[\cdot]$  via  $T(p) := T[\widehat{F}(p)]$ . Similarly, we define the *smooth target function*  $t(\cdot)$  via  $t(p) := t[\widehat{F}(p)]$ . In particular:

$$t_{MCD}(p) = \log(\det(\text{cov}[\widehat{F}(p)]))$$

while:

$$t_{Cook}(p) = (ks^2)^{-1}(\widehat{\beta}(p) - \widehat{\beta})^T X^T X(\widehat{\beta}(p) - \widehat{\beta})$$

where  $\widehat{\beta}(p) := \beta[\widehat{F}(p)]$ .

The final relaxation step is to embed Problem 1 in:

**Problem 2** ( $O(n)$  smooth reformulation of Problem 1)

Optimise  $t(p)$  over  $p \in \mathbb{P}_{-m}^n \equiv \mathbb{P}_h^n$ .

It follows at once that any concave (respectively, convex) smooth target function  $t(\cdot)$  attains its minimum (respectively, maximum) over the feasible region  $\mathbb{P}_{-m}^n \equiv \mathbb{P}_h^n$  of Problem 2 at a member of the feasible region  $\mathbb{V}_{-m}^n \equiv \mathbb{V}_h^n$  of Problem 1 and, in the strict case, only at such a vertex.

In particular, [7] show that  $t_{MCD}(\cdot)$  is concave, exploiting this in their *smooth-MCD* algorithms. Although its convexity in a neighbourhood of  $p_0$  need not extend to the whole feasible region of Problem 2, [4] present numerical results which support the conjecture that  $p$ -generalised Cook's distance  $t_{Cook}(\cdot)$  enjoys similar extremal properties (as they note, it would be helpful to have either a proof of – or counterexample to – such a conjecture). We note, in passing, that further positive evidence for it turns up in Figure 3 of the following section.

## 5 Visual displays of multiple case effects

One outcome of the above relaxation strategy is the availability of visual displays offering insight into the nature and variety of multiple case effects that can occur in different contexts. We focus here on graphs of  $t_{\text{Cook}}(\cdot)$  in the linear model, following [10] from which Figure 3 is taken.

For all but the smallest values of  $n$ , direct visualisation of the graph of any smooth target function  $t(\cdot)$  over  $\mathbb{P}^n$  – or one of its subsets  $\mathbb{P}_{-m}^n$  – is prevented by the fact that each has dimension  $(n - 1)$ . Instead, the approach adopted here uses tripartitions of  $N$  as devices providing informative triangular subsets of  $\mathbb{P}^n$ , over which the graph of  $t(\cdot)$  can then be displayed. The key idea is to attach *equal* probability weight to cases in the same member of a tripartition. This turns out to be a rich enough structure to provide insight into a range of multiple case effects – allowing us, in effect, to *see* the nature of each, and their variety.

### 5.1 Tripartitions

Suppose then that  $\mathcal{M} := \{M_r : r = 1, 2, 3\}$  is a given partition of  $N$  into three disjoint subsets, with  $m_r := |M_r| > 0$  and  $\sum_r m_r = n$ , and let

$$\mathbb{T} = \mathbb{T}(\mathcal{M}) := \{p \in \mathbb{P}^n : [i \in M_r, j \in M_r] \Rightarrow p_i = p_j\}.$$

It follows that  $\mathbb{T}$  is the convex hull of  $\{p_{M_r} : r = 1, 2, 3\}$ . That is,  $\mathbb{T}$  is the triangle which has these three points as vertices which, when convenient, we abbreviate to  $\{M_r\}$ . Otherwise said,  $p \in \mathbb{P}^n$  belongs to  $\mathbb{T}$  if and only if, for some  $\pi \equiv (\pi_r) \in \mathbb{P}^3$ ,  $p = \sum_r \pi_r p_{M_r}$ . In this case,  $\pi = \pi(p)$  is *unique*,  $\pi_r(p)$  being the total probability assigned (equally) by  $p$  to the  $m_r$  cases in  $M_r$ .

Accordingly, we may identify  $\mathbb{T}$  with  $\mathbb{P}^3$  via the bijection  $p \leftrightarrow \pi(p)$ . For example,  $p_0 \leftrightarrow (\kappa_r)$ , where  $\kappa_r := m_r/n$  is the proportion of cases in  $M_r$ . However, whereas  $\mathbb{P}^3$  is a fixed equilateral, the shape and size of  $\mathbb{T}$  vary with the  $\{m_r\}$ . Nevertheless, important inclusions, collinearities and orthogonalities in  $\mathbb{P}^3$  survive in  $\mathbb{T}$  for every  $\mathcal{M}$ .

Two obvious cyclic permutations applying, the identity:

$$p_{-M_1} = (1 - \kappa_1)^{-1}(\kappa_2 p_{M_2} + \kappa_3 p_{M_3})$$

shows that  $p_{-M_1}$  lies on the  $M_2 M_3$  side of  $\mathbb{T}$ , being closer to whichever vertex labels the larger number of cases. In particular, writing  $p_r(\lambda) := (1 - \lambda)p_{-M_r} + \lambda p_{M_r}$ , the line segment  $\mathbb{L}_r := \{p_r(\lambda) : \lambda \in [0, 1]\}$  lies in  $\mathbb{T}$ , all three such meeting at  $p_0$  by (8). Again, using Section 4.2, each  $\mathbb{L}_r$  is orthogonal to the side of  $\mathbb{T}$  containing  $p_{-M_r}$ ,  $\mathbb{S}_{-r}$  say, along which probability weight is traded between the other two subsets. Thus, the probability attached to  $M_r$  increases linearly along  $\mathbb{L}_r$  from zero at the  $p_{-M_r}$  end to unity at the other. Indeed, for each  $\lambda \in [0, 1]$ , this probability is constant at the value  $\lambda$  for all points in  $\mathbb{T}$  on the line through  $p_r(\lambda)$  parallel to  $\mathbb{S}_{-r}$ . In particular, it vanishes on  $\mathbb{S}_{-r}$ .

## 5.2 Four multiple case effects in the linear model

[3] and [11] discuss a variety of possible effects that a pair of cases may have on Cook's distance. Here, with  $M_3$  representing a convenient 'null' data set, and restricting ourselves to the special case  $m_1 = m_2 = 1$  (for a fuller account, see [10]), we consider four effects defined in the table below, and illustrated in the corresponding rows of Figure 3:

	<b>Effect</b>	<b>Joint presence of <math>M_1</math> and <math>M_2</math> ...</b>
(a)	Masking	... conceals presence of either
(b)	Cancellation	... has no effect on fitted line
(c)	Swing	... swings fitted line, (intercept $\sim$ unchanged)
(d)	Raise & Lower	... translates fitted line, (slope $\sim$ unchanged)

For clarity, stylised simple linear regression data sets are used, shown in the middle column of Figure 3. In each case,  $M_3$  contains  $m_3 = 20$  points, comprising five replicates at each corner of the square  $\{\pm 1\}^2$ , whose fitted line is the horizontal axis. Both  $M_1$  and  $M_2$  consist of a single point at the corner of  $\{\pm 4\}^2$  indicated.

### [FIGURE 3 ABOUT HERE]

The righthand column of Figure 3 gives the corresponding graph of  $t_{Cook}(\cdot)$  over  $\mathbb{T}$ , limits being used where needed (since, of course, a line cannot be fitted to a single case). Some linear rescaling between plots has been applied, both vertically and horizontally, to enhance their visual clarity (a minor cost being some loss of visual perception that the angle at  $M_3$  exceeds  $87^\circ$ ). Note that  $p_0$  (corresponding to  $\hat{F}$ ) is close to  $M_3$ , being just one-eleventh of the way along the line  $\mathbb{L}_3$  joining  $M_3$  to the midpoint of the opposite side. The inbuilt  $M_1 - M_2$  symmetry is evident throughout. Overall, the four graphs have visibly different *shapes*, discussed next:

**(a) Masking.** The 'spike' at  $M_3$  reflects the dominant effect of removing both  $M_1$  and  $M_2$ , while the parallelism of the contours to  $\mathbb{S}_{-3}$  corresponds to the fact that there is, of course, no effect here in trading weight between these sets.

**(b) Cancellation.** The contours of  $t_{Cook}(\cdot)$  here are straight lines fanning out from  $M_3$ . In particular,  $\mathbb{L}_3$  is the zero height contour, since varying  $\pi_3$  while keeping  $\pi_1 = \pi_2$  has no effect on the fitted line. Trading weight between  $M_1$  and  $M_2$  now has a quadratic, globally dominant, effect.

**(c) Swing.** The overall shape of the surface here is very similar, but not identical, to that in the masking case. The 'spike' at  $M_3$  remains dominant, but the surface contours are no longer parallel to  $\mathbb{S}_{-3}$ .

**(d) Raise & Lower.** This is perhaps the most interesting graph. As is intuitive from the data, the dominant global effect occurs along  $\mathbb{S}_{-3}$ . Looking at the surface, we see two 'troughs'. These run along  $\mathbb{L}_1$  and  $\mathbb{L}_2$ , showing that varying the weight on one of these subsets alone has little effect. The contours of  $t_{Cook}(\cdot)$  are parallel to  $\mathbb{S}_{-3}$  when there is little weight on  $M_3$ , but become

more curved as  $\pi_3$  increases. Locally to  $p_0$ , trading weight between  $M_1$  and  $M_2$  produces the largest effects.

## 6 A relaxed diagnostic approach to detecting heavy mutual masking

Multiple case effects can be strong and yet intrinsically hard to detect with standard diagnostic procedures, while the burden of full enumeration increases combinatorially with  $m$ . Heavy mutual masking is a well-known example, challenge data sets comprising 60% of cases from one distribution and 40% from a second, suitably remote from the first. [4] present a widely applicable, relaxed, two-stage approach to detecting such effects (cf. [2]), briefly reviewed here.

Adopting the standard assumption in the literature that at most half the cases are discordant from a common pattern followed by the rest, Stage I consists on maximising (say) a suitable target function  $t(\cdot)$  over  $\mathbb{P}_{-m}^n$ , with  $m$  the integer part of  $n/2$ , the optimum being known or assumed to occur at a vertex. This corresponds precisely to missing out a specified subset  $\widehat{M}$  of  $m$  cases. The (in)equality constraints defining  $\mathbb{P}_{-m}^n$  being linear, this relaxed optimisation can be carried out with standard software (or some alternative, as indicated in Section 7). The assumed internal consistency of the cases in  $\widehat{H} := \widehat{M}^c$  may also be checked.

Stage II back-checks for swamping. That is, for cases in  $\widehat{M}$  which are *not* inconsistent with the pattern followed by the majority. [4] envisage doing this *separately* for each case in  $\widehat{M}$ , although a *sequential* approach is possible. Having augmented  $\widehat{H}$  with any such cases, a final check on their internal consistency can be made while, if required, the possibility of further structure within the cases in  $\widehat{M}$  may be made.

[4] report encouraging results for this general strategy, using regression as a test problem and several forms of challenge data set. Specifically, they maximise  $t_{Cook}(\cdot)$  in Stage I, using the mean shift outlier test in Stage II.

Finally, a remark on local maxima. On those occasions when the final check for a common pattern fails, the possibility that this is because omission of  $\widehat{M}$  is a particular form of non-trivial local maximum can be easily explored as follows. The value of  $t(\cdot)$  there can be compared to that where  $\widehat{M}$  is held onto. If this is greater, replacing  $\widehat{M}$  by its complement, and then continuing as before, is indicated. On the relatively few occasions where it was needed in their regression study, [4] report that this simple strategy was successful. The original  $\widehat{M}$  containing no mutually masked cases, moving to its complement produced a large increase in  $t_{Cook}$  and led again to correct identification of the structure in the data.

## 7 Developments in relaxed robust computation

Consider now minimisation over  $\mathbb{P}_h^n$  of the particular function  $t(\cdot) = t_{MCD}(\cdot)$  as an exemplar of the class of robust estimation procedures that can be defined in this way. Algorithms for the MCD problem include those reported in [1], [8], [9] and [12]. These are all *discrete* in the sense that they address Problem 1, iteratively ‘jumping’ between members of  $\mathbb{V}_h^n$ .

We briefly sketch here some of the work reported in [6] and, more fully, in [7], recalling that these papers show that  $t_{MCD}(\cdot)$  is, indeed, concave. Collectively, the new approaches reported therein are referred to as *smooth-MCD* algorithms.

**[FIGURE 4 ABOUT HERE]**

Figure 4 shows two views of the same  $t_{MCD}$  surface over  $\mathbb{P}_2^3$  for univariate data. This simple example offers some general geometric insight: the graph of  $t_{MCD}$  contains multiple local minima, separated by hills, with corresponding limitations for any purely descent algorithm. In particular, it motivates the use of *swapping* strategies aimed at ‘getting you over a hill to a lower valley’. At the same time, the swapping strategy employed by the feasible subsets algorithm – while optimal in its own terms – is relatively expensive to perform and may not always be needed, in the sense that not every vertex is a local minimum.

Again, [4] note the benefits of using explicit gradient information, when this is available. [5] develop local *projected* (here, *centred*) Taylor expansions in generality. They show that such expansions are possible even when, as here, one or more constraints (here,  $p^T \mathbf{1}_n = 1$ ) imply that there are *no* open sets in a function’s domain (here, a subset of  $\mathbb{P}^n$ ). Indeed, they exist uniquely under mild conditions and can be used to guide algorithms downhill, in the usual way. They also provide also a useful necessary and sufficient condition for a vertex in  $\mathbb{V}_h^n$  to be a local minimum, for any  $t$ . In the  $t_{MCD}$  case, it is shown that these are precisely the points where the C-steps of FAST-MCD converge.

Now, conditional on robustness, there are two key performance criteria in any problem such as this: speed and optimality. Perfection (*i.e.* instant, global optimality!) being unachievable, different algorithms aim for it, while striking different trade-offs between these criteria. Accordingly, the state-of-the-art can be thought of as a boundary of limiting speed/optimality trade-offs that are currently feasible, the different algorithms appearing at different points along it.

[6] and [7], to which the reader is referred for further details, exploit features of the case sensitivity function approach – in particular, insights from (convex) geometry, the power of analysis, and a unifying structure – both to understand better *why* current algorithms occur where they do along this boundary, and to add new algorithms that fill it out and/or nudge it nearer to perfection.

## References

- [1] Agulló J. (1998). *Computing the minimum covariance determinant estimator*. Technical report, Universidad de Alicante.
- [2] Atkinson A.C. (1986). *Masking unmasked*. *Biometrika* **73**, 533 - 541.
- [3] Barrett B.E. and Gray J.B. (1997). *Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression*. *Computational Statistics and Data Analysis* **26**, 39 - 52.
- [4] Critchley F., Atkinson R.A., Lu G. and Biazi E. (2001). *Influence analysis based on the case sensitivity function*. *J. Royal Statistical Society*, **B 63**, 307 - 323.
- [5] Critchley F., Lu G., Atkinson R.A. and Wang D.Q. (2003). *Projected Taylor expansions for use in Statistics*. Under consideration.
- [6] Critchley F., Schyns M. and Haesbroeck G. (2003). *Smooth optimization for the MCD estimator*. *International Conference on Robust Statistics*, Antwerp, 29 - 30.
- [7] Critchley F., Schyns M., Haesbroeck G., Lu G., Atkinson R.A. and Wang D.Q. (2004). *A convex geometry approach to algorithms for the MCD method of robust statistics*. Under consideration.
- [8] Hawkins D.M. (1994). *A feasible solution algorithm for the minimum covariance determinant estimator in multivariate data*. *Computational Statistics and Data Analysis* **17**, 197 - 210.
- [9] Hawkins D.M. and Olive D.J. (1999). *Improved feasible solution algorithms for high breakdown estimation*. *Computational Statistics and Data Analysis* **30**, 1 - 11.
- [10] Kinns D.J. (2001). *Multiple case influence analysis with particular reference to the linear model*. PhD thesis, University of Birmingham.
- [11] Lawrance A.J. (1995). *Deletion influence and masking in regression*. *J. Royal Statistical Society*, **B 57**, 181 - 189.
- [12] Rousseeuw P.J. and Van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant estimator*. *Technometrics* **41**, 212 - 223.

*Acknowledgements:* The UK authors are grateful for EPSRC support under research grant GR/K08246 and to D.Q. Wang for helpful discussions.

*Addresses* (in author order): The Open University, Milton Keynes; University of Namur; University of Liège; (formerly) University of Birmingham; University of Birmingham and University of Bristol.

*E-mail:* F.Critchley@open.ac.uk