The Annals of Statistics
1993, Vol. 21, No. 4, 1843–1899

ASPECTS OF ROBUST LINEAR REGRESSION[1]

By P. L. Davies

University of Essen

Section 1 of the paper contains a general discussion of robustness. In Section 2 the influence function of the Hampel–Rousseeuw least median of squares estimator is derived. Linearly invariant weak metrics are constructed in Section 3. It is shown in Section 4 that S-estimators satisfy an exact Hölder condition of order $1/2$ at models with normal errors. In Section 5 the breakdown points of the Hampel–Krasker dispersion and regression functionals are shown to be 0. The exact breakdown point of the Krasker–Welsch dispersion functional is obtained as well as bounds for the corresponding regression functional. Section 6 contains the construction of a linearly equivariant, high breakdown and locally Lipschitz dispersion functional for any design distribution. In Section 7 it is shown that there is no inherent contradiction between efficiency and a high breakdown point. Section 8 contains a linearly equivariant, high breakdown regression functional which is Lipschitz continuous at models with normal errors.

1. A long introduction.

1.1. *Contents*. Subsections 1.2 to 1.6 contain some general and also critical remarks on robustness and optimality, on the properties of influence functions, on the calculation of breakdown points, on the use of metrics and on the inherent instability of S-estimators. They are intended to motivate the remainder of the paper. The general thrust of the paper is that stability of inference is to be obtained by the construction of functionals with specified properties rather than by the derivation of functionals which are optimal in some sense. Stability of inference is taken to include insensitivity to minor changes in the data, insensitivity to minor changes in the model, uniform asymptotic behaviour and resistance to outliers. The word "minor" will be quantified in terms of metrics. One result of the paper is that the existence and boundedness of the influence function will not guarantee stability. This is done in Section 2 where the influence functions of the middle of the shortest half and the Hampel–Rousseeuw least median of squares estimators are derived. This is of some interest as it has been generally assumed that they do not exist. It turns out that the influence function of the middle of the shortest half is bounded but that of the least median of squares estimator is not. The rather surprising reason for this latter result is the effect of inliers. This may

Received May 1990; revised December 1992.
[1]This work was partially supported by DFG-Grant 51167.
AMS 1991 *subject classifications*. 62F35, 62J05.
Key words and phrases. Influence function, Lipschitz continuity, global definability, breakdown point, linear equivariance.

go some way in explaining some results of Hettmansperger and Sheather (1992) regarding the effect of inlying observations on the least median of squares functional. This section also contains a criticism of the derivation of optimal bounded influence regression functionals. It is argued that such functionals are not GM-estimators as the effect of alterations in the dispersion matrix are not considered.

In Section 3 linearly invariant weak metrics are constructed and used to calculate breakdown points. The metrics are defined in terms of Vapnik–Cervonenkis classes and permit, for example, direct comparisons of theoretical and empirical distributions. The linear invariance of the metrics reflects the invariance of the regression problem under certain linear transformations. The metrics can be related to the gross error model. The use of metrics for one-dimensional data is well established; in particular, Huber (1981) makes considerable use of them. It seems that metrics have not been applied to higher-dimensional problems, perhaps because of the lack of suitable candidates. It is hoped that the metrics introduced in Section 3 and other parts of the paper will go some way to redressing this situation.

Section 4 is devoted to the properties of S-estimators such as existence, uniqueness, continuity and breakdown points. It is shown that they are Hölder continuous of exact order $1/2$ at normal models but that uniqueness can only be obtained by imposing conditions on the distribution under consideration. Furthermore, these conditions are such that, given any distribution for which they hold, there are other distributions arbitrarily close for which they do not hold. This reflects an inherent instability of S-estimators. Although S-estimators are shown to be capable of obtaining the highest possible breakdown point of $1/2$, this can only be done by choosing the estimator according to the distribution. If this is not done, then it is surmised that the highest breakdown point is $1/3$.

In Section 5 the breakdown point of the Hampel–Krasker dispersion functional is shown to be 0. The exact breakdown point of the Krasker–Welsch dispersion functional is not 0 and is given by Theorem 5.2. All breakdown points are calculated in terms of metrics as constructed in Section 3. The breakdown point of the Hampel–Krasker regression functional is also shown to be 0, whilst nontrivial bounds for the breakdown point of the Krasker–Welsch regression functional are obtained, again using metrics constructed as in Section 3.

The most important positive result of the paper is the construction in Section 6 of a globally defined, high breakdown, linearly equivariant and locally Lipschitz continuous dispersion functional. This is the first step in the programme mentioned previously of constructing functionals with given properties.

In Section 7 the relationship between efficiency, high breakdown and Lipchitz continuity is discussed. It is shown that there is no conflict between efficiency and high breakdown but that there is such a conflict between Lipschitz continuity and efficiency.

Finally, Section 8 is devoted to the construction of a regression functional which has a high breakdown point, is linearly equivariant and Lipschitz continuous at models with normal errors. This again is intended as a part of the programme of constructing functionals with given properties but the problem is more difficult than that of construction dispersion functionals with given properties. The solution is less satisfactory.

1.2. *Huber's approach to robustness.* Huber (1981) has given a beautiful minimax approach to the problem of estimating a location parameter for one-dimensional data. He considered and solved the problem of minimizing the maximum bias over a neighbourhood of the normal model. The problem of minimizing the maximum asymptotic variance was not completely solved but Huber showed that there are M-estimators which perform remarkably well in this respect. Recently Riedel (1989a, 1989b, 1991) has extended the first part of Huber's theory by determining minimax bias estimators in some very general situations. The problem of extending Huber's partial result on asymptotic variance to more complex situations has proved intractable. Nevertheless, the importance of obtaining exact optimality results should not be overstated. Optimal estimators are not for applications to data; they are border posts delimiting the possible. This is amply illustrated by the median which is optimal in the sense of minimizing the maximum bias but whose asymptotic variability in a full neighbourhood of any distribution exhibits the whole gamut of possibilities. These range from almost sure convergence in a finite number of steps to no convergence at all with asymptotic normality somewhere in between. Any estimator intended for use on data must be a compromise; it would seem to me that this is always the case. A discussion may be found on pages 44–47 of Hampel, Rousseeuw, Ronchetti and Stahel (1986). Given this, the practical problem is that of producing an estimator which behaves well, or at least not disastrously, in several different respects. In order to do this, it is not necessary to know what the limits are for any particular aspect. At the risk of labouring the point, we consider the location problem for one-dimensional data with known scale. The M-estimator with psi-function ψ = atan may well be judged to be satisfactory without knowing the minimum obtainable values for the maximum bias and the maximum asymptotic variance. The bias and asymptotic variance of the atan-M-estimator may be calculated, or at least upper bounds may be obtainable, and judged to be sufficiently small for the purpose in mind.

At first sight the problem of producing a compromise estimator may not seem particularly difficult, but no such estimator exists for the linear regression model. Huber's problem may be stated as follows. We take a given model $\mathbb{P}_0$, say the $\mathfrak{N}(0,1)$-distribution, and a full neighbourhood of this model defined by some metric, say the Kolmogorov metric d_K. The neighbourhood is given by

$$B_\varepsilon(\mathbb{P}_0, d_K) = \{\mathbb{P}: d_K(\mathbb{P}_0, \mathbb{P}) < \varepsilon\}.$$

The bias problem is to determine a translation equivariant functional T_1 such that

$$\sup_{\mathbb{P}\in B_\varepsilon(\mathbb{P}_0,d_K)} |T_1(\mathbb{P}) - T_1(\mathbb{P}_0)| = \inf_T \sup_{\mathbb{P}\in B_\varepsilon(\mathbb{P}_0,d_K)} |T(\mathbb{P}) - T(\mathbb{P}_0)|,$$

where the infimum is taken over all translation equivariant functionals T.

The asymptotic variance problem is the following. We denote by $\hat{\mathbb{P}}_n(\mathbb{P})$ the empirical distribution of n independently and identically distributed random variables with common distribution $\mathbb{P}$. We restrict attention to translation equivariant functionals T such that $\sqrt{n}\,(T(\hat{\mathbb{P}}_n(\mathbb{P})) - T(\mathbb{P}))$ is asymptotically normally distributed with mean 0 and variance $A(T,\mathbb{P})$. We now wish to determine a functional T_2 with

$$\sup_{\mathbb{P}\in B_\varepsilon(\mathbb{P}_0,d_K)} A(T_2,\mathbb{P}) = \inf_T \sup_{\mathbb{P}\in B_\varepsilon(\mathbb{P}_0,d_K)} A(T,\mathbb{P}).$$

A further important demand which should be placed on a functional T is that the convergence of $\sqrt{n}\,(T(\hat{\mathbb{P}}_n(\mathbb{P})) - T(\mathbb{P}))$ to $\mathfrak{N}(0, A(T,\mathbb{P}))$ be uniform for $\mathbb{P} \in B_\varepsilon(\mathbb{P}_0, d_K)$. This is explicitly mentioned on page 51 of Huber (1981). In his book Huber obtains the uniform convergence to the normal distribution by placing smoothness conditions on the psi-function of the M-estimator. Essentially these conditions are such as to cause the estimator to be Fréchet differentiable at each point $\mathbb{P}$ of $B_\varepsilon(\mathbb{P}_0, d_K)$ [see Clarke (1983)] with an influence function $I(x,\mathbb{P},T)$ which is bounded with respect to x *and* $\mathbb{P}$ in $B_\varepsilon(\mathbb{P}_0, d_K)$. There are therefore good reasons for demanding of a functional that it should be Fréchet differentiable at each point of the neighbourhood.

If we return to the M-estimator with psi-function atan, then it may be checked that it behaves well in all respects. For any fixed $\varepsilon < 1/2$ we have the following. The functional is uniquely defined at all $\mathbb{P}$ in $B_\varepsilon(\mathbb{P}_0, d_K)$, it is translation equivariant, its bias remains bounded, it is Fréchet differentiable at each $\mathbb{P}$ in $B_\varepsilon(\mathbb{P}_0, d_K)$, $\sqrt{n}\,(T(\hat{\mathbb{P}}_n(\mathbb{P})) - T(\mathbb{P}))$ converges to a normal distribution and its asymptotic variance is not too large. Furthermore, the convergence to the normal distribution is uniform over the ball $B_\varepsilon(\mathbb{P}_0, d_K)$.

To state the corresponding problem for the linear regression model, we require some notation. A regression distribution is a probability distribution $\mathbb{Q}$ on the Borel sets $\mathfrak{B}(\mathbb{R}^{k+1})$ of $\mathbb{R}^{k+1}$. The first k components are the x's and the $(k+1)$th component is the y. The marginal distribution $\mathbb{Q}_d$ defined by $\mathbb{Q}_d(B) = \mathbb{Q}(B \times \mathbb{R})$ for $B \in \mathfrak{B}(\mathbb{R}^k)$ is the distribution of the design points and will be called the design distribution. If the x's in the linear regression model are linearly dependent, it is clear that the problem of estimating the regression coefficients has no unique solution. Attention is consequently restricted to the regression distribution $\mathbb{Q}$ for which the design distribution $\mathbb{Q}_d$ is nonsingular, that is, the support of $\mathbb{Q}_d$ is not contained in a plane of dimension less than k. We denote the set of all such regression distributions by $\mathfrak{W}(\mathbb{R}^{k+1})$. A regression functional T is a mapping of $\mathfrak{W}(\mathbb{R}^{k+1})$ into $\mathbb{R}^{k+1}$. The first k components are estimates of the regression coefficients and the last component is an estimate of the dispersion of the errors. The translation equivariance of the functionals for the one-dimensional location problem is replaced by linear equivariance. The regression problem is not invariant under all linear transformations but only under a certain group of linear translations, essentially those which leave the "y" on the correct side of the equation. We consider only such functionals which are linearly equivariant in this sense. In place of $\mathbb{P}_0$ we take some regression distribution $\mathbb{Q}_0$ with, for example, errors which are normally distributed with mean 0 and variance σ^2. In order to define a full neighbourhood of $\mathbb{Q}_0$, we require a metric d, the choice of which will be discussed below. A direct translation of the properties of the atan-estimator for the location problem would lead to the search for a functional T which has low bias over $B_\varepsilon(\mathbb{Q}_0, d)$, which is Fréchet differentiable at each $\mathbb{Q}$ in $B_\varepsilon(\mathbb{Q}_0, d)$ and whose influence function is uniformly bounded. In particular, it is necessary for T to be well defined, that is, to exist and be uniquely defined at each $\mathbb{Q}$ in $B_\varepsilon(\mathbb{Q}_0, d)$. Indeed, if this is not the case $\mathbb{Q}$ cannot even be continuous let alone Fréchet differentiable.

We are not yet able to produce a functional which satisfies these demands as they stand. One reason is the fact that the linear regression model is only partially parametric. In general, the design distribution $\mathbb{Q}_{0d}$ of $\mathbb{Q}_0$ is regarded as given, there does not exist a parametric model for it. Nevertheless, one is interested in identifying leverage points and this leads to the stipulation that T should be well defined for any $\mathbb{Q}$ within a neighbourhood of $\mathbb{Q}_0$ with $\mathbb{Q}_{0d}$ arbitrary. From this it is a short step to require that T be globally defined, that is, $T(\mathbb{Q})$ should be well defined for all regression distributions $\mathbb{Q}$. We note that this requirement is not trivial. If we consider, for example, M-estimators, it seems that conditions have to be placed on $\mathbb{Q}$ to guarantee the existence and uniqueness of an M-estimator. Theorem 2.2 of Maronna and Yohai (1981) gives sufficient conditions for the uniqueness of a general M-estimator, a so-called GM-estimator. In particular, they assume that the error at the design point x has a symmetric distribution for each x. In any neighbourhood of such a regression distribution, there are other distributions where the errors are not symmetric and where consequently the theorem is not applicable. It is therefore not clear whether GM-estimators can be uniquely defined at all regression distributions.

The demand that the bias of the estimator remain small requires the calculation of bias. Although this may be possible we restrict ourselves to the weaker stipulation that the estimator should have a large breakdown point. The breakdown point $\varepsilon^*(T, \mathbb{Q}, d)$ of T at the distribution Q is defined by

$$\varepsilon^*(T, \mathbb{Q}, d,) = \inf\left\{\varepsilon > 0: \sup_{\mathbb{Q}' \in B_\varepsilon(\mathbb{Q}, d)} \|T(\mathbb{Q}')\| = \infty\right\}.$$

The breakdown point depends on the distribution $\mathbb{Q}$ where it is evaluated and is therefore a local property of T. We require that T should have a high local breakdown point at each $\mathbb{Q}$.

Finally, we require that T should be locally Fréchet differentiable at each $\mathbb{Q}$ or, failing that, at least locally Lipschitz continuous. A functional which is globally defined and locally Lipschitz continuous will have the following properties. Small changes in the distribution will cause only small changes in the value of the functional. In particular, the functional will be well defined at any empirical distribution and will react smoothly to small changes in the data. These include large changes in a small number of data points, small changes in a large number of data points or a combination of the two. It has been pointed out by Hettmansperger and Sheather (1992) that the Hampel–Rousseuw least median of squares estimator does not react smoothly to small changes in centrally located data points. A possible explanation of this is offered in Section 2.3 where the influence function is shown to be unbounded with respect to inlying design points. In a recent paper Simpson, Ruppert and Carroll (1992) consider the problem of obtaining stable inferences in linear regression. The idea is to combine high breakdown, efficiency and bounded influence. This has some similarities with the approach adopted here although we shall argue strongly that bounded influence alone is not sufficient to give stable inferences.

As already mentioned we are not able to exhibit a functional T with the required properties. On the positive side we can construct a functional T which fulfills the demands for those $\mathbb{Q}$ which have independent normal errors. This is done in Section 7 and corresponds to the construction given in Davies (1992b) for multivariate location and dispersion functionals. Another and more important positive result is the construction of a dispersion function T^D on the space of design distributions $\mathfrak{W}_d(\mathbb{R}^k) = \{\mathbb{Q}_d : \mathbb{Q} \in \mathfrak{W}(\mathbb{R}^{k+1})\}$. A dispersion functional T^D associates with each $\mathbb{Q}_d$ in $\mathfrak{W}_d(\mathbb{R}^k)$ a positive definite symmetric $k \times k$ matrix $T^D(\mathbb{Q}_d)$. The problem of dispersion functionals is somewhat easier, and we construct one in Section 6 with the following properties: (i) global definability, (ii) linear equivariance, (iii) high local breakdown point and (iv) local Lipschitz continuity. It turns out that the local breakdown point is at most $1/3$ in contrast to the figure of $1/2$ which is the highest local breakdown point for linearly equivariant functionals. The reason for the difference is the global definability of the functional, discussed in more detail in subsection 1.4.

1.3. *The approach based on influence functions.* Another approach to robustness is that based on the concept of influence functions. It is associated with Hampel [(1968), (1971), (1974)] and expounded in Hampel, Rousseeuw, Ronchetti and Stahel (1986). The idea is to place an upper bound on the supremum of the influence function and then to maximize the efficiency subject to this bound. An extension is to consider the change of variance function, to place an upper bound on its supremum and again to maximize the efficiency subject to this bound. The influence function measures the effect of an infinitesimal one-point contamination which, from the point of view of robustness, is somewhat problematic. The analysis is therefore augmented by an investigation of the breakdown point of the functional. The idea is that the local bias will be reflected by the influence function and the global bias by the breakdown point. We refer to Figure 2 on page 42 of Hampel, Rousseeuw, Ronchetti and Stahel (1986). In Section 2 we give an example where this does not hold, namely, the middle of the shortest half sample. We show that at the normal distribution the influence function exists and is bounded. In spite of this the influence function does not reflect the local behaviour of the functional.

The approach based on influence functions has the advantage of leading to optimal functionals in situations where Huber's minimax approach is not possible. In the case of a one-dimensional location parameter and the model of a normal distribution, the influence function approach and the change of variance approach both lead to the median as the most robust functional or to M-estimators with Huber's psi-function which maximize the efficiency subject to the respective bounds. The breakdown point of these functionals is $1/2$. In k dimensions the optimal location and dispersion estimators are M-estimators which are known to have a breakdown point of at most $1/(k+1)$. Similarly in the linear regression model the approach leads to general M-estimators, GM-estimators, such as the Krasker–Welsch estimator with low breakdown points in high dimensions. In Section 2 we criticize the usual derivation of the influence function for such estimators. Given the low breakdown points of the optimal estimators, the statistician is once again forced to look for estimators which are not optimal, but which have given properties.

In the influence function approach the model plays a much more important role than it does in Huber's minimax approach. The concept of "efficiency at the model" is not a robust concept in the following sense. Given any location model such as the normal distribution, there exist other location models which are empirically indistinguishable from the normal model and for which no likelihood function exists. Such models permit a super efficient estimator of the location parameter. The very concept of "efficiency" is seen to be nonrobust and the influence function approach is presumably restricted to "sensible" models. Huber's approach does not suffer from this defect. If the normal model is replaced by an empirically indistinguishable model, then the resulting optimal estimators are hardly altered. In this sense Huber's approach is "model robust," whereas the influence function approach is not.

It is clear that for regression problems with normal errors there is a tradeoff between gross error sensitivity and efficiency. The example of S-estimators, whose influence function is unbounded in the factor space, shows that there can be a tradeoff between breakdown point and efficiency. This fact caused a search for regression functionals which do not have this deficiency. Jurečková and Portnoy (1987), Yohai (1987) and Yohai and Zamar (1988) succeeded in showing how to combine high breakdown point and efficiency in certain situations. These situations effectively exclude large leverage points. It has recently been suggested by Morgenthaler (1989) and Stefanski (1991) that in general there is a conflict between breakdown point and efficiency, namely any high breakdown estimator can have arbitrarily low efficiency with respect to least squares. In Section 7 we discuss this problem and show that the claim is not justified. We do, however, show that the idea of Morgenthaler and Stefanski can be made precise for locally uniformly Lipschitz regression functionals. This shows that in the regression problem the efficiency of an estimator is related to its continuity and not to its breakdown point. This point can not be stressed too greatly; an efficiency deriving from extreme leverage points is to be treated with the utmost caution.

1.4. *Breakdown points.* One of the desirable properties of an estimator is that it should have a high breakdown point. Breakdown points are calculated in three ways. The easiest and, for nonstatisticians, most understandable definition is that of the finite sample breakdown point due to Donoho and Huber (1983). It has, however, disadvantages. It does not allow for any "wobbling" of the data, and it is not defined at the model describing the data. Furthermore, it cannot be related in a natural way to the continuity of the functional. For these reasons we consider it only when comparing the results of this paper with those in the literature.

The second method of calculating the breakdown point at a distribution $\mathbb{Q}$ is to consider the "gross error neighbourhood" $(1-\varepsilon)\mathbb{Q} + \varepsilon\mathbb{Q}'$ of $\mathbb{Q}$. This also violates the spirit of robustness as it implies that a proportion $(1-\varepsilon)$ of the data actually are distributed exactly according to $\mathbb{Q}$. One reason why it is possible to calculate breakdown points using the gross error neighbourhood is that the neighbourhood is affinely equivariant. Indeed, if this were not the case, the breakdown point of a functional T would depend on the parameterization and it is clear that such a dependence would greatly complicate any calculation of the breakdown point.

Finally, breakdown points can be calculated using metrics. This is in keeping with the spirit of robustness as it allows a full topological neighbourhood of any distribution $\mathbb{Q}$. In Huber (1981) breakdown points are defined in terms of metrics (pages 11–13) and those of the one-dimensional location and scale estimators in the Lévy, Kolmogorov and Prohorov metrics are calculated (pages 52–54 and 109–110). As far as I am aware there are no corresponding results for location and dispersion functionals in higher dimensions. One reason for this may be the lack of appropriate metrics. The Prohorov metric topologizes weak convergence and has been treated favourably in the literature on robustness as it formalizes the idea of measurement error. It is, however, not linearly invariant in the following sense.

Let $A: \mathbb{R}^k \to \mathbb{R}^k$ be a nonsingular linear transformation and let $\mathbb{Q}$ and $\mathbb{P}$ be two distributions over $\mathbb{R}^k$. The transformed distribution $\mathbb{Q}^A$ is defined by

$$\mathbb{Q}^A(B) = \mathbb{Q}(A^{-1}(B)), \qquad B \in \mathfrak{B}(\mathbb{R}^k).$$

A metric d is called affinely invariant if $d(\mathbb{P}^A, \mathbb{Q}^A) = d(\mathbb{P}, \mathbb{Q})$ for all $\mathbb{P}$ and $\mathbb{Q}$ and all nonsingular affine transformations A. The Prohorov metric, being based as it is on the Euclidean norm, is not linearly invariant. To show the effect of this in the linear regression model, we consider the following example. Let $\mathbb{Q}_1$ denote a design distribution where all measurements are made at the point $x_1^T = (1, 0)^T$. This distribution is singular and it is clear that it is not possible to obtain consistent estimates of the regression coefficients. The second design distribution $\mathbb{Q}_2$ is such that half the observations are performed at x_1 and the other half at $x_2^T = (1, 1)^T$. This design distribution is not singular and consistent estimates of the regression coefficients can be obtained. Assume that the second component is measured in inches and the units are now changed to miles. The transformed design points are $x_1^T = (1, 0)^T$ and $x_2^T = (1, 1/63360)^T$. In terms of the Prohorov metric, the transformed distribution is now closer to the singular distribution than it is to the nonsingular one from which it is derived. This is at odds with the fact that the regression problem has not been altered by the linear transformation indicating that the Prohorov metric is not appropriate for the linear regression model. What are required are linearly invariant metrics. In Section 3 we introduce such metrics and give one which formalizes the idea of proportional measurement error.

Consider the finite sample breakdown point $\varepsilon_{\mathrm{FS}}^*(T)$ [Donoho and Huber (1983)] of a linearly equivariant regression functional T. It follows from the arguments given in Rousseeuw and Leroy (1987), page 125 [see also Davies (1987) and Lopuhaä and Rousseeuw (1991)], that

$$\varepsilon_{\mathrm{FS}}^*(T) \leq [(n-p)/2 + 1]/n, \tag{1.1}$$

where p is the maximum number of points on some $k - 1$-dimensional plane. The upper bound $[(n-p)/2 + 1]/n$ can be obtained by an appropriate change in the definition of the Hampel–Rousseeuw least median of square estimator, namely by minimizing the h_nth-order statistic of the absolute residuals where $h_n = [n/2] + [(p+1)/2]$. It is clear, however, that p is a function of the distribution and that altering the definition of the Hampel–Rousseeuw estimator in the manner described above is effectively choosing the functional in a manner which depends on the distribution being considered. If global functionals are considered, this is no longer possible. In Section 6 we give a global dispersion measure for design distribution. The factor $1/2$ which is usually associated with optimal breakdown points is replaced by $1/3$. We conjecture that this is best possible for global functionals.

1.5. *Metrics.* As argued in subsection 1.2 we are interested in functionals which are locally Lipschitz. As the very definition of Lipschitz continuity requires a metric, we see that metrics are an essential part of the paper. Furthermore, as argued in subsection 1.4, the notion of a breakdown point can be most satisfactorily formulated using metrics and these should be linearly invariant. This is at odds with part of the folklore of robustness, namely, that part which emphasizes the importance of the weak topology and the corresponding weak continuity of the functionals. It is easily proved that no linearly invariant metric can topologize weak convergence, giving rise to a conflict between the emphasis on linear equivariance and the use of the weak topology.

Besides being linearly invariant, we claim that a metric should be weak. No formal definition of "weakness" is given, but the following discussion should make it clear what is meant. The total deviation metric $d_{\mathfrak{T}\mathfrak{D}}$ is defined by

$$d_{\mathfrak{T}\mathfrak{D}}(\mathbb{P}, \mathbb{Q}) = \sup_{B \in \mathfrak{B}(\mathbb{R}^k)} |\mathbb{P}(B) - \mathbb{Q}(B)|. \tag{1.2}$$

It is clear that $d_{\mathfrak{T}\mathfrak{V}}$ is linearly invariant, but it is so strong as to be useless for robustness purposes. To see this, we note that the total variation distance between any continuous model and a derived empirical distribution is always 1. Furthermore, if we have an empirical distribution $\hat{\mathbb{P}}$ and move all points slightly to give another distribution $\hat{P}'$, then $d_{\mathfrak{T}\mathfrak{V}}(\hat{\mathbb{P}}, \hat{\mathbb{P}}') = 1$.

The total deviation is obtained by taking the supremum in (1.2) over all Borel sets. Replacing $\mathfrak{B}(\mathbb{R}^k)$ by a subset $\mathfrak{K}$ of Borel sets gives rise to a metric $d_{\mathfrak{T}\mathfrak{K}}$ defined by

$$d_{\mathfrak{T}\mathfrak{K}}(\mathbb{P}, \mathbb{Q}) = \sup_{B \in \mathfrak{K}} |\mathbb{P}(B) - \mathbb{Q}(B)|. \tag{1.3}$$

In general, $d_{\mathfrak{T}\mathfrak{K}}$ is a pseudometric but this will be of no relevance for our present purposes. The smaller the set $\mathfrak{K}$ in (1.3) the weaker is the metric. It seems to be the case that all useful metrics derive from classes $\mathfrak{K}$ which are Vapnik–Cervonenkis classes (classes with polynomial discrimination). We refer to Pollard (1984) for details.

If $\mathfrak{K}$ is a Vapnik–Cervonenkis class and $\hat{\mathbb{P}}_n$ is the empirical distribution of n i.i.d. random variables with distribution $\mathbb{P}$, it follows from general results to be found in Pollard (1984) that

$$d_{\mathfrak{T}\mathfrak{K}}\left(\hat{\mathbb{P}}_n, \mathbb{P}\right) = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.4}$$

One consequence of this is the following. If a function T is Fréchet differentiable at $\mathbb{P}$ with respect to $d_{\mathfrak{T}\mathfrak{K}}$ and the metric $d_{\mathfrak{T}\mathfrak{K}}$ satisfies (1.4), then, following the arguments of Huber (1981), pages 38 and 39, we immediately obtain a central limit theorem for $T(\hat{\mathbb{P}}_n)$. We note that this does not hold for the Prohorov metric [Kersting (1978), and Huber (1981)].

In subsection 1.4 we argued that metrics should be linearly invariant to reflect the linear invariance of the linear model. Metrics defined by (1.3) are linearly invariant if $\mathfrak{K} \cup \mathfrak{K}^c$ is linearly invariant, where $\mathfrak{K}^c$ denotes the class of complements of sets in $\mathfrak{K}$.

Not all useful metrics are of the form given by (1.3). We shall introduce a metric below which is based on a Vapnik–Cervonenkis class $\mathfrak{K}$ and which is linearly invariant but is not of this form. It can be regarded as a variant of the Prohorov metric with the absolute error being replaced by a proportional error. A similar metric was introduced in Davies (1992a) to study the behaviour of Rousseeuw's minimum volume ellipsoid.

We take the point of view that a metric should also reflect the structure of the problem. In Davies (1992a) one metric was based on ellipsoids and used to study Rousseeuw's minimum volume ellipsoid. We shall introduce a metric d in Section 3 in order to study the Krasker–Welsch regression functional. This has the advantage of indicating which properties of distribution have an influence on the metric. Thus, if a breakdown point with respect to some metric d is obtained, it will indicate what sort of deviation from the model will cause the estimate to fail. Similarly, if Fréchet differentiability or Lipschitz continuity is proved, this will show in what sense distributions have to be close together to give rise to functional values which are also close together.

1.6. *S-estimators.* The first high breakdown linearly equivariant functional for the regression problem was the least median of square proposed by Hampel (1975) and further developed by Rousseeuw (1984). Since then, several other such estimators have been proposed but all suffer from the defect of requiring regularity conditions for their uniqueness and continuity. The author has not yet succeeded in obtaining a global smooth high breakdown regression functional and therefore has to rely on existing estimators. S-estimators were introduced by Rousseeuw and Yohai (1984) in the context of linear regression and have since been investigated in more detail [Kim and Pollard (1990) and Davies (1990).] S-estimators can also be used for elliptical multivariate distributions [Rousseeuw (1986), Davies (1987) and Lopuhaä (1989)]. In Section 4 we study problems of existence, uniqueness and Hölder continuity. In Section 8 we show how k-step M-estimators can be used to obtain a high breakdown regression functional which is Lipschitz continuous at models with normal errors. Such an approach was taken in Davies (1992b) to construct a multivariate location and dispersion functional with the same properties. In spite of this last result, the author is convinced that a globally defined functional will have to be obtained by other means. S-estimators suffer from a serious flaw which probably cannot be removed, namely that uniqueness and continuity can only be proved under certain conditions. Hettmansperger and Sheather (1992) have shown that the least median of squares functional is not stable, a fact which can be deduced from its lack of continuity. The lack of continuity will persist even if an S-estimator with a smooth rho-function is used. The situation is similar to that for dispersion functionals. Section 6 contains an example where no S-estimator of dispersion can possibly be uniquely defined. This will cause all S-estimators to be unstable in the neighbourhood of such a distribution. Because of this we take a different approach to construct a dispersion functional for the design distribution.

1.7. *Notation.* A generic point of a linear regression will be denoted by $(x^T, y)^T$ with $y \in \mathbb{R}$ and the design point $x \in \mathbb{R}^k$. A regression problem is defined in terms of a probability measure $\mathbb{Q}$ on the Borel sets $\mathfrak{B}(\mathbb{R}^{k+1})$ of $\mathbb{R}^{k+1}$, giving the joint distribution of y and x. The empirical regression problem based on n data points $(x_i^T, y_i)^T$, $1 \le i \le n$, is described by the measure

$$\hat{\mathbb{Q}}_n = \frac{1}{n} \sum_1^n \delta_{(x_i^T, y_i)^T},$$

where δ_x here and in the following denotes the Dirac measure at the point x. We shall refer to probability measures as regression distributions, although it is to be emphasized that no linear relationship between y and the x's is intended, nor that the errors are independent of the x's. Given a measure $\mathbb{Q}$

on $\mathfrak{B}(\mathbb{R}^{k+1})$, we shall require the marginal distribution $\mathbb{Q}_d$ of the design points, $\mathbb{Q}_d(B) = \mathbb{Q}(B \times \mathbb{R})$, $B \in \mathfrak{B}(\mathbb{R}^k)$. Such a measure will be called a design measure. We shall restrict attention to nonsingular designs for which

$$\Delta(\mathbb{Q}) = \Delta(\mathbb{Q}_d) := \sup_{\|\theta\|=1} \mathbb{Q}_d(x^T\theta = 0) < 1. \tag{1.5}$$

It is clear that this condition is necessary in order to be able to obtain a uniquely defined regression functional. The set of regression distributions which satisfies (1.5) will be denoted by $\mathfrak{W}(\mathbb{R}^{k+1})$ and the set of design distributions by $\mathfrak{W}_d(\mathbb{R}^k)$.

The functional Δ: $\mathfrak{W}(\mathbb{R}^{k+1}) \to [0,1]$ plays an important role in this paper. It will turn out that the maximal attainable breakdown point of any linearly equivariant regression or dispersion functional T at the regression distribution $\mathbb{Q}$ is essentially determined by $\Delta(\mathbb{Q})$. It is immediately clear that the breakdown point is at most $1 - \Delta(\mathbb{Q})$ as by moving $1 - \Delta(\mathbb{Q})$ of mass to a plane of measure $\Delta(\mathbb{Q})$ we can obtain a $\mathbb{Q}'$ with $\Delta(\mathbb{Q}') = 1$. For this $\mathbb{Q}'$ all the design points lie on a plane of dimension at most $k - 1$. This implies that the set of solutions of the regression problem lies on a plane of dimension at least 1 and so can be arbitrarily large. If we consider a data sample of size n in $k + 1$ dimensions and with linearly independent design points, then $\Delta(\mathbb{Q}) \geq k/n$. This is because any k design points are linearly dependent and hence lie on a plane of dimension at least 1. In regression problems it is not unusual for several measurements to be made at some or all design points. If m measurements are made at some design point, then we can find $k - 1$ further design points so that the resulting $m + k - 1$ design points lie on a plane. This gives $\Delta(\mathbb{Q}) \geq (m + k - 1)/n$. In the metrics we consider, Δ is Lipschitz continuous but not differentiable and it is partly this which forces a choice between global definability and high breakdown point on the one hand and differentiability on the other.

The usual linear regression models specify

$$y = x^T\beta + e, \tag{1.6}$$

where e is the error whose distribution $\mathbb{Q}_e$ is independent of x and belongs to some scale family of distributions on $\mathbb{R}$. Such a model corresponds to a regression measure $\mathbb{Q}$ of the form

$$\mathbb{Q}(B_1 \times B_2) = \int_{B_2} \mathbb{Q}_e\big(B_1 - x^T\beta\big)\, d\mathbb{Q}_d(x) \tag{1.7}$$

and will be written $\mathbb{Q} = \mathbb{Q}_d * \mathbb{Q}_e$.

As we do not always obtain a functional which is uniquely defined at all points of $\mathfrak{W}(\mathbb{R}^{k+1})$, we consider set valued functionals

$$T: \mathfrak{W}(\mathbb{R}^{k+1}) \to \mathfrak{P}(\mathbb{R}^{k+1}),$$

where $\mathfrak{P}(\mathbb{R}^{k+1})$ denotes the set of subsets of $\mathbb{R}^{k+1}$. If T is not defined for some regression distribution $\mathbb{Q}$, we set $T(\mathbb{Q}) = \varnothing$.

The unit sphere in $\mathbb{R}^k$ will be denoted by $S^k = \{\theta: \theta \in \mathbb{R}^k, \|\theta\| = 1\}$ and Lebesgue measure in $\mathbb{R}^k$ by m_k. The set of all strictly positive definite symmetric $k \times k$-matrices will be denoted by PDS(k). For any positive definite matrix Σ, the smallest and largest eigenvalues will be denoted by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$.

2. The influence function.

2.1. *Generalities.* Let T be a functional well defined at a distribution $\mathbb{P}$ and at all distributions of the form $(1-\varepsilon)\mathbb{P} + \varepsilon\delta_x$. The influence function IF($\cdot$: $T, \mathbb{P}$) is defined by Hampel, Rousseeuw, Ronchetti and Stahel (1986), page 84:

$$\text{IF}(x: T, \mathbb{P}) = \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)\mathbb{P} + \varepsilon\delta_x\big) - T(\mathbb{P})}{\varepsilon} \tag{2.1}$$

if the limit exists.

T is said to have bounded influence if the gross error sensitivity $\gamma^*(T, \mathbb{P})$ defined by

$$\gamma^*(T, \mathbb{P}) = \sup_x |\text{IF}(x: T, \mathbb{P})| \tag{2.2}$$

is finite.

The heuristics of the influence function are heuristics and not theorems. Nevertheless, the failure of the heuristics for the Hampel–Rousseeuw least median of squares estimator has led to statements about the nonexistence of the influence function of this estimator [see, e.g., Rousseeuw and Leroy (1987) page 188].

The first heuristic conclusion for a bounded influence functional is that the effect of a small amount ε of point contamination is bounded by ε. More exactly, we expect for ε sufficiently small

$$\sup_x \big|T\big((1-\varepsilon)\mathbb{P} + \varepsilon\delta_x\big) - T(\mathbb{P})\big| \le 2\varepsilon\gamma^*(T, \mathbb{P}). \tag{2.3}$$

The constant 2 can be replaced by any constant greater than 1 at the possible cost of having to make the ε-values smaller.

If T is well defined at all empirical distributions $\hat{\mathbb{P}}_n$ deriving from n independently and identically distributed random variables $(X_j)_1^n$ with common distribution $\mathbb{P}$, then we expect [Hampel, Rousseeuw, Ronchetti and Stahel (1986), page 85]

$$n^{1/2}\big(T(\hat{\mathbb{P}}_n) - T(\mathbb{P})\big) = n^{-1/2}\sum_1^n \text{IF}(X_j: T, \mathbb{P}) + o_p(1). \tag{2.4}$$

In particular, we obtain a central limit theorem for $T(\hat{\mathbb{P}}_n)$ if T has bounded influence or, more generally, $\int \text{IF}(x: T, \mathbb{P})^2\, d\mathbb{P} < \infty$.

Finally, if we consider one-dimensional data, the stylized sensitivity curve SC_n may be defined by

$$(2.5)\qquad \mathrm{SC}_n(x) = n\left(T\left((1-n^{-1})\mathbb{P}^*_{n-1} + \frac{1}{n}\delta_x\right) - T(\mathbb{P}^*_{n-1})\right),$$

where

$$\mathbb{P}^*_{n-1} = \frac{1}{n-1}\sum_1^{n-1}\delta_{x_i} \quad \text{with } x_i = F_{\mathbb{P}}^{-1}\left(\frac{i}{n}\right), \qquad 1 \le i \le n-1,$$

and $F_{\mathbb{P}}$ denotes the distribution function associated with $\mathbb{P}$. In general, one expects

$$(2.6)\qquad \lim_{n\to\infty} \mathrm{SC}_n(x) = \mathrm{IF}(x\colon T, \mathbb{P}).$$

In going from (2.1) and (2.2) to (2.3), there has been an interchange of limits which is not always valid. The same holds for (2.1) and (2.5). In general, there is no relationship between (2.1) and (2.4) at least not without further conditions [Bednarski, Clarke and Kolkiewicz (1991)]. We show below that (2.1) and (2.2) hold for the middle of the shortest half estimator but that none of (2.3), (2.4) or (2.5) holds.

2.2. *Middle of the shortest half.* Let $\mathbb{P}$ be a distribution on $\mathfrak{B}(\mathbb{R})$. The middle of the shortest half functional T_{MSH} is defined as $\frac{1}{2}(a+b)$, where a and b are such as to minimize $b-a$ subject to $\mathbb{P}([a,b]) \ge \frac{1}{2}$. For the $\mathfrak{N}(0,1)$ distribution, we have $a = \Phi^{-1}(1/4)$, $b = \Phi^{-1}(3/4)$, giving $T_{\mathrm{MSH}}(\mathfrak{N}(0,1)) = 0$. Consider now x with $|x| > \Phi^{-1}(3/4)$. Then it is not difficult to show that for sufficiently small ε, depending on x, the shortest half interval for the distribution $(1-\varepsilon)\mathfrak{N}(0,1) + \varepsilon\delta_x$ is

$$\left[-\Phi^{-1}\left(\frac{3-2\varepsilon}{4(1-\varepsilon)}\right), \Phi^{-1}\left(\frac{3-2\varepsilon}{4(1-\varepsilon)}\right)\right],$$

giving $T_{\mathrm{MSH}}((1-\varepsilon)\mathfrak{N}(0,1) + \varepsilon\delta_x) = 0$. Similarly, if $|x| < \Phi^{-1}(0.75)$, then for sufficiently small ε the shortest half interval is

$$\left[-\Phi^{-1}\left(\frac{3-4\varepsilon}{4(1-\varepsilon)}\right), \Phi^{-1}\left(\frac{3-4\varepsilon}{4(1-\varepsilon)}\right)\right]$$

and again we deduce $T_{\mathrm{MSH}}((1-\varepsilon)\mathfrak{N}(0,1) + \varepsilon\delta_x) = 0$. Finally, if $x = \Phi^{-1}(3/4)$, then the smallest half interval is now

$$\left[\Phi^{-1}\left(\frac{1+\varepsilon}{4(1-\varepsilon)}\right), \Phi^{-1}\left(\frac{3}{4}\right)\right]$$

and we may deduce

$$\text{IF}(x\colon T_{\text{MSH}}, \mathfrak{N}(0,1)) \tag{2.7}$$
$$= \begin{cases} -(\pi/8)^{1/2}\exp\big(\Phi^{-1}(3/4)^2/2\big), & x = -\Phi^{-1}(3/4),\\ (\pi/8)^{1/2}\exp\big(\Phi^{-1}(3/4)^2/2\big), & x = \Phi^{-1}(3/4),\\ 0, & \text{otherwise}. \end{cases}$$

It follows that the middle of the shortest half sample estimator has a bounded influence function. Nevertheless, none of (2.3), (2.4) or (2.6) holds. To start with (2.6), it can be shown that

$$\lim_{n\to\infty} \text{SC}_n(x) = 4\,\text{IF}(x\colon T_{\text{MSH}}, \mathfrak{N}(0,1)). \tag{2.8}$$

This conflicts with a statement in Rousseeuw and Leroy (1987), page 189, where it is claimed that $\lim_{n\to\infty}\text{SC}_n(x) = \infty$ for $x = \pm\Phi^{-1}(3/4)$, a result suggested by Figure 4 on page 190 of Rousseeuw and Leroy (1987). However, (2.8) can be made plausible by computer calculation or, for the doubting, it can be proved analytically. Figure 4 on page 190 of Rousseeuw and Leroy (1987) shows very dramatically that the pointwise convergence in (2.8) is not uniform. The factor 4 in (2.8) follows from the definition of the x_i as $F_{\mathbb{P}}^{-1}(i/n)$ in the definition of the stylized sensitivity curve. Another factor would be obtained if the x_i were to be defined for example as $F_{\mathbb{P}}^{-1}(i/(n-1/2))$, $1 \le i \le n-1$.

To show that (2.4) does not hold, we refer to the cube root convergence of T_{MSH} proved in Kim and Pollard (1990) and Davies (1990). Finally, to show that (2.3) does not hold, it is sufficient to set $x = x(\varepsilon) = \Phi^{-1}(3/4) + \alpha(\varepsilon)$, where $\alpha(\varepsilon)$ satisfies

$$\Phi\big(\Phi^{-1}(3/4) + \alpha(\varepsilon)\big) - \Phi\big(\Phi^{-1}(1/4) + \alpha(\varepsilon)\big) = \frac{1-2\varepsilon}{2(1-\varepsilon)}.$$

The shortest half interval is seen to be $[\Phi^{-1}(1/4) + \alpha(\varepsilon), \Phi^{-1}(3/4) + \alpha(\varepsilon)]$, giving $T_{\text{MSH}}((1-\varepsilon)\mathfrak{N}(0,1) + \varepsilon\delta_{x(\varepsilon)}) = \alpha(\varepsilon)$. From a Taylor expansion we deduce $\alpha(\varepsilon) \simeq ((2\pi)^{1/2}\exp(\Phi^{-1}(3/4)^2/2)/(2\Phi^{-1}(3/4)))\varepsilon^{1/2}$, from which we conclude

$$\sup_x \big|T_{\text{MSH}}\big((1-\varepsilon)\mathfrak{N}(0,1) + \varepsilon\delta_{x(\varepsilon)}\big) - T_{\text{MSH}}(\mathfrak{N}(0,1))\big| \ge c_1\varepsilon^{1/2}, \tag{2.9}$$

for some constant $c_1 > 0$, in contrast to (2.3). If we specialize Theorem 6 of Davies (1992a) to the case $p = 1$ and the Kolmogorov metric on $\mathbb{R}$, we obtain

$$|T_{\text{MSH}}(\mathbb{Q}) - T_{\text{MSH}}(\mathfrak{N}(0,1))| \le c_2 d_K(\mathbb{Q}, \mathfrak{N}(0,1))^{1/2}.$$

This in conjunction with (2.9) shows that T_{MSH} satisfies a local Hölder condition of order $1/2$ at the normal distribution and that this cannot be improved upon. One of the results of this paper is that the Hampel–Rousseeuw least median of squares estimator also satisfies a local Hölder condition of order $1/2$ in an appropriate metric. This result is perhaps somewhat surprising as the estimator does not depend on an explicit down-weighting of leverage points.

2.3. *Least median of squares.* Let $\mathbb{Q}$ be a regression distribution. The Hampel–Rousseeuw least median of squares functional T_{LMS} is defined to be $(\tilde{b}^T, \tilde{s})^T \in \mathbb{R}^{k+1}$, where $(\tilde{b}^T, \tilde{s})^T$ is the solution of the following problem. Choose $(b^T, s)^T \in \mathbb{R}^{k+1}$ so as to minimize s subject to $\mathbb{Q}(\{(x^T, y)^T\colon |y - x^T b| \le s\}) \ge 1/2$.

We now specialize to the case of a finite number of design points $x_1, \dots, x_n$ and where the y's are independently and identically distributed with common distribution $\mathfrak{N}(0,1)$. Furthermore, we assume that the x_i's are not concentrated on a plane of dimension less than k. For this $\mathbb{Q}$ it follows from Theorem 4.7 [see also Davies (1990), Theorem 3] that $T_{\mathrm{LMS}}(\mathbb{Q}) = (0^T, \Phi^{-1}(3/4))^T$. Consider now a point contamination of mass ε of the point $(x_0^T, y_0)^T$. Theorem 4.1 shows that for each ε, $0 \le \varepsilon < 1$, there exists a solution $(b(\varepsilon)^T, s(\varepsilon))^T$ of the minimization problem. From Theorem 4.7 we obtain

$$\lim_{\varepsilon \downarrow 0} \left(b(\varepsilon)^T, s(\varepsilon)\right)^T = \left(0^T, \Phi^{-1}(3/4)\right)^T. \tag{2.10}$$

Suppose now that $|y_0| > \Phi^{-1}(3/4)$. We see from (2.10) that $|y_0 - x_0^T b(\varepsilon)| > s(\varepsilon)$ for all ε sufficiently small and hence

$$(1-\varepsilon)\frac{1}{n}\sum_1^n \left(\Phi\left(x_i^T b(\varepsilon) + s(\varepsilon)\right) - \Phi\left(x_i^T b(\varepsilon) - s(\varepsilon)\right)\right) = \frac{1}{2}.$$

As each individual summand is maximized by setting $b(\varepsilon) = 0$, we see that for ε sufficiently small $T_{\mathrm{LMS}}((1-\varepsilon)\mathbb{Q} + \varepsilon\delta_{(x_0^T, y)^T}) = (0, s(\varepsilon))^T$. Similarly, if $|y_0| < \Phi^{-1}(3/4)$, $(b(\varepsilon)^T, s(\varepsilon))^T$ minimizes $s(\varepsilon)$ subject to

$$(1-\varepsilon)\frac{1}{n}\sum_1^n \left(\Phi\left(x_i^T b(\varepsilon) + s(\varepsilon)\right) - \Phi\left(x_i^T b(\varepsilon) - s(\varepsilon)\right)\right) + \varepsilon = \frac{1}{2}.$$

Again $s(\varepsilon)$ is minimized by setting $b(\varepsilon) = 0$. If $y_0 \ne \pm\Phi^{-1}(3/4)$ we obtain, using a standard Taylor expansion,

$$\lim_{\varepsilon \downarrow 0} \frac{T_{\mathrm{LMS}}\left((1-\varepsilon)\mathbb{Q} + \varepsilon\delta_{(x_0^T, y_0)^T}\right) - T_{\mathrm{LMS}}(\mathbb{Q})}{\varepsilon} = \left(0^T, \pm\frac{1}{4\varphi(\Phi^{-1}(3/4))}\right)^T,$$

with the $+$ for $|y_0| > \Phi^{-1}(3/4)$ and the $-$ for $|y_0| < \Phi^{-1}(3/4)$.

It remains to consider the case $y_0 = \pm\Phi^{-1}(3/4)$. Suppose $y = \Phi^{-1}(3/4)$. Then it is easily seen that the solution $(b(\varepsilon)^T, s(\varepsilon))^T$ satisfies

$$(1-\varepsilon)\frac{1}{n}\sum_1^n \Phi\left(x_i^T b(\varepsilon) + s(\varepsilon)\right) - \Phi\left(x_i^T b(\varepsilon) - s(\varepsilon)\right) + \varepsilon = \frac{1}{2} \tag{2.11}$$

and

$$y_0 - x_0^T b(\varepsilon) = s(\varepsilon). \tag{2.12}$$

Given $s(\varepsilon)$, it is clear that $b(\varepsilon)$ maximizes (2.11) subject to (2.12). A Taylor expansion of (2.11) shows that $b(\varepsilon)$ maximizes

$$\sum_1^n \left(x_i^T b(\varepsilon)\right)^2 + o\left(\|b(\varepsilon)\|^2\right)$$

subject to (2.12). The solution is

$$b(\varepsilon) = \frac{\left(y_0 - s(\varepsilon)\right) \cdot V^{-1}x_0(1 + o(1))}{x_0^T V^{-1} x_0},$$

with $V = \sum_1^n x_i x_i^T$. A short calculation shows

$$s(\varepsilon) = \Phi^{-1}\left(\frac{3}{4}\right) - \frac{\varepsilon}{4\varphi(\Phi^{-1}(3/4))} + o(\varepsilon^2)$$

and this gives

$$\text{(2.13)} \quad \mathrm{IF}\left(\left(x_0^T, y_0\right)^T : T_{\mathrm{LMS}}, \mathbb{Q}\right) = \left(\frac{1}{4\varphi(\Phi^{-1}(3/4))}\right)\left(\pm\frac{\left(V^{-1}x_0\right)^T}{x_0^T V^{-1} x_0}, -1\right)^T,$$

for $y_0 = \pm\Phi^{-1}(3/4)$. Thus we have shown that T_{LMS} does indeed have an influence function. As V is nonsingular the influence function is bounded for x_0 with $\|x_0\| \geq \delta > 0$. It is not, however, bounded for small x_0. Hettmansperger and Sheather (1992) have noted that T_{LMS} is sensitive towards inliers. Equation (2.13) may help to explain this result.

2.4. *Bounded influence regression.* In this section we consider regression distributions with known error dispersion so that the regression functional T takes values in $\mathbb{R}^k$. A generalized M-estimator T_η is defined by the implicit equation

$$\text{(2.14)} \quad \int \eta\left(x, y - T_\eta(\mathbb{Q})^T x\right) x \, d\mathbb{Q} = 0$$

[see Hampel, Rousseeuw, Ronchetti and Stahel (1986), page 315]. Under certain regularity conditions $T_\eta(\mathbb{Q})$ exists, is uniquely defined and has an influence function given by

$$\text{(2.15)} \quad \mathrm{IF}\left(\left(x^T, y\right)^T : T_\eta, (\mathbb{Q})\right) = \eta\left(x, y - T_\eta(\mathbb{Q})^T x\right) M^{-1}(\eta, \mathbb{Q}) x,$$

where

$$M(\eta, \mathbb{Q}) = \int \eta'\left(x, y - T_\eta(\mathbb{Q})^T x\right) x x^T \, d\mathbb{Q}.$$

In the literature there seems to be a general assumption that standard bounded regression functionals are generalized M-estimators [Hampel, Rousseeuw, Ronchetti and Stahel (1986), Simpson, Ruppert and Carroll (1992), and Maronna and Yohai (1991)]. As an example we mention the Krasker–Welsch [Krasker and Welsch (1982)] estimator which is given by

(2.14) with

$$\eta(x, r) = \frac{1}{\|Ax\|}\psi_c(r\|Ax\|), \tag{2.16}$$

where ψ_c is Huber's ψ-function and A is determined implicitly by

$$\mathbb{E}\big(\|Ax\|^{-2}\psi_c^2(r\|Ax\|)xx^T\big) = (A^TA)^{-1}. \tag{2.17}$$

The expectation in (2.16) is evaluated under the assumption of independently distributed $\mathfrak{N}(0,1)$ errors. It is clear that A as defined is a dispersion functional for the design distribution of the regression $\mathbb{Q}$. It will depend on $\mathbb{Q}$ and it is this dependency which means that the Krasker–Welsch estimator is not a generalized M-estimator as defined by (2.14). Given this, it is no longer obvious that the influence function of the Krasker–Welsch functional is given by (2.15). Huber (1983) claims that the bounded influence regression of Krasker and Welsch (1982) is not concerned with errors in the independent variables. In their comment on Huber (1983), Krasker and Welsch disputed this. The argument given above supports Huber's position. We examine this in more detail.

Consider a point contamination of size ε at the point $(x^T, y)^T$. The implicit definition of A given by (2.17) will give rise to a new dispersion matrix $A(\varepsilon)$. If we now substitute this into (2.14) and go through the calculations, we obtain the following. The influence function of the Krasker–Welsch estimator is indeed given by (2.15) as long as the following two conditions hold:

$$\lim_{\varepsilon \downarrow 0} A(\varepsilon) = A \tag{2.18}$$

and

$$\int \eta\big(x, (y - x^T\beta)\lambda\big)\, d\mathbb{Q}(x, y) = 0, \tag{2.19}$$

for all λ. In particular, (2.19) will hold if the errors have a symmetrical distribution and $\eta(x, r)$ is an odd function of r. The latter is indeed the case for η given by (2.15).

Suppose now that (2.19) does not hold and that the dispersion part has an influence function $\mathrm{IF}((x^T, y)^T\!: A, \mathbb{Q}) = \Gamma$ so that $A(\varepsilon) = A + \varepsilon\Gamma + o(\varepsilon)$. If we now recalculate the influence function of the Krasker–Welsch estimator, we obtain

$$\begin{aligned}\mathrm{IF}\big((x^T, y)^T : T_{\mathrm{KW}}, \mathbb{Q}\big) &= \psi_c\Big(\big(y - T_{\mathrm{KW}}(\mathbb{Q})^T x\big)\|Ax\|\Big)x/\|Ax\| \\ &\quad - \int \frac{z^TA^T\Gamma z}{\|Az\|^3}\psi_c\Big(\big(v - T_{\mathrm{KW}}(\mathbb{Q})^T z\big)\|Az\|\Big)z\, d\mathbb{Q}(z, v) \\ &\quad + \int\left(\frac{z^TA^T\Gamma z}{\|Az\|^2}\right)\big(v - T_{\mathrm{KW}}(\mathbb{Q})^T z\big)\psi_c^{(1)} \\ &\quad \times\Big(\big(v - T_{\mathrm{KW}}(\mathbb{Q})^T z\big)\|Az\|\Big)^Z d\mathbb{Q}(z, v).\end{aligned}$$

This again reduces to (2.15) if the errors are symmetrically distributed under $\mathbb{Q}$.

Maronna and Yohai (1991) calculated the breakdown point (gross error model) of the generalized M-estimator with unknown scale. They state that for a symmetrical design distribution the Krasker–Welsch functional is optimal with respect to breakdown. However, they assumed that the dispersion matrix A is known so that the possibility of a breakdown of the dispersion functional as defined by (2.17) leading to a breakdown of the regression functional does not arise.

Simpson, Ruppert and Carroll (1992) suggest using a high breakdown dispersion functional in place of the matrix A of (2.17). In particular, they suggest using Rousseeuw's minimum volume estimator [Rousseeuw (1986), Nolan (1991) and Davies (1992a)] or multivariate S-estimators of dispersion [Davies (1987)]. The idea of using a high breakdown functional is surely correct; the problem is that those suggested are strongly linked to a parametric model, one of an elliptical distribution. The existence of solutions can be proved quite generally, but all uniqueness theorems require that the distribution be elliptical [Davies (1987)]. Lopuhaä (1989) has shown that smooth S-estimators of dispersion have a well-defined influence function, but again this has only been shown for elliptical distributions. One of the main problems with linear regression is that the design distribution is given; I know of no practical case where this is elliptical. We therefore have to deal with an essentially nonparametric situation. The results of Simpson, Ruppert and Carroll (1992) are therefore of limited value unless it is possible to obtain a well-defined smooth high breakdown dispersion functional. It is exactly this problem which is considered in Section 6.

3. Linear equivariance, metrics and breakdown points.

3.1. *Linear equivariance for regression functionals.* The regression problem remains invariant under the group $\mathfrak{A}$ of linear transformations $A\colon \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}$ of the form $A((x^T, y)^T) = ((\Gamma x)^T, (\alpha y - x^T\gamma))^T$, where $\alpha \in \mathbb{R} \setminus \{0\}$, $\gamma \in \mathbb{R}^k$ and $\Gamma\colon \mathbb{R}^k \to \mathbb{R}^k$ is a nonsingular linear transformation. Let $\mathbb{Q}^A$ denote the regression distribution of the transformed problem so that $\mathbb{Q}^A(B) = \mathbb{Q}(A^{-1}(B))$, $B \in \mathfrak{B}(\mathbb{R}^{k+1})$. We shall call an estimator $T\colon \mathfrak{W}(\mathbb{R}^{k+1}) \to \mathfrak{P}(\mathbb{R}^{k+1})$ linearly equivariant if, for all $\mathbb{Q} \in \mathfrak{W}(\mathbb{R}^{1+k})$ and for all $A \in \mathfrak{A}$,

$$T(\mathbb{Q}^A) = A^{\#}(T(\mathbb{Q})), \tag{3.1}$$

where $A^{\#}((b^T, s)^T) = ((\alpha b - \gamma)^T \Gamma^{-1}, |\alpha| s)^T$. It should be noted that this is a rather weak sense of equivariance as, in the case of more than one solution, there is no guarantee that the solution obtained by one method will transform in an equivariant manner. If, however, $T(\mathbb{Q})$ is a one-point set, then so is the $T(\mathbb{Q}^A)$ for any A in $\mathfrak{A}$, and these estimators transform in an equivariant manner.

3.2. *Linearly equivariant metrics for linear regression.* Given that the regression problem is invariant under $\mathfrak{A}$, it seems natural to make use of metrics on the space of regression distributions which are metric invariant with respect to $\mathfrak{A}$. The most obvious invariant is the total variation metric $d_{\mathfrak{T}\mathfrak{V}}$ of (1.2) but, as mentioned in subsection 1.5, this is too strong. To define a weaker metric, we consider the family $\mathfrak{H}$ of strips in $\mathbb{R}^{k+1}$ defined by $\mathfrak{H} = \{H(c,\gamma): c \ge 0, \gamma \in \mathbb{R}^{k+1}\}$, where $H(c,\gamma) = \{z: |z^T\gamma| \le c\}$. For any $H = H(c,\gamma)$ and any $\eta > 0$, we write $H^\eta = \{z: |z^T\gamma| \le ce^\eta\}$. The invariant pseudometric $d_{\mathfrak{H}}$ on $\mathfrak{W}(\mathbb{R}^{k+1})$ is defined by

$$\begin{aligned} d_{\mathfrak{H}}(\mathbb{Q}_1,\mathbb{Q}_2) = \inf\{\eta > 0: \mathbb{Q}_1(H) \le \mathbb{Q}_2(H^\eta) + \eta \text{ and} \\ \mathbb{Q}_2(H) \le \mathbb{Q}_1(H^\eta) + \eta \text{ for all } H \in \mathfrak{H}\}, \end{aligned} \tag{3.2}$$

for all regression measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ in $\mathfrak{W}(\mathbb{R}^{k+1})$. It is easily checked that $d_{\mathfrak{H}}(\mathbb{Q}_1^A, \mathbb{Q}_2^A) = d_{\mathfrak{H}}(\mathbb{Q}_1, \mathbb{Q}_2)$ for all $\mathbb{Q}_1, \mathbb{Q}_2$ in $\mathfrak{W}(\mathbb{R}^{k+1})$ and for all $A \in \mathfrak{A}$.

It is perhaps worth making a few comments on the pseudometric $d_{\mathfrak{H}}$. If H^η is replaced by $\{z: |z^T\gamma| \le c + \eta\}$, then we obtain a form of Prohorov pseudometric and η has the interpretation of an inaccuracy in the measurements of z. The definition of H^η may be interpreted as a proportional error corresponding to the absolute error of the Prohorov metric.

Although the pseudometric $d_{\mathfrak{H}}$ is sufficient for much of the analysis, we shall require a more complicated metric when dealing with the Krasker–Welsch bounded influence functional. The reasons for this were discussed in subsection 1.5. We define the metric $d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}_1, \mathbb{Q}_2)$ by

$$d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}_1,\mathbb{Q}_2) = \sup_{c \in \mathfrak{C}} |\mathbb{Q}_1(C) - \mathbb{Q}_2(C)|, \tag{3.3}$$

where $\mathfrak{C}$ is the family of sets of the form $C_1 \cup C_2$ with

$$C_i = \left\{(x^T, y)^T: \theta_i^T x > a_i, (\alpha_i y - \gamma_i^T x)\|\Gamma x\| < c_i\right\}, \tag{3.4}$$

with $a_i, \alpha_i, c_i \in \mathbb{R}, \theta_i, \gamma_i \in \mathbb{R}^k$ and Γ a nonsingular $k \times k$-matrix. The class $\mathfrak{C}$ is invariant under the group $\mathfrak{A}$ and hence $d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}_1^A, \mathbb{Q}_2^A) = d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}_1, \mathbb{Q}_2)$ for all $A \in \mathfrak{A}$. The metric $d_{\mathfrak{K}\mathfrak{W}}$ is weak as the class $\mathfrak{C}$ is a Vapnik–Cervonenkis class. This follows from Lemma 18 on page 20 of Pollard (1984).

The gross error model $\mathbb{Q}' = (1-\varepsilon)\mathbb{Q} + \varepsilon\mathbb{Q}''$ is popular in robust statistics and the ε may be related to $d_{\mathfrak{H}}$ and $d_{\mathfrak{K}\mathfrak{W}}$ by noting

$$\sup_{Q''} d(\mathbb{Q},\mathbb{Q}') = \varepsilon, \tag{3.5}$$

for $d = d_{\mathfrak{H}}$ or $d_{\mathfrak{K}\mathfrak{W}}$.

3.3. *Breakdown points of regression functionals.* We now define the breakdown point $\varepsilon^*(T,\mathbb{Q},d)$ of a regression estimator T at the point $\mathbb{Q}$ of $\mathfrak{W}(\mathbb{R}^{k+1})$ and with respect to the metric d as follows:

$$\varepsilon^*(T,\mathbb{Q},d) = \inf\{\varepsilon > 0: \sup(\|b(\mathbb{Q}')\| + s(\mathbb{Q}')) = \infty\}, \tag{3.6}$$

where the supremum is taken over all $\mathbb{Q}'$ with $d(\mathbb{Q}, \mathbb{Q}') < \varepsilon$ and all $(b^T(\mathbb{Q}'), s(\mathbb{Q}'))^T$ in $T(\mathbb{Q}')$ with the convention $\|b(\mathbb{Q}')\| = \infty$ if $T(\mathbb{Q}') = \varnothing$.

THEOREM 3.1. *Let T be any linearly equivariant functional and d any linearly invariant metric with $d \le d_{\mathfrak{T}\mathfrak{D}}$. Then for all $\mathbb{Q}$ and $A \in \mathfrak{A}$ we have* $\varepsilon^*(T, \mathbb{Q}, d) = \varepsilon^*(T, \mathbb{Q}^A, d) \le (1 - \Delta(\mathbb{Q}))/2$.

PROOF. Let $H = \{x: x^T\theta = 0\}$ with $\|\theta\| = 1$ be such that $\mathbb{Q}(H) = \delta(\mathbb{Q})$ and consider the linear transformation $A(\tau)$ given by $A(\tau)((x^T, y)^T) = (x^T, y - \tau x^T\theta)^T$. We define the measures W_1, W_2 and $\mathbb{Q}_\tau$ by $W_1(B) = \mathbb{Q}(B \cap (\mathbb{R} \times H))$, $W_2(B) = \mathbb{Q}(B) - W_1(B)$ and $\mathbb{Q}_\tau(B) = (W_2 + W_2^{A(\tau)})/2 + W_1$. As $W_1^{A(\tau)} = W_1^{A(-\tau)} = W_1$ and $(W_2^{A(\tau)})^{A(-\tau)} = W_2$, we obtain $\mathbb{Q}_\tau^{A(-\tau)} = (W_2^{A(-\tau)} + W_2)/2 + W_1$. As T is linearly equivariant we have $(b^T + \tau\theta^T, s)^T \in T(\mathbb{Q}_\tau^{A(-\tau)})$ for some $(b^T, s)^T \in T(\mathbb{Q}_\tau)$ and hence not both of the sets $T(\mathbb{Q}_\tau^{A(\tau)})$ and $T(\mathbb{Q}_\tau^{A(-\tau)})$ can be bounded. It remains to show that $d(\mathbb{Q}, \mathbb{Q}_\tau)$ and $d(\mathbb{Q}, \mathbb{Q}_\tau^{A(-\tau)})$ are both at most $(1 - \Delta(\mathbb{Q}))/2$. Because of the symmetry of the situation and the fact that d is linearly invariant, it is sufficient to consider $\mathbb{Q}_\tau$. The claim follows on noting $|\mathbb{Q}_\tau(B) - \mathbb{Q}(B)| \le |W_2(B) - W_2^{A(\tau)}(B)|/2 \le (1 - W_2(\mathbb{R}^{k+1}))/2 = (1 - \Delta(\mathbb{Q}))/2$. □

We point out that the finite sample breakdown point version of this result is (1.1) and may be found in Rousseeuw and Leroy (1987), Theorem 4, page 125. The "p" in their theorem is the dimension of the design space. A close examination of the proof shows, however, that it remains valid if p is the largest number design points on some lower-dimensional plane. In our notation we then have $\Delta(\mathbb{Q}) = p/n$. The slight difference between (1.1) and Theorem 3.1 is due to the fact that the finite sample breakdown point definition allows one only to move whole points of measure $1/n$. In the metric definition a point of mass $1/n$ may be split into two "points," each of measure $1/(2n)$. Similar results are to be found in Davies (1987) and Lopuhaä and Rousseeuw (1991).

3.4. *Linear equivariance for dispersion functionals.* In order to determine the leverage points of a design distribution, it would seem to be necessary to obtain a robust dispersion measure. We therefore consider dispersion functionals T^D, T^D: $\mathfrak{W}_d(\mathbb{R}^k) \to \mathfrak{P}(\mathrm{PDS}(k))$ with the convention that $T^D(\mathbb{Q}_d) = \varnothing$ if T^D is not defined at the point $\mathbb{Q}_d$. A dispersion functional T^D will be called linearly equivariant if, with the obvious convention, $T^D(\mathbb{Q}_d^G) = GT^D(\mathbb{Q}_d)G^T$ for all G in the group $\mathfrak{G}$ of nonsingular $k \times k$-matrices. Again, this is a weak definition of linear equivariance as, unless $T^D(\mathbb{Q}_d)$ consists of just one point, there is no guarantee that actual versions of $T^D(\mathbb{Q}_d)$ will transform in a linearly equivariant manner.

As for the regression problem it seems natural to use metrics on $\mathfrak{W}_d(\mathbb{R}^k)$ which reflect the linear invariance of the dispersion problem. A metric d^D on $\mathfrak{W}_d(\mathbb{R}^k)$ will be called linearly invariant if $d^D(\mathbb{Q}_{1d}^G, \mathbb{Q}_{2d}^G) = d^D(\mathbb{Q}_{1d}, \mathbb{Q}_{2d})$ for all nonsingular G. Corresponding to the pseudometric $d_{\mathfrak{H}}$ on $\mathfrak{W}(\mathbb{R}^{k+1})$, we define the pseudometric $d_{\mathfrak{H}}^D$ on the space of design distributions by (3.2) but on $\mathbb{R}^k$.

3.5. *Breakdown points of dispersion functionals.* We define the breakdown point $\varepsilon^*(T^D, \mathbb{Q}_d, d^D)$ of the dispersion functional T^D at the point $\mathbb{Q}_d$ of $\mathfrak{W}_d(\mathbb{R}^k)$ with respect to the metric d^D as follows:

$$\varepsilon^*\left(T^D, \mathbb{Q}_d, d^D\right) = \inf\left\{\varepsilon > 0\colon \sup\left(\lambda_{\max}(D) + \lambda_{\min}(D)^{-1}\right) = \infty\right\},$$

where the supremum is taken over all $\mathbb{Q}'_d$ with $d^D(\mathbb{Q}_d, \mathbb{Q}'_d) < \varepsilon$ and all D in $T^D(\mathbb{Q}'_d)$ with the convention $\lambda_{\max}(D) = \infty$ if $T^D(\mathbb{Q}'_d) = \varnothing$.

THEOREM 3.2. *Let T^D be any linearly equivariant dispersion functional and d^D any linearly invariant metric which satisfies $d^D \le d_{\mathfrak{T}\mathfrak{D}}$. Then for all $\mathbb{Q}_d$ and $G \in \mathfrak{G}$ we have*

$$\varepsilon^*\left(T^D, \mathbb{Q}_d, d^D\right) = \varepsilon^*\left(T^D, \mathbb{Q}_d^G, d^D\right) \le \left(1 - \Delta(\mathbb{Q}_d)\right)/2.$$

PROOF. Let $H = \{x\colon x^T\theta = 0\}$ with $\theta \in S^k$ be such that $\mathbb{Q}_d(H) = \Delta(\mathbb{Q}_d)$. Without loss of generality we may assume that $H = \{x\colon x_1 = 0\}$. We consider the linear transformation $G(\tau)$ given by $G(\tau)(x) = (\tau x_1, x_2, \dots, x_k)^T$ with $\tau \ne 0$. We define the measures W_{1d}, W_{2d} and $\mathbb{Q}_{\tau d}$ by $W_{1d}(B) = \mathbb{Q}_d(B \cap H)$, $W_{2d}(B) = \mathbb{Q}_d(B) - W_{1d}(B)$ and $\mathbb{Q}_{\tau d}(B) = (W_{d2} + W_{d2}^{G(\tau)})/2 + W_{1d}$.

As $W_{1d}^{G(\tau)} = W_{1d}^{G(\tau^{-1})} = W_{1d}$ and $(W_{2d}^{G(\tau)})^{G(\tau^{-1})} = W_{2d}$, we obtain

$$\mathbb{Q}_{\tau d}^{G(\tau^{-1})} = \left(W_{2d}^{G(\tau^{-1})} + W_{2d}\right)/2 + W_{1d}.$$

The linear equivariance of T^D implies $\det(\Gamma) = \tau^{-2}\det(\Gamma')$ for some $\Gamma \in T^D(D_{\tau d}^{G(\tau^{-1})})$ and for some $\Gamma' \in T^D(\mathbb{Q}_{\tau d})$. Thus, as $\tau \to 0$, it is not possible for the eigenvalues in the sets $T^D(\mathbb{Q}_{\tau d}^{G(\tau^{-1})})$ and $T^D(\mathbb{Q}_{\tau d})$ to be bounded away from 0 and ∞. That $d(\mathbb{Q}_{\tau d}, \mathbb{Q}_d) \le (1 - \Delta(\mathbb{Q}_d))/2$ follows as in the proof of Theorem 3.1. □

4. S-estimators.

4.1. *Existence.* S-estimators for the linear regression model were introduced by Rousseeuw and Yohai (1984). To define S-estimators, we consider a function ρ, $\rho\colon \mathbb{R} \to [0, 1]$, with the following properties [Rousseeuw and Yohai (1984) and Davies (1990)]:

R1. ρ is symmetric.
R2. $\rho\colon \mathbb{R}_+ \to [0, 1]$ is nonincreasing.
R3. $\rho(0) = 1$.
R4. ρ is continuous at 0.
R5. $\rho\colon \mathbb{R}_+ \to [0, 1]$ is continuous on the left.
R6. $\rho(u) > 0$ if $0 \le u < c_0$ and $\rho(u) = 0$ if $u \ge c_0$ for some $c_0 > 0$.

We note that R1, R2 and R5 imply

$$\limsup_{u' \to u} \rho(u') \le \rho(u). \tag{4.1}$$

Let ε, $0 < \varepsilon < 1$, be fixed and suppose the function ρ satisfies R1–R6. For any distribution $\mathbb{Q}$ we define an S-estimator $T_S(\mathbb{Q})$ to be the set of solutions $(b^T, s)^T$ of the following problem.

Choose $(b^T, s)^T$, $s \ge 0$, so as to minimize s subject to

$$\int_{\mathbb{R}^{k+1}} \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q}(y, x) \ge 1 - \varepsilon \tag{4.2}$$

[Rousseeuw and Yohai (1984) and Davies (1990)].

We denote this problem by $\mathfrak{P}(\mathbb{Q})$.

THEOREM 4.1. *$T_S(\mathbb{Q}) \ne \varnothing$ for any distribution $\mathbb{Q}$.*

PROOF. It follows from R1–R4 that $\lim_{s \uparrow \infty} \int \rho(y/s)\, d\mathbb{Q} = 1 > 1 - \varepsilon$ so that the set

$$\mathfrak{S}_0(\mathbb{Q}) = \left\{ (b^T, s)^T : \int_{\mathbb{R}^{k+1}} \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q}(y, x) \ge 1 - \varepsilon \right\}$$

is nonempty. We define $s^* = \inf\{s : (b^T, s)^T \in \mathfrak{S}_0(\mathbb{Q})\}$ and consider a sequence $((b_n^T, s_n)^T)_1^\infty$ in $\mathfrak{S}_0(\mathbb{Q})$ with $\lim s_n = s^*$. If there exists a convergent subsequence $((b_{n_i}^T, s_{n_i})^T)_1^\infty$ with $\lim b_{n_i} = b^*$, then it follows from (3.1) that

$$1 - \varepsilon \le \limsup_{i \to \infty} \int \rho\left(\frac{y - x^T b_{n_i}}{s_{n_i}}\right) d\mathbb{Q} \le \int \limsup_{i \to \infty} \rho\left(\frac{y - x^T b_{n_i}}{s_{n_i}}\right) d\mathbb{Q}$$
$$\le \int \rho\left(\frac{y - x^T b^*}{s^*}\right) d\mathbb{Q},$$

from which we see that $(b^{*T}, s^*)^T$ is a solution of $\mathfrak{P}(\mathbb{Q})$.

If $\|b_n\|$ tends to ∞ we consider first the case

$$\lim_{n \to \infty} |x^T b_n| = \infty, \tag{4.3}$$

for all $x \ne 0$. We obtain

$$\begin{aligned} 1 - \varepsilon &\le \limsup_{n \to \infty} \int \rho\left(\frac{y - x^T b_n}{s_n}\right) d\mathbb{Q} \\ &\le \lim_{n \to \infty} \int_{\{x = 0\}} \rho\left(\frac{y - x^T b_n}{s_n}\right) d\mathbb{Q} + \limsup_{n \to \infty} \int_{\{x \ne 0\}} \rho\left(\frac{y - x^T b_n}{s_n}\right) d\mathbb{Q} \\ &= \lim_{n \to \infty} \int_{\{x = 0\}} \rho\left(\frac{y}{s_n}\right) d\mathbb{Q} \\ &\le \int \rho\left(\frac{y}{s^*}\right) d\mathbb{Q}. \end{aligned} \tag{4.4}$$

This implies $(0^T, s^*)^T$ is a solution.

If (4.3) does not hold, then there exists an x_1 with $\liminf_{n\to\infty}|x_1^T b_n| < \infty$. We choose a subsequence of $(b_n)_1^\infty$ which we continue to denote by $(b_n)_1^\infty$ such that $\lim_{n\to\infty} x_1^T b_n = a_1$. Let L_1 denote the linear subspace spanned by x_1 and $L_1^\perp$ its orthogonal complement. If $\lim_{n\to\infty}|x^T b_n| = \infty$ for all $x \in L_1^\perp$, we set $L = L_1$. Otherwise there exists an $x_2 \in L_1^\perp$ with $\liminf_{n\to\infty}|x_2^T b_n| < \infty$ and we may find a further subsequence of $(b_n)_1^\infty$ such that $\lim_{n\to\infty} x_2^T b_n = a_2$. Let L_2 denote the linear space spanned by x_1 and x_2 and $L_2^\perp$ its orthogonal complement. If $\lim_{n\to\infty}|x^T b_n| = \infty$ for all $x \in L_2^\perp$, $x \neq 0$, we set $L = L_2$. Otherwise we choose $x_3 \in L_2^\perp$ and continue as above. Eventually, we obtain a linear subspace L spanned by $x_1, \dots, x_l$ such that

$$\lim_{n\to\infty} x_i^T b_n = a_i, \qquad 1 \le i \le l,$$

and $\lim_{n\to\infty}|x^T b_n| = \infty$ for all $x \in L^\perp$, $x \neq 0$. The linearity of limits shows that there exists a $b \in \mathbb{R}^k$ such that

$$\lim_{n\to\infty} x^T b_n = x^T b,$$

for all $x \in L$. We have, for any $x \in \mathbb{R}^k \setminus L$, $\lim_{n\to\infty}|x^T b_n| = \infty$ as may be seen by writing $x = x' + x''$ with $x' \in L$ and $x'' \in L^\perp$, $x'' \neq 0$. The argument leading to (4.4) now gives

$$1 - \varepsilon \le \int_{\{x\in L\}} \rho\left(\frac{y - x^T b}{s^*}\right) d\mathbb{Q} \le \int \rho\left(\frac{y - x^T b}{s^*}\right) d\mathbb{Q},$$

which shows that $(b^T, s^*)^T$ is a solution. □

4.2. *Bounded solutions.* It is not difficult to construct examples where the set of solutions $T_S(\mathbb{Q})$ is unbounded. The following theorem gives a sufficient condition for the solution set to be compact.

THEOREM 4.2. *If $\Delta(\mathbb{Q}) < 1 - \varepsilon$, then $T_S(\mathbb{Q})$ is compact.*

PROOF. We give a proof for $k = 1$. If there exists a sequence $((b_n, s^*)^T)_1^\infty$ in $T_S(\mathbb{Q})$ with $\lim|b_n| = \infty$, then the proof of Theorem 4.1 shows that $1 - \varepsilon \le \mathbb{Q}(\{x = 0\}) \le \Delta(\mathbb{Q})$, contradicting the assumption placed on $\Delta(\mathbb{Q})$. The general case follows on writing $\theta_n = b_n/\|b_n\|$ and choosing a convergent subsequence, which we continue to denote by θ_n, for which $\lim \theta_n = \theta$. The same argument now gives $1 - \varepsilon \le \mathbb{Q}(\{x^T\theta = 0\}) \le \Delta(\mathbb{Q})$, proving the theorem. □

If $\Delta(\mathbb{Q}) \ge 1 - \varepsilon$, then the following example shows that $T_S(\mathbb{Q})$ may be unbounded.

EXAMPLE 4.1. We take $k = 2$ and a distribution $\mathbb{Q}$ such that $\mathbb{Q}(\{(x^T, y)^T: y = x_1 = x_2\}) = 1 - \varepsilon$. Then $T_S = \{(b^T, 0)^T: b_1 + b_2 = 1\}$ and is unbounded.

4.3. *Breakdown points.* We now calculate the breakdown points of T_S. We require the following lemma.

LEMMA 4.1. *Let* ρ *satisfy* R1–R6. *Then, for any* $b \in \mathbb{R}^k$, $s > 0$ *and* $\mathbb{Q}_1$ *and* $\mathbb{Q}_2$ *in* $\mathfrak{W}(\mathbb{R}^{k+1})$, *we have*

$$\int \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q}_1 \le \int \rho\left(\frac{y - x^T b}{\exp(d_{\mathfrak{H}}(\mathbb{Q}_1, \mathbb{Q}_2)) s}\right) d\mathbb{Q}_2 + d_{\mathfrak{H}}(\mathbb{Q}_1, \mathbb{Q}_2).$$

PROOF. For $u > 0$ it follows from R1, R2, R5 and R6 that $\rho(u) = -\int\{u \le v < \infty\}\, d\rho(v)$. The claim can now be proved using Fubini and

$$\int \left\{\left|\frac{y - x^T b}{s}\right| \le v < \infty\right\} d\mathbb{Q}_1 \le \int \left\{\left|\frac{y - x^T b}{s}\right| \le v e^{\eta} < \infty\right\} d\mathbb{Q}_2 + \eta,$$

for any $\eta > d_{\mathfrak{H}}(\mathbb{Q}_1, \mathbb{Q}_2)$. □

The next theorem gives a lower bound for $\varepsilon^*(T_S, \mathbb{Q}, d_{\mathfrak{H}})$.

THEOREM 4.3.

$$\varepsilon^*(T_s, \mathbb{Q}, d_{\mathfrak{H}}) \ge \min(\varepsilon, 1 - \varepsilon - \Delta(\mathbb{Q})).$$

PROOF. If $\Delta(\mathbb{Q}) \ge 1 - \varepsilon$, then the claim is trivial. We therefore suppose that $\Delta(\mathbb{Q}) < 1 - \varepsilon$, choose $\eta < \min\{\varepsilon, 1 - \varepsilon - \Delta(\mathbb{Q})\}$ and consider the set of distributions $\mathbb{Q}'$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta$. From R2, R3 and R4 it follows that there exists an $s(\eta) > 0$ such that $1 - \varepsilon + \eta \le \int \rho(y/s(\eta))\, d\mathbb{Q}$ and hence, on using Lemma 4.1, we obtain

$$1 - \varepsilon - \eta \le \int \rho\left(\frac{y}{s(\eta)\exp(\eta)}\right) d\mathbb{Q}' + \eta,$$

which implies

$$s^*(\mathbb{Q}') \le s(\eta)\exp(\eta), \tag{4.5}$$

for all $\mathbb{Q}'$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta$ and where $(b^*(\mathbb{Q}')^T, s^*(\mathbb{Q}'))^T \in T_S(\mathbb{Q}')$ for some $b^*(\mathbb{Q}')$.

Suppose now that there exists a sequence $(\mathbb{Q}_n)_1^\infty$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}_n) < \eta$ and a corresponding sequence of solutions $((b_n^T, s_n)^T)_1^\infty$ of the problems $(\mathfrak{B}(\mathbb{Q}_n))_1^\infty$ with $\lim_{n \to \infty} \|b_n\| = \infty$. Then

$$\begin{aligned} 1 - \varepsilon &\le \limsup_{n \to \infty} \int \rho\left(\frac{y - x^T b_n}{s_n}\right) d\mathbb{Q}_n \le \limsup_{n \to \infty} \int \rho\left(\frac{y - x^T b_n}{s_n \exp(\eta)}\right) d\mathbb{Q} + \eta \\ &\le \limsup_{n \to \infty} \int \rho\left(\frac{y - x^T b_n}{s(\eta)\exp(\eta)}\right) d\mathbb{Q} + \eta, \end{aligned}$$

where we have used (4.5) and Lemma 4.1. The argument of the proof of Theorem 4.2 now gives $1 - \varepsilon \le \Delta(\mathbb{Q}) + \eta$, contradicting the choice of η. Thus for all $\eta < \min\{\varepsilon, 1 - \varepsilon - \Delta(\mathbb{Q})\}$ there exists a bounded subset $C(\eta)$ of $\mathbb{R}^{k+1}$ such that $T_S(\mathbb{Q}') \subset C(\eta)$ for all $\mathbb{Q}'$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta$, proving the theorem. □

The expression $\min\{\varepsilon, 1-\varepsilon-\Delta(\mathbb{Q})\}$ is maximized by setting $\varepsilon = (1 - \Delta(\mathbb{Q}))/2$ and in this case we obtain $\varepsilon^*(T_S, \mathbb{Q}, d_{\mathfrak{H}}) \geq (1-\Delta(\mathbb{Q}))/2$. This together with Theorem 3.1 gives the following result.

THEOREM 4.4. *If* $\varepsilon = (1-\Delta(\mathbb{Q}))/2$, *then* $\varepsilon^*(T_S, \mathbb{Q}, d_{\mathfrak{H}}) = (1-\Delta(\mathbb{Q}))/2$.

Theorem 4.4 is a sleight of hand as the choice of ε now depends on $\mathbb{Q}$. A similar sleight of hand in the calculation of finite sample breakdown points was commented upon in subsection 1.4. We therefore define a new functional T_S^G with ε in (4.2) given by $(1-\Delta(\mathbb{Q}))/2$. We have the following result.

THEOREM 4.5. *The breakdown point of* T_S^G *is* $\varepsilon^*(T_S^G, \mathbb{Q}, d_{\mathfrak{H}}) = (1 - \Delta(\mathbb{Q}))/3$.

PROOF. Let $\mathbb{Q}'$ be such that $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta$. Then $\Delta(\mathbb{Q}) \leq \Delta(\mathbb{Q}') + \eta$ and the proof of Theorem 4.4 may be repeated to give $\varepsilon^*(T_S^G, \mathbb{Q}, d_{\mathfrak{H}}) \geq (1 - \Delta(\mathbb{Q}))/3$. Let $\mathbb{Q}_\eta$ be a distribution of the form

$$\mathbb{Q}_\eta(B) = \eta\delta_z(B) + \mathbb{Q}\big(B \cap \{x: x^T\theta_0 = 0\}\big) + \big(1 - \eta/(1-\Delta(\mathbb{Q}))\big)\mathbb{Q}\big(B \cap \{x: \theta_0^T x \neq 0\}\big),$$

where $z^T = (\theta_1^T, y)$, θ_0 and θ_1 are in S^k, $\theta_1^T\theta_0 = 0$ and $\Delta(\mathbb{Q}) = \mathbb{Q}_d(\{x: \theta_0^T x = 0\})$. Then $\Delta(\mathbb{Q}_\eta) = \Delta(\mathbb{Q}) + \eta$ and for any $(b^T, s)^T \in T_S^G(\mathbb{Q}_\eta)$ we have

$$\big(1 + \Delta(\mathbb{Q}_\eta)\big)/2 \leq \eta\rho\left(\frac{y - \theta_1^T b}{s}\right) + 1 - \eta.$$

If $\varepsilon^*(T_S^G, \mathbb{Q}, d_{\mathfrak{H}}) > (1-\Delta(\mathbb{Q}))/3$, then we can choose η so that $(1-\Delta(\mathbb{Q}))/3 < \eta < \varepsilon^*(T_S^G, \mathbb{Q}, d_{\mathfrak{H}})$ and the set $T_S^G(\mathbb{Q}_\eta)$ is bounded. It is therefore possible to choose y so that $\rho((y - \theta_1^T b)/s) = 0$ and this implies $(1 + \Delta(\mathbb{Q}) + \eta)/2 \leq 1 - \eta$, giving $\eta \leq (1 - \Delta(\mathbb{Q}))/3$. This is a contradiction and therefore $\varepsilon^*(T_S^G, \mathbb{Q}, d_{\mathfrak{H}}) \leq (1 - \Delta(\mathbb{Q}))/3$, proving the theorem. □

The expression $(1 - \Delta(\mathbb{Q}))/3$ recurs several times in this article. It raises the question as to whether this is the best possible breakdown point for globally defined linearly equivariant functionals in the sense that a higher breakdown point can only be obtained at the expense of reducing the breakdown point below $(1 - \Delta(\mathbb{Q}))/3$ at some points or not having the functional defined at all points.

4.4. *Uniqueness and continuity of S-estimators.* We now give sufficient conditions for the S-estimator T_S to be uniquely defined. We restrict attention to distributions of the form (1.7) which arise from the usual linear model (1.6). We require the following conditions to be placed on the density function of the errors [Davies (1990)]:

F1. $f\colon \mathbb{R} \to \mathbb{R}_+$ is bounded.
F2. f is symmetric.
F3. $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ is nonincreasing.
FR1. For some $u > 0$ and for all η,
$0 < \eta < u$, $(f(u+\eta) - f(u-\eta))(\rho(u+\eta) - \rho(u-\eta)) > 0$.
FR2. $\int \rho(u) f(u)\, du = 1 - \varepsilon$.

We note here that the slightly weaker condition FR1 of Davies (1990) is not quite sufficient and should be replaced by the FR1 above. The condition FR1 is automatically satisfied if $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ is strictly nonincreasing as is the case for the normal distribution. The condition FR2 is a normalization to ensure Fisher consistency at the assumed model. Given any ρ satisfying R1–R6, one may define ρ_c by $\rho_c(u) = \rho(u/c)$ and then choose $c > 0$ so that FR2 is satisfied. This is always possible as the integral in FR2 is a continuous function of c. The function ρ_c also satisfies R1–R6.

THEOREM 4.6. *Suppose that* $\mathbb{Q}$ *is of the form* (1.7) *and that* $\mathbb{Q}_e$ *has a density* f_σ *of the form* $f_\sigma(u) = f(u/\sigma)/\sigma$, *where* f *satisfies* F1–F3, FR1 *and* FR2. *Then* $T_S = \{(\beta^T, \sigma)^T\}$.

PROOF. We have

$$\int \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q} = \int \left(\int \rho\left(\frac{u - x^T(b-\beta)}{s}\right) f_\sigma(u)\, du \right) d\mathbb{Q}_d(x) \tag{4.6}$$
$$\leq \int \left(\int \rho\left(\frac{u}{s}\right) f_\sigma(u)\, du \right) d\mathbb{Q}_d(x),$$

because of R1, R2, F2 and F3 [see Lemma 1(ii) of Davies (1990)]. The integral in (4.6) is a strictly increasing function of s and it therefore follows from FR2 that $s = \sigma$ minimizes s subject to the constraint (4.2). For $s = \sigma$ we obtain strict inequality in (4.2) because of FR1. This implies that $x^T(b - \beta) = 0$ $\mathbb{Q}_d$-a.e. x and hence, as $\Delta(\mathbb{Q}_d) < 1$, we must have $b = \beta$, proving the theorem. □

We now turn to the continuity of T_S and introduce the conditions

FR3 $$\lim_{v \to 0} \frac{1}{v} \int \left(\rho\left(\frac{y}{1+v}\right) - \rho(y) \right) f(y)\, dy = c_3 > 0,$$

FR4 $$\lim_{(u,v) \to 0} \frac{1}{u^2} \int \left(\rho\left(\frac{y+u}{1+v}\right) - \rho\left(\frac{y}{1+v}\right) \right) f(y)\, dy = -c_4 < 0.$$

We have the following result.

THEOREM 4.7. *Suppose that* $\mathbb{Q}$ *is as in Theorem* 4.6 *with* $T_S(\mathbb{Q}) = \{(\beta^T, \sigma)^T\}$, *that* FR3 *and* FR4 *hold and that* $\min\{\varepsilon, 1 - \varepsilon - \Delta(\mathbb{Q})\} > 0$. *Then,* *for all* η_0, $0 < \eta_0 < \min\{\varepsilon, 1 - \varepsilon - \Delta(\mathbb{Q})\}$, *there exists a constant* $c_5 = c_5(\eta)$ *such that*

$$T_S(\mathbb{Q}') \subset \left\{(b^T, s)^T : \|b - \beta\|^2 + |s - \sigma| < c_5 d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}')\right\},$$

for all $\mathbb{Q}'$ *with* $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta_0$.

PROOF. As the functional T_S is linearly equivariant, it is sufficient to consider the case $\beta = 0$ and $\sigma = 1$. We note first that by Theorem 4.2 there exists a compact subset $K(\eta_0)$ of $\mathbb{R}^{k+1}$ such that $T_S(\mathbb{Q}') \subset K$ for all $\mathbb{Q}'$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') < \eta_0$. If $(b^T, s)^T \in T_S(\mathbb{Q}')$ we have, on writing $\eta = d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}')$,

$$1 - \varepsilon \le \int \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q}' \le \int \rho\left(\frac{y - x^T b}{\exp(\eta) s}\right) d\mathbb{Q} + \eta \le \int \rho\left(\frac{y}{\exp(\eta) s}\right) d\mathbb{Q} + \eta$$
$$= \int \rho\left(\frac{y}{\exp(\eta) s}\right) f(y)\, dy + \eta$$

by Lemma 4.1 and the argument leading to (4.6). Thus

$$\int \rho\left(\frac{y}{\exp(\eta) s}\right) f(y)\, dy \ge 1 - \varepsilon - \eta$$

and hence, by FR2 and FR3, $c_6(\exp(\eta)s - 1) \ge -\eta$ for some constant c_6 which implies

$$s - 1 \ge -c_7 \eta, \tag{4.7}$$

for some constant $c_7 > 0$. In the other direction FR3 gives

$$1 - \varepsilon + \eta \le \int \rho\left(\frac{y}{\exp(c_8 \eta)}\right) f(y)\, dy,$$

for some c_8 and hence

$$1 - \varepsilon + \eta \le \int \rho\left(\frac{y}{\exp(c_8 \eta)}\right) d\mathbb{Q} \le \int \rho\left(\frac{y}{\exp((c_8 + 1)\eta)}\right) d\mathbb{Q}' + \eta,$$

which yields $s \le \exp(c_9 \eta)$ for some constant $c_9 > 0$. This together with (4.7) yields

$$|s - 1| \le c_{10} d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}'). \tag{4.8}$$

We now turn to the location part. Using (4.8), we have

$$1 - \varepsilon \le \int \rho\left(\frac{y - x^T b}{s}\right) d\mathbb{Q}' \le \int \rho\left(\frac{y - x^T b}{\exp(c_{11}\eta)}\right) d\mathbb{Q}' \le \int \rho\left(\frac{y - x^T b}{\exp(c_{12}\eta)}\right) d\mathbb{Q} + \eta$$
$$= \int \left(\int \rho\left(\frac{y - x^T b}{\exp(c_{11} 2\eta)}\right) f(y)\, dy\right) d\mathbb{Q}_d(x) + \eta.$$

On using FR2 and FR3 we obtain

$$\int\left(\int\left(\rho\left(\frac{y}{\exp(c_{12}\eta)}\right)-\rho\left(\frac{y-x^Tb}{\exp(c_{12}\eta)}\right)\right)f(y)\,dy\right)d\mathbb{Q}_d(x)\geq c_{13}\eta,$$

with $c_{13}>0$. From FR4 we may deduce

$$\int\frac{b^Txx^Tb}{1+b^Txx^Tb}\,d\mathbb{Q}_d\leq c_{14}\eta,$$

with $c_{14}>0$. As $\Delta(\mathbb{Q})<1$ there exists a c_{15} such that $\int\{\|x\|\leq c_{15}\}xx^T\,d\mathbb{Q}_d$ is strictly positive definite and this is seen to imply $\|b\|^2\leq c_{16}\eta$, proving the theorem. □

Theorems 4.6 and 4.7 also hold for the functional T_S^G.

The following example shows that in general we cannot improve on a Hölder condition of order $1/2$ even for smooth ρ.

EXAMPLE 4.2. For simplicity, we assume that $k=1$, that $\mathbb{Q}$ has compact support and that the conditions of Theorem 4.7 are satisfied. Without loss of generality we may take $T_S(\mathbb{Q})=(0,1)^T$. Let ρ have a continuous second derivative and choose c with $\rho^{(1)}(c)>0$. We now consider the effect of a point contamination at $(u,c)^T$. Let $\mathbb{Q}_\varepsilon=(1-\varepsilon)\mathbb{Q}+\varepsilon\delta_{(u,c)^T}$ and write $T_s(\mathbb{Q}_\varepsilon)=(b(\varepsilon),s(\varepsilon))^T$. From Theorem 4.7 it follows that

$$|b(\varepsilon)|+|1-s(\varepsilon)|=O(\sqrt{\varepsilon})\tag{4.9}$$

uniformly in u. Now $b(\varepsilon)$ maximizes

$$(1-\varepsilon)\int\rho\left(\frac{y-xb}{s(\varepsilon)}\right)d\mathbb{Q}+\varepsilon\rho\left(\frac{c-ub(\varepsilon)}{s(\varepsilon)}\right),$$

with respect to b so that, on differentiating,

$$(1-\varepsilon)\int\frac{x}{s(\varepsilon)}\rho^{(1)}\left(\frac{y-xb(\varepsilon)}{s(\varepsilon)}\right)d\mathbb{Q}+\frac{\varepsilon u}{s(\varepsilon)}\rho^{(1)}\left(\frac{c-ub(\varepsilon)}{s(\varepsilon)}\right)=0.$$

Because of F2 we have

$$\int x\rho^{(1)}\left(\frac{y}{s(\varepsilon)}\right)d\mathbb{Q}=0$$

and a Taylor expansion gives

$$-(1-\varepsilon)b(\varepsilon)\int\frac{x^2}{s(\varepsilon)}\rho^{(2)}\left(\frac{y-\theta xb(\varepsilon)}{s(\varepsilon)}\right)d\mathbb{Q}+\varepsilon u\rho^{(1)}\left(\frac{c-ub(\varepsilon)}{s(\varepsilon)}\right)=0,$$

for some θ, $0<\theta<1$. For small ε we have, on using (4.9),

$$c'b(\varepsilon)(1+o(1))=\varepsilon u\rho^{(1)}\left(\frac{c-ub(\varepsilon)}{s(\varepsilon)}\right),\tag{4.10}$$

with $c' \neq 0$. We now set $u = 1/\sqrt{\varepsilon}$ and write $b(\varepsilon) = \gamma(\varepsilon)\sqrt{\varepsilon}$. This gives

$$c'\gamma(\varepsilon)(1 + o(1)) = \rho^{(1)}\left(\frac{c - \gamma(\varepsilon)}{s(\varepsilon)}\right).$$

As $\lim_{\varepsilon \to 0} s(\varepsilon) = 1$ and $\rho^{(1)}(c) \neq 0$, we see that $\liminf|\gamma(\varepsilon)|/\sqrt{\varepsilon} > 0$ which shows that the result of Theorem 4.7 cannot be improved upon.

We note that the continuity of T_S was proved only for theoretical distributions which have density and for which FR3 and FR4 hold. Theorem 4.7 applies to the Hampel–Rousseeuw least median of squares functional T_{LMS}. The fact that one needs strong conditions to obtain continuity indicates that the functional will not in general be continuous at any given regression distribution $\mathbb{Q}$. Indeed it is easy to construct distributions where T_{LMS} is not continuous and the work of Hettmansperger and Sheather (1992) suggests that this is not only a theoretical possibility.

5. Bounded influence regression.

5.1. *Definition.* Assuming the scale parameter to be known, M-estimators T_ψ of the form

$$\int x\psi\big(y - x^T T_\psi(\mathbb{Q})\big)\, d\mathbb{Q} = 0$$

have an influence function given by

$$\mathrm{IF}\big(x^T, y; T_\psi, \mathbb{Q}\big) = \psi\big(y - x^T T_\psi(\mathbb{Q})\big) M^{-1}(\psi, \mathbb{Q}) x, \tag{5.1}$$

where

$$M(\psi, \mathbb{Q}) = \int x x^T \psi^{(1)}\big(y - x^T T_\psi(\mathbb{Q})\big)\, d\mathbb{Q}$$

[Hampel, Rousseeuw, Ronchetti and Stahel (1986), page 316]. Assuming $M(\psi, \mathbb{Q})$ to be finite, it is clear from (5.1) that the influence function is bounded in the residual space but not in the factor space. Moreover, it may be shown that the breakdown point is 0.

Proposals to modify M-estimators so as to obtain bounded influence functions have been made by several authors [Hampel (1978), Krasker (1980), Krasker and Welsch (1982) and Ronchetti and Rousseeuw (1985)]. Some of these proposals have certain optimality properties such as minimizing the trace of the asymptotic covariance of the estimator of the regression coefficients subject to a bound on the maximal influence. All proposals are based on robust measures of dispersion for the design distribution $\mathbb{Q}_d$, permitting a downweighting of infinitesimal contamination at the leverage points. This leads to estimators of the form

$$\int x w(T^D(\mathbb{Q}_d)x)\psi\Big(\big(y - x^T T_\psi(\mathbb{Q})\big)/w(T^D(\mathbb{Q}_d)x)\Big)\, d\mathbb{Q} = 0,$$

where w is some weight function and T^D is a dispersion functional. An example is $w(x) = 1/\|x\|$.

5.2. *Breakdown points of dispersion functionals.* Both the Hampel–Krasker and Krasker–Welsch dispersion functionals are M-estimators and as such have a breakdown point of at most $1/(k+1)$ [Maronna (1976)]. As far as I am aware the actual breakdown points have not been published, at least not for full metric neighbourhoods.

The Hampel–Krasker dispersion estimator T^D_{HK} is DD^T, where D is a symmetric positive definite $k \times k$-matrix which is a solution of

$$\int (2\Phi(c/\|Dx\|) - 1) Dx(Dx)^T \, d\mathbb{Q}_d = D, \tag{5.2}$$

where Φ denotes the $\mathfrak{N}(0,1)$-distribution function. If $\int \|x\| \, d\mathbb{Q}_d < \infty$ it is known that a solution exists for c sufficiently large and we shall assume that this is the case.

THEOREM 5.1. *Let T^D_{HK} be the Hampel–Krasker estimator and $\mathbb{Q}_d$ a design distribution with $\Delta(\mathbb{Q}_d) = 0$ and $\int \|x\| \, d\mathbb{Q}_d < \infty$. Then $\varepsilon^*(T^D_{\mathrm{HK}}, \mathbb{Q}_d, d^D) = 0$ for any linearly invariant metric d^D which satisfies* (3.5).

PROOF. For η, $0 < \eta < 1$, we set

$$\mathbb{Q}_{\eta d} = \mathbb{Q}_{\eta d}(\eta, R, \theta_1, \dots, \theta_k) = (1-\eta)\mathbb{Q}_d + \eta \sum_1^k \delta_{R\theta_j}/k,$$

where $\theta_1, \dots, \theta_k$ are orthogonal vectors in $\mathbb{R}^k$ with $\|\theta_j\| = 1$, $1 \le j \le k$. Let $D = D(\eta, R, \theta_1, \dots, \theta_k)$ be the solution of (5.2) with $\mathbb{Q}_{\eta d}$ in place of $\mathbb{Q}_d$. On multiplying on the left by θ_j^T and on the right by θ_j, we obtain

$$\eta R^2 \|D\theta_j\|^2 \left(2\Phi\left(\frac{c}{R\|D\theta_j\|} \right) - 1 \right) \le k\|D\theta_j\|.$$

This is only possible if $\lim_{R \to \infty} R\|D\theta_j\| = 0$, implying that the estimator has broken down. □

For the proof of the theorem it is only necessary to consider $(1-\eta)\mathbb{Q}_d + \eta\delta_{R\theta}$ for some fixed θ. However, the contamination we have chosen implies

$$\lim_{R \to \infty} \|Dx\| = 0, \tag{5.3}$$

for all x as may be seen by noting $\lim_{k \to \infty} R\|D\theta_j\| = 0$ for all j.

The next theorem covers the Krasker–Welsch dispersion estimator T^D_{KW} which is defined as DD^T, where D is the solution of

$$\int \left(\int \psi_c(u\|Dx\|)^2 \, d\mathfrak{N}(0,1) \right) Dx(Dx)^T / \|Dx\|^2 \, d\mathbb{Q}_d = I_k, \tag{5.4}$$

where ψ_c denotes the Huber ψ-function with cutoff point c. Using the results of Maronna (1976), it can be shown that a solution of (5.4) exists if $c^2 > k$ and $\Delta(\mathbb{Q}_d) = 0$.

THEOREM 5.2. *Let T^D_{KW} be the Krasker–Welsch dispersion estimator with tuning constant c, $c^2 > k$, and $\mathbb{Q}_d$ a design distribution with $\Delta(\mathbb{Q}_d) = 0$. Then*

$$\varepsilon^*\big(T^D_{\mathrm{KW}}, \mathbb{Q}_d, d^D_{\mathfrak{H}}\big) = \min\{1/c^2, 1 - k/c^2\} \le 1/(k+1).$$

PROOF. Let η, $c^{-2} < \eta < 1$, be fixed and consider the distribution $\mathbb{Q}_{\eta d}$, $\mathbb{Q}_{\eta d} = \mathbb{Q}_{\eta d}(\eta, R, \theta) = (1-\eta)\mathbb{Q}_d + \eta\delta_{R\theta}$. On multiplying (5.4) on the left by $(D\theta)^T/\|D\theta\|$ and on the right by $D\theta/\|D\theta\|$, we obtain $\eta\int\psi_c(uR\|D\theta\|)^2\, d\mathfrak{N}(0,1))$ ≤ 1 which is only possible if $R\|D\theta\|$ remains bounded as $R \to \infty$, implying that the estimator has broken down. We have therefore shown

$$\varepsilon^*\big(T^D_{\mathrm{KW}}, \mathbb{Q}_d, d^D_{\mathfrak{H}}\big) \le 1/c^2. \tag{5.5}$$

If now $\eta > 1 - k/c^2$ we have, on taking traces in (5.4), letting $R \to 0$ and assuming that the eigenvalues of D remain bounded,

$$k = \lim_{R\to 0}(1-\eta)\int\left(\int\psi_c(u\|Dx\|)^2\, d\mathfrak{N}(0,1)\right) d\mathbb{Q}_d \le (1-\eta)c^2,$$

which contradicts the choice of η. This shows $\varepsilon^*(T^D_{\mathrm{KW}}, \mathbb{Q}_d, d^D_{\mathfrak{H}}) \le 1 - k/c^2$ which together with (5.5) implies

$$\varepsilon^*\big(T^D_{\mathrm{KW}}, \mathbb{Q}_d, d^D_{\mathfrak{H}}\big) \le \min\{1/c^2, 1 - k/c^2\}. \tag{5.6}$$

In the other direction we argue as follows. Let $(\mathbb{Q}_{dn})_1^\infty$ be a sequence of design distributions with corresponding matrices $(D_n)_1^\infty$. Let $(\lambda_{\min}(n))_1^\infty$ denote the smallest eigenvalues of the $(D_n)_1^\infty$ with eigenvectors $(\theta_n)_1^\infty$ in S^k. Suppose that $\lim_{n\to\infty}\lambda_{\min}(n) = 0$. Then

$$\begin{aligned}1 &= \int\left(\int\psi_c(u\|D_n x\|)^2\, d\mathfrak{N}(0,1)\right)\frac{\left(\theta_n^T D_n x\right)^2}{\|D_n x\|^2}\, d\mathbb{Q}_{dn}\\ &\le c^2\int_{B_n}\frac{\left(\lambda_{\min}(n)\theta_n^T x\right)^2}{\|D_n x\|^2}\, d\mathbb{Q}_{dn} + \int_{B_n^c}\left(\lambda_{\min}(n)\theta_n^T x\right)^2 d\mathbb{Q}_{dn},\end{aligned}$$

where $B_n = \{x\colon |\theta_n^T x| \ge \lambda_{\min}(n)^{-1/2}\}$ and B_n^c denotes the complement of B_n. On noting that $\|D_n x\| \ge \lambda_{\min}(n)\|x\|$, we obtain

$$\begin{aligned}1 \le c^2\mathbb{Q}_{dn}(B_n) + \lambda_{\min}(n) &\le c^2\Big(\mathbb{Q}_d\Big(\big\{x\colon |\theta_n^T x| \ge \lambda_{\min}(n)^{-1/2}\exp(-\eta)\big\}\Big) + \eta\Big)\\ &\quad + \lambda_{\min}(n).\end{aligned}$$

On letting n tend to ∞ and remembering $\delta(\mathbb{Q}_d) = 0$, we deduce $1 \le c^2\eta$. This shows that if $\eta < c^{-2}$, then T^D_{KW} cannot break down in that the smallest eigenvalues tend to 0. It remains to show that if $\eta < 1 - k/c^2$, then it cannot break down in that the largest eigenvalues tend to ∞.

Let $(\mathbb{Q}_{dn})_1^\infty$ be as above but with the maximum eigenvalues $\lambda_{\max}(n)$ tending to ∞. With $\theta_n \in S^k$ as the corresponding eigenvectors, we obtain on taking traces

$$\begin{aligned} k &= \lim_{n\to\infty} \int \left(\int \psi_c(u\|D_n x\|)^2 \, d\mathfrak{N}(0,1) \right) d\mathbb{Q}_{dn} \\ &\geq \lim_{n\to\infty} \int \left(\int \psi_c(u|\lambda_{\max}(n)\theta_n^T x|)^2 \, d\mathfrak{N}(0,1) \right) d\mathbb{Q}_{dn} \\ &\geq c^2 \liminf_{n\to\infty} \mathbb{Q}_{dn}\left(\left\{x \colon |\theta_n^T x| > \lambda_{\max}(n)^{-1/2}\right\}\right) \\ &\geq c^2 \liminf_{n\to\infty} \left(\mathbb{Q}_d\left(\left\{x \colon |\theta_n^T x| > \lambda_{\max}(n)^{-1/2} \exp(\eta)\right\}\right) - \eta \right) \\ &= c^2(1-\eta). \end{aligned}$$

This is only possible if $\eta \geq 1 - k/c^2$ and hence $\varepsilon^*(T_{\mathrm{KW}}^D, \mathbb{Q}_d, d_{\mathfrak{H}}^D) \geq \min\{1/c^2, 1 - k/c^2\}$ which together with (5.6) proves the theorem. □

5.3. *Breakdown points of regression functionals.* Maronna and Yohai (1991) calculate the breakdown points of general M-estimators in the linear regression model. [See also Maronna, Bustos and Yohai (1979)]. However, their calculations do not take into account a possible breakdown of the dispersion functional which we considered in subsection 5.1. They do not define breakdown in terms of metrics but use the gross error model. Their results are therefore not directly comparable with those obtained here.

The Hampel–Krasker regression functional is defined to be the solution b of

$$\int x\psi_c((y - b^T x)\|Dx\|)/\|Dx\| \, d\mathbb{Q} = 0, \tag{5.7}$$

where $DD^T = T_{\mathrm{HK}}^D(\mathbb{Q}_d)$. We have the following result.

THEOREM 5.3. *Let T_{HK} be the Hampel–Krasker estimator and $\mathbb{Q}$ a regression distribution with $\Delta(\mathbb{Q}) = 0$. Then $\varepsilon^*(T_{\mathrm{HK}}, \mathbb{Q}, d) = 0$ for any metric d which satisfies* (3.5).

PROOF. Let

$$\mathbb{Q}_\eta(\eta, y_1, \ldots, y_k, R, \theta_1, \ldots, \theta_k) = (1-\eta)\mathbb{Q} + \frac{\eta}{k}\sum_1^k \delta_{(R\theta_j^T, y_j)^T}$$

and consider now the defining equation (5.7) for $T_{\mathrm{HK}}(\mathbb{Q}_\eta) = b$. We have

$$\begin{aligned} &(1-\eta)\int Dx\psi_c((y - b^T x)\|Dx\|)/\|Dx\| \, d\mathbb{Q} \\ &\quad = \frac{\eta}{k}\sum D\theta_j \psi_c\left(\left(y_j - Rb^T\theta_j\right)R\|D\theta_j\|\right)/\|D\theta_j\| \end{aligned} \tag{5.8}$$

and as $\|b\|$ remains bounded we can choose the y_j so that $\psi_c((y_j - Rb^T\theta_j)R\|D\theta_j\|) = \pm c$, $1 \leq j \leq k$. By an appropriate choice of the $\pm$ signs, we can arrange that the right-hand side of (5.8) has modulus at least $\eta k^{-1/2}$. As $\|Dx\|$ tends to 0 for all x as R tends to ∞, by (5.3) the left-hand side of (5.8) tends to 0 by dominated convergence. Thus $\|b\|$ cannot remain bounded and the estimator has broken down, proving the theorem. □

We now turn to the Krasker–Welsch estimator T_{KW} which is defined by (5.7) with $DD^T = T_{\mathrm{KW}}^D(\mathbb{Q}_d)$. We prove the following result.

THEOREM 5.4. *Let T_{KW} be the Krasker–Welsch estimator with tuning constant c, $c^2 > k$, and $\mathbb{Q}$ a regression design distribution with $\Delta(\mathbb{Q}) = 0$. Then always*

$$\frac{1}{2c^2} \le \varepsilon^*(T_{\mathrm{KW}}, \mathbb{Q}, d_{\mathfrak{K}\mathfrak{W}}) \le \min\left\{\frac{k}{c^2}, \max\left\{1 - \frac{k}{c^2}, \frac{c^2 - k + 1}{2c^2 - k + 1}\right\}\right\} \le \frac{1}{2}.$$

If $\int \|x\|(\|x\| + |y|)\, d\mathbb{Q} < \infty$, then

$$\varepsilon^*(T_{\mathrm{KW}}, \mathbb{Q}, d_{\mathfrak{K}\mathfrak{W}}) \le \min\left\{\frac{1}{c^2}, \frac{c^2 + 1 - k}{2c^2 + 1 - k}\right\} \le \frac{2}{k + 1 + \sqrt{k^2 - 2k + 5}}.$$

PROOF. Let $(\mathbb{Q}_n)_1^\infty$ be a sequence of regression distributions for which $d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}, \mathbb{Q}_n) < \eta < 1/2c^2$. We denote the corresponding Krasker–Welsch dispersion estimates by $(D_n D_n^T)_1^\infty$ and the corresponding regression estimates by $(b_n)_1^\infty$ and suppose that $\|b_n\|$ tends to ∞ with n. From the definition of the Krasker–Welsch estimator with $\theta_n = D_n^{-1} b_n / \|D_n^{-1} b_n\|$, we have

$$\int \theta_n^T D_n x \psi_c\big((y - R_n \theta_n^T D_n x)\|D_n x\|\big)/\|D_n x\|\, d\mathbb{Q}_n = 0,$$

where $R_n = \|D_n^{-1} b_n\|$. We define the sets $(B_n)_1^\infty$ by

$$B_n = \left\{(x^T, y)^T : \theta_n^T D_n x > 0, (y - R_n \theta_n^T D_n x)\|D_n x\| \le -c\right\}$$
$$\cup \left\{(x^T, y)^T : \theta_n^T D_n x < 0, (y - R_n \theta_n^T D_n x)\|D_n x\| \ge c\right\}.$$

This implies $\int_{B_n} |\theta_n^T D_n x| / \|D_n x\|\, d\mathbb{Q}_n \le \mathbb{Q}_n(B_n^c)$, where B_n^c denotes the complement of B_n. On noting that $d_{\mathfrak{K}\mathfrak{W}}(\mathbb{Q}, \mathbb{Q}_n) < \eta$, we obtain

$$\limsup_{n \to \infty} \int_{B_n} |\theta_n^T D_n x| / \|D_n x\|\, d\mathbb{Q}_n \le \mathbb{Q}(B_n^c) + \eta.$$

The eigenvalues of D_n are bounded away from 0 (Theorem 5.2), giving $\lim_{n \to \infty} R_n = \infty$. From this and $\Delta(\mathbb{Q}) = 0$, it follows that $\limsup_{n \to \infty} \mathbb{Q}(B_n^c) = 0$ and hence

$$\limsup_{n \to \infty} \int_{B_n} |\theta_n^T D_n x| / \|D_n x\|\, d\mathbb{Q}_n \le \eta.$$

This implies

$$\limsup_{n\to\infty}\int_{B_n}|\theta_n^T D_n x|^2/\|D_n x\|^2\,d\mathbb{Q}_n \le \eta,$$

from which we may deduce

$$\limsup_{n\to\infty}\int_{B_n}\left(\int\psi_c(u\|D_n x\|)^2\,d\mathfrak{N}(0,1)\right)|\theta_n^T D_n x|^2/\|D_n x\|^2\,d\mathbb{Q}_n \le c^2\eta.$$

The definition of the Krasker–Welsch dispersion estimator gives

$$\left(1-\liminf_{n\to\infty}\int_{B_n^c}\left(\int\psi_c(u\|D_n x\|)^2\,d\mathfrak{N}(0,1)\right)|\theta_n^T D_n x|^2/\|D_n x\|^2\,d\mathbb{Q}_n\right)\le c^2\eta$$

and we therefore obtain $(1 - c^2 \liminf_{n\to\infty}\mathbb{Q}_n(B_n^c)) \le c^2\eta$. As $\liminf_{n\to\infty}\mathbb{Q}_n(B_n^c)\le\eta$ by the argument given above, we may deduce $(1 - c^2\eta)\le c^2\eta$, contradicting the choice of η. We have therefore proved $\varepsilon^*(T_{\mathrm{KW}},\mathbb{Q},d_{\mathfrak{RW}})\ge 1/2c^2$.

We now consider the opposite inequality. Let $\eta > k/c^2$ and $\mathbb{Q}_\eta$ be a regression distribution of the form

$$\mathbb{Q}_\eta = (1-\eta)\mathbb{Q} + \frac{\eta}{k}\sum_{j=1}^{k}\delta_{(R\theta_j^T,y_j)^T}, \tag{5.9}$$

where $\theta_j\in S^k$ for $1\le j\le k$ and the θ_j are orthogonal. Let $b = b(\eta,R,\theta_1,\dots,\theta_k)$ denote the corresponding regression estimate and DD^T with $D = D(\eta,R,\theta_1,\dots,\theta_k)$ the Krasker–Welsch dispersion estimate. We note that

$$\lim_{R\to\infty} Dx = 0, \tag{5.10}$$

for all x. Indeed, the proof of Theorem 5.2 shows $\lim D\theta_j = 0$ for all j from which (5.10) follows. Suppose that the $\|b\|$ remain bounded as R tends to ∞. As

$$\begin{aligned}(1-\eta)\int Dx\psi_c\big((y-b^Tx)\|Dx\|\big)/\|Dx\|\,d\mathbb{Q}\\ = -\frac{\eta}{k}\sum D\theta_j\psi_c\big(\big(y_j - Rb^T\theta_j\big)R\|D\theta_j\|\big)/\|D\theta_j\|,\end{aligned} \tag{5.11}$$

we can arrange, as in the proof of Theorem 5.3, that the right-hand side of (5.11) has modulus at least $\eta k^{-1/2}$. The left-hand side tends to 0 as R tends to ∞ by (5.10) and dominated convergence so that $\|b\|$ cannot be bounded and the Krasker–Welsch estimator has broken down.

Consider now the case

$$\eta > \max\left\{1-\frac{k}{c^2},\frac{c^2-k+1}{2c^2-k+1}\right\}.$$

We set

$$\mathbb{Q}_\eta = (1-\eta)\mathbb{Q} + \eta\delta_{(y_0,R\theta^T)^T},$$

for some $\theta \in S^k$ and $y_0 \in \mathbb{R}$ and write $DD^T = T^D_{\mathrm{KW}}(\mathbb{Q}_{\eta d})$. The proof of Theorem 5.2 shows $\lim_{R \to 0}\|D\theta\| = \infty$ which together with $\delta(\mathbb{Q}) = 0$ implies

$$\lim_{R \to 0} \|Dx\| = \infty, \tag{5.12}$$

for all x. On writing $b = T_{\mathrm{KW}}(\mathbb{Q}_\eta)$, we have

$$\begin{aligned} &(1-\eta)\int Dx\psi_c\big((y - b^T x)\|Dx\|\big)/\|Dx\|\,d\mathbb{Q} \\ &\qquad = -\eta D\theta\psi_c\big((y_0 - Rb^T\theta)R\|D\theta\|\big)/\|D\theta\|. \end{aligned} \tag{5.13}$$

If now $\|b\|$ remains bounded as R tends to 0, we can choose y_0 so that $\psi_c((y_0 - Rb^T\theta)R\|D\theta\|) = -c$. We write $\varphi = D\theta/\|D\theta\|$ and, on multiplying both sides of (5.13) by φ^T, we obtain

$$(1-\eta)\left|\int \theta^T Dx\psi_c\big((y - b^T x)\|Dx\|\big)/\|Dx\|\,d\mathbb{Q}\right| = c\eta.$$

On squaring this, using (5.12) and the Cauchy–Schwarz inequality, we may deduce

$$\begin{aligned} \eta^2 c^2 &\le \lim_{R\to 0}(1-\eta)^2\int c^2|\varphi^T Dx|^2/\|Dx\|^2\,d\mathbb{Q} \\ &= (1-\eta)\left(1 - \lim_{R\to 0}\eta\int \psi_c(uR\|D\theta\|)^2\,d\mathfrak{N}(0,1)\right), \end{aligned}$$

where we have used the definition of D and again (5.12). On taking traces in (5.4) we have, once again using (5.12),

$$(1-\eta)c^2 + \lim_{R\to 0}\eta\int\psi_c(uR\|D\theta\|)^2\,d\mathfrak{N}(0,1) = k$$

and hence $\eta^2c^2 \le (1-\eta)(1 - k + (1-\eta)c^2)$ which contradicts $\eta > (c^2 - k + 1)/(2c^2 - k + 1)$. Thus the $\|b\|$ cannot remain bounded and the Krasker–Welsch estimator has broken down. This proves the first set of inequalities.

We now turn to the case $\int\|x\|(\|x\| + |y|)\,d\mathbb{Q} < \infty$ and consider the measure $\mathbb{Q}_\eta = (1-\eta)\mathbb{Q} + \eta\delta_{(y_0,\, R\theta^T)^T}$. Arguing as before, we obtain

$$\lim_{R\to\infty}\|D\theta\| = 0, \tag{5.14}$$

where $DD^T = T^D_{\mathrm{KW}}(\mathbb{Q}_{\eta d})$. If $b = T_{\mathrm{KW}}(\mathbb{Q}_\eta)$ remains bounded, we may choose y_0 so that

$$(1-\eta)\int\theta^T x\psi_c\big((y - b^T x)\|Dx\|\big)/\|Dx\|\,d\mathbb{Q} = \eta c/\|D\theta\|.$$

From this and the assumption of the theorem, we have $(1-\eta)K\int\|x\|(\|x\| + |y|)\,d\mathbb{Q} \ge \eta/\|D\theta\|$ for some constant K. This, however, conflicts with (5.14) and hence T_{KW} has broken down. □

6. Global dispersion measures.

6.1. *The problem.* A dispersion functional T^D is defined on the set $\mathfrak{W}_d(\mathbb{R}^k)$ of all nondegenerate design distributions over $\mathfrak{B}(\mathbb{R}^k)$ and takes values in the space PDS(k) of symmetric, strictly positive definite $k \times k$-matrices. In subsection 1.2 it was argued that such a functional T^D should have the following "desirable properties:"

DP1: T^D: $\mathfrak{W}_d(\mathbb{R}^k) \to$ PDS(k) should be well defined.
DP2: T^D should be linearly equivariant, $T^D(\mathbb{Q}_d^L) = LT^D(\mathbb{Q}_d)L^T$ for all nonsingular linear transformations L: $\mathbb{R}^k \to \mathbb{R}^k$.
DP3: T^D should have a high breakdown point at each $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$.
DP4: T^D should be Fréchet differentiable at each point $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$.

We shall not be able to exhibit a functional T^D satisfying DP1–DP4. If, however, we weaken DP4 to

DP4′: T^D should be locally Lipschitz at every $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$,

then we shall be able to exhibit a functional satisfying DP1, DP2, DP3 and DP4′. This section is devoted to the construction of such a functional.

It is perhaps worth making a comment as to why DP2 is restricted to nonsingular linear mappings. First, if nonsingular mappings L are allowed, then we must replace PDS(k) by the set of symmetric nonnegative definite $k \times k$-matrices. If this is done and T^D is any globally defined dispersion functional which is linearly equivariant for all linear mappings L, the following hold. If $\mathbb{Q}_d = (1/n)\Sigma_1^n \delta_{x_j}$ is an empirical distribution, then $T^D(\mathbb{Q}_d)$ is a quadratic polynomial in the components of the x_j. This corresponds to a result of Donoho (1982), Obenchain (1971) and Rousseeuw (1986) for the case of linearly equivariant location functionals. Furthermore, for any such T^D we have $T^D(\mathbb{Q}_d) = 0$ for any $\mathbb{Q}_d$ with either a Cauchy or a Gaussian marginal distribution. These partial results are sufficient to indicate that demanding linear equivariance for singular linear transformations is unreasonable; it is simply too strong.

One ad hoc method of obtaining a functional which satisfies DP1 and DP2 is to consider equivalence classes of distributions which may be transformed into each other by nonsingular linear mappings. The functional T^D may then be defined in an arbitrary way for one member of each equivalence class and then by linear equivariance for the remaining members of the class. Such a method would not in general fulfill DP3 or DP4′.

One use of high breakdown dispersion functionals is in detecting outliers in the factor space. Rousseeuw and van Zomeren (1990) proposed using Rousseeuw's minimum volume ellipsoid functional [Rousseeuw (1986)]. Simpson, Ruppert and Carroll (1992) proposed using S-functionals as defined in Davies (1987) in order to obtain a high breakdown. From the present point of view, the problem with such dispersion functionals is that they require regularity conditions on the design distributions $\mathbb{Q}_d$ [Davies (1987), (1992a)]. S-functionals are also proposed because they have an influence function at the model. The reader is referred to Davies (1987) and Lopuhäa (1989). However, both these papers place restrictive conditions on the design distribution $\mathbb{Q}_d$ in order to obtain a uniquely defined functional with an influence function. In particular, $\mathbb{Q}_d$ is assumed to have an elliptical distribution with a unimodal density f. This condition is only a sufficient one, but the fact that it has to be imposed for the proof of uniqueness does indicate that some condition is required. This is confirmed by Example 6.1 which gives a design distribution at which no S-estimator can be uniquely defined in a linearly equivariant manner.

EXAMPLE 6.1. We set $k = 2$ and let $\mathbb{Q}_d = 0.4\mathbb{Q}_{\mathrm{ud}}(0, r) + 0.15\Sigma_1^4\mathbb{Q}_{\mathrm{ud}}(x_j, r)$, where $\mathbb{Q}_{\mathrm{ud}}(x, r)$ denotes the uniform distribution over the ball with centre x and radius r. The points x_j are given by $x_1 = (1, 0)^T$, $x_2 = (0, 1)^T$, $x_3 = (-1, 0)^T$ and $x_4 = (0, -1)^T$. Without going into details it is clear that for sufficiently small r any minimum volume ellipsoid will concentrate on the component of $\mathbb{Q}_d$ located at the origin and two diametrically opposed components of the remaining four. There will therefore be two solutions and, depending on which one is chosen, there will be two different sets of leverage points. Because of the nature of $\mathbb{Q}_d$ any uniquely defined linearly equivariant dispersion functional evaluated at $\mathbb{Q}_d$ must be a multiple of the unit matrix I_2.

It is easy, of course, to produce a discrete design distribution where an S-functional is uniquely defined and even Fréchet differentiable. However, Example 6.1 shows that it is not possible to obtain a general result for all design distributions. In practice, this will mean that for any given data set it is not known whether the functional is well defined or not. This would seem to be a fundamental difficulty with S-functionals. In order to construct a dispersion functional which satisfies DP1, DP2, DP3 and DP4′, we therefore take another approach.

6.2. *The Donoho–Stahel dispersion functional.* Independently of each other Donoho (1982) and Stahel (1981) proposed the first affine equivariant dispersion functional for empirical distributions. It has an asymptotic breakdown point of $1/2$ for distributions with zero measure on all lower-dimensional hyperplanes. This is the best possible result for affinely equivariant functionals [Theorem 3.2 above, Davies (1987), and Lopuhaä and Rousseeuw (1991)]. We now show that a suitable modification of the Donoho–Stahel functional will lead to a solution of the problem discussed in the previous section.

Let $\mathbb{R}_+ = [0, \infty)$ and $\chi\colon \mathbb{R}_+ \to [0, 1]$ be a strictly increasing twice continuously differentiable function with

$$\chi(0) = 0 \quad \text{and} \quad \chi(\infty) = 1, \tag{6.1}$$

$$\sup_{r \ge 0} r\chi^{(1)}(r) < \infty \tag{6.2}$$

and

$$\sup_{r\geq 0} r^2|\chi^{(2)}(r)| < \infty, \tag{6.3}$$

where $\chi^{(1)}$ and $\chi^{(2)}$ denote, respectively, the first and second derivatives of χ.

For any $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ and for any $\theta \in S^k$, we define the function $\gamma(\mathbb{Q}_d, \theta)$ by

$$\int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}_d, \theta)}\right) d\mathbb{Q}_d = \frac{1 - \Delta(\mathbb{Q}_d)}{2}.$$

As

$$\lim_{s\uparrow\infty} \int \chi\left(\frac{|\theta^T x|}{s}\right) d\mathbb{Q}_d = 0$$

and

$$\lim_{s\downarrow 0} \int \chi\left(\frac{|\theta^T x|}{s}\right) d\mathbb{Q}_d(x) = \mathbb{Q}_d(\{x: x^T\theta \neq 0\}) \geq 1 - \Delta(\mathbb{Q}_d),$$

it follows that $\gamma(\mathbb{Q}_d, \theta)$ is uniquely defined and strictly positive for all $\mathbb{Q}_d$ and θ.

If we metricize $\mathfrak{W}_d(\mathbb{R}^k)$ with the metric $d_{\mathfrak{H}}$, we have the following result.

Lemma 6.1. *For all* $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ *and* η, $0 < \eta < (1 - \Delta(\mathbb{Q}_d))/3$, *there exists a constant* $c_{17} = c_{17}(\mathbb{Q}_d, \eta) > 0$ *and a continuous nondecreasing function* h *depending on* $\mathbb{Q}_d$ *with* $h(0) = 0$ *such that*

$$|\gamma(\theta', \mathbb{Q}'_d) - \gamma(\theta, \mathbb{Q}_d)| < c_{17}(h(\|\theta' - \theta\|) + d_{\mathfrak{H}}(\mathbb{Q}'_d, \mathbb{Q}_d)),$$

for all θ' *and* θ *in* S^k *and all* $\mathbb{Q}'_d$ *with* $d_{\mathfrak{H}}(\mathbb{Q}'_d, \mathbb{Q}_d) < \eta$.

Proof. It follows from dominated convergence and the continuity of χ that $\gamma(\cdot, \mathbb{Q}_d)$: $S^k \to [0, \infty]$ is continuous and that

$$0 < \inf_\theta \gamma(\theta, \mathbb{Q}_d) \leq \sup_\theta \gamma(\theta, \mathbb{Q}_d) < \infty. \tag{6.4}$$

On differentiating $\int \chi(|\theta^T x|/s)\, d\mathbb{Q}_d(x)$ twice with respect to s and using (6.2) and (6.3), we obtain

$$\left|\int \chi\left(\frac{|\theta^T x|}{s}\exp(-\alpha)\right) d\mathbb{Q}_d(x) - \int \chi\left(\frac{|\theta^T x|}{s}\right) d\mathbb{Q}_d(x) - \xi(s, \theta, \mathbb{Q}_d)\alpha\right| = O(\alpha^2), \tag{6.5}$$

where

$$0 < \inf \xi(s, \theta, \mathbb{Q}_d) \leq \sup \xi(s, \theta, \mathbb{Q}_d) < \infty, \tag{6.6}$$

the infimum and supremum being taken over all $\theta \in S^k$ and all s bounded away from 0 and ∞. The term $O(\alpha^2)$ is uniform in θ and s for such values of θ

and s. As in the proof of Lemma 4.1, we have

$$\int \chi\left(\frac{|\theta^T x|}{s}\exp(-d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\right)d\mathbb{Q}_d(x) \le \int \chi\left(\frac{|\theta^T x|}{s}\right)d\mathbb{Q}'_d(x) + d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d),$$

which implies

$$\begin{aligned}\int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}'_d,\theta)}\exp(-d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\right)d\mathbb{Q}_d(x)\\ &\le (1-\Delta(\mathbb{Q}'_d))/2 + d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)\\ &\le (1-\Delta(\mathbb{Q}_d))/2 + 3d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)/2.\\ &< 1-\Delta(\mathbb{Q}_d).\end{aligned}$$

From this, (6.4), (6.5) and (6.6) it follows that, for some $c_{18}, c_{19} > 0$,

$$\begin{aligned}\gamma(\mathbb{Q}'_d,\theta) &\ge \gamma(\mathbb{Q}_d,\theta)\exp(-c_{18}d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\\ &\ge \gamma(\mathbb{Q}_d,\theta) - c_{19}d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)\end{aligned} \tag{6.7}$$

uniformly in θ and $\mathbb{Q}'_d$, $d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d) < \eta < (1-\Delta(\mathbb{Q}_d))/3$.

In the opposite direction we choose $c_{20} = c_{20}(\mathbb{Q}_d,\eta) < \infty$ so that

$$\begin{aligned}\int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}_d,\theta)}\exp(-c_{20}d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\right)d\mathbb{Q}_d(x)\\ \le (1-\Delta(\mathbb{Q}_d))/2 - 3d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)/2,\end{aligned}$$

for all $\theta \in S^k$ and $\mathbb{Q}'_d$ with $d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d) < \eta < (1-\Delta(\mathbb{Q}_d))/3$. This gives

$$\begin{aligned}\int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}_d,\theta)}\exp(-(c_{20}+1)d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\right)d\mathbb{Q}'_d(x)\\ &\le \int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}_d,\theta)}\exp(-c_{20}d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d))\right)d\mathbb{Q}_d(x) + d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)\\ &\le (1-\Delta(\mathbb{Q}_d))/2 - d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)/2 \le (1-\Delta(\mathbb{Q}'_d))/2,\end{aligned}$$

which implies

$$\gamma(\mathbb{Q}'_d,\theta) \ge \gamma(\mathbb{Q}_d,\theta)\exp(-(c_{20}+1)d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d)). \tag{6.8}$$

This and (6.7) yield

$$|\gamma(\mathbb{Q}'_d,\theta) - \gamma(\mathbb{Q}_d,\theta)| \le c_{21}d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d) \tag{6.9}$$

uniformly in $\theta \in S^d$ and $\mathbb{Q}'_d$, $d_{\mathfrak{H}}(\mathbb{Q}_d,\mathbb{Q}'_d) < \eta < (1-\Delta(\mathbb{Q}_d))/3$.

Suppose $s_i = s_i(\mathbb{Q}_d)$, $i = 1, 2$, are such that $s_1 \le \gamma(\mathbb{Q}_d,\theta) \le s_2$ for all θ. We define

$$h(u) = \sup_{\theta',\theta,s}\left|\int\left(\chi\left(\frac{\theta'^T x}{s}\right) - \chi\left(\frac{\theta^T x}{s}\right)\right)d\mathbb{Q}_d\right|,$$

where the supremum is taken over all θ', θ and s with $\|\theta'-\theta\| \le u$ and

$s_1 \le s \le s_2$. Dominated convergence shows that $\int \chi(\theta^T x/s)\, d\mathbb{Q}$ is, as a function of θ and s, continuous on $S^k \times [s_1, s_2]$. It is therefore uniformly continuous on $S^k \times [s_1, s_2]$ from which it follows that h is continuous, non-decreasing with $h(0) = 0$. It follows from (6.5) that

$$|\gamma(\mathbb{Q}_d, \theta') - \gamma(\mathbb{Q}_d, \theta)| \le c_{22} h(\|\theta' - \theta''\|),$$

for some c_{22} and all θ', θ in S^k. On combining this with (6.9) we obtain the claim of the lemma. □

For $x \in \mathbb{R}^k$ we define the outlyingness $O(\mathbb{Q}_d, x)$ of x by

$$O(\mathbb{Q}_d, x) = \sup_{\theta \in S^k} |\theta^T x| / \gamma(\mathbb{Q}_d, \theta). \tag{6.10}$$

We note that

$$\|x\| / \sup_\theta \gamma(\mathbb{Q}_d, \theta) \le O(\mathbb{Q}_d, x) \le \|x\| / \inf_\theta \gamma(\mathbb{Q}_d, \theta). \tag{6.11}$$

The Donoho–Stahel estimator is now defined as a weighted average of the matrices xx^T with x in the support of $\mathbb{Q}_d$. To do this, we introduce a weight function w: $\mathbb{R}_+ \to (0, 1]$ which satisfies the following conditions:

W1: w: $\mathbb{R}_+ \to [0, 1]$ is continuous, strictly decreasing on its support and $w(0) = 1$.

W2: $w(u) \le 1/u^2$ for all u.

For a function w which satisfies W1 and W2, the Donoho–Stahel dispersion functional T^D_{DS} is defined by

$$T^D_{\mathrm{DS}}(\mathbb{Q}_d) = 2 \int w(\zeta(\mathbb{Q}_d) O(\mathbb{Q}_d, x)) xx^T\, d\mathbb{Q}_d / (1 + \Delta(\mathbb{Q}_d)), \tag{6.12}$$

where $\zeta(\mathbb{Q}_d)$ is determined by

$$\int w(\zeta(\mathbb{Q}_d) O(\mathbb{Q}_d, x))\, d\mathbb{Q}_d = (1 + \Delta(\mathbb{Q}_d))/2. \tag{6.13}$$

Condition W1 shows that $\zeta(\mathbb{Q}_d)$ is uniquely defined. From (6.4) and W2 it follows that $w(O(\mathbb{Q}_d, x)) \le c_{23}/\|x\|^2$ for some constant c_{23} so that T^D_{SD} is uniquely defined for all $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ and we have the following theorem.

THEOREM 6.1. *T^D_{DS}: $\mathfrak{W}_d(\mathbb{R}^k) \to PDS(k)$ is well defined and*

$$T^D_{\mathrm{DS}}(\mathbb{Q}_d^L) = L T^D_{\mathrm{DS}}(\mathbb{Q}_d) L^T, \tag{6.14}$$

for all nonsingular linear transformations L: $\mathbb{R}^k \to \mathbb{R}^k$.

PROOF. We prove only (6.14). The transformation formula for integrals gives

$$\frac{1 - \Delta(\mathbb{Q}_d^L)}{2} = \int \chi\left(\frac{|\theta^T x|}{\gamma(\mathbb{Q}_d^L, \theta)}\right) d\mathbb{Q}_d^L = \int \chi\left(\frac{|(L^T\theta)^T x| / \|L^T\theta\|}{\gamma(\mathbb{Q}_d^L, \theta)/\|L^T\theta\|}\right) d\mathbb{Q}_d$$

and as $1 - \Delta(\mathbb{Q}_d) = 1 - \Delta(\mathbb{Q}_d^L)$ it follows that $\gamma(\mathbb{Q}_d, L^T\theta/\|L^T\theta\|) = \gamma(\mathbb{Q}_d^L, \theta)/\|L^T\theta\|$. This gives $O(\mathbb{Q}_d^L, x) = O(\mathbb{Q}_d, L^{-1}x)$, where we have used the fact that L is nonsingular. Using the transformation formula again, we obtain $(1 + \Delta(\mathbb{Q}_d^L))/2 = \int w(\zeta(\mathbb{Q}_d^L)O(\mathbb{Q}_d, x))\, d\mathbb{Q}_d$ which together implies $\zeta(\mathbb{Q}_d^L) = \zeta(\mathbb{Q}_d)$. One final application of the transformation formula gives

$$T_{\mathrm{DS}}^D(\mathbb{Q}_d^L) = 2\int w\big(\zeta(\mathbb{Q}_d^L)O(\mathbb{Q}_d^L, x)\big)xx^T\, d\mathbb{Q}_d^L/\big(1 + \Delta(\mathbb{Q}_d^L)\big) = LT_{\mathrm{DS}}^D(\mathbb{Q}_d^L)L^T,$$

as was to be shown. □

In order to study the breakdown behaviour of T_{DS}^D, we introduce the metric $d_{\mathfrak{Q}}$ defined by

$$d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}'_d) = \sup\{\eta > 0: \mathbb{Q}_d(C) \le \mathbb{Q}'_d(\exp(\eta)C) + \eta, \quad \mathbb{Q}'_d(C) \le \mathbb{Q}_d(\exp(\eta)C) + \eta \text{ for all } C \in \mathfrak{C}\}, \tag{6.15}$$

where

$$\mathfrak{C} = \big\{C: C = \{x: |\theta_i^T x| \le c_i, 1 \le i \le k\} \text{ for some } \theta_i \in S^k \text{ and } c_i \ge 0, 1 \le i \le k\big\}. \tag{6.16}$$

We note that $\mathfrak{C}$ is a Vapnik–Cervonenkis class and that the following theorem holds.

THEOREM 6.2.

(i) $d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}'_d) \le d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}'_d)$.
(ii) $\sup_{Q''_d} d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}'_d) = \eta$, *where* $\mathbb{Q}'_d = (1 - \eta)\mathbb{Q}_d + \eta\mathbb{Q}''_d$.
(iii) *If* $h: \mathbb{R}_+ \to [0, 1]$ *is nondecreasing, then*

$$\int h\Big(\max_{1 \le j \le k} |\theta_j^T x|\Big)\, d\mathbb{Q}_d \le \int h\Big(\max_{1 \le j \le k} |\theta_j^T x| \exp(d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}'_d))\Big)\, d\mathbb{Q}'_d + d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}'_d).$$

PROOF. The proof follows the lines of the corresponding statements for the metric $d_{\mathfrak{H}}$. □

We require the following lemma.

LEMMA 6.2. *If* $\eta < (1 - \Delta(\mathbb{Q}_d))/3$, *then* $0 < \inf \zeta(\mathbb{Q}_{\eta d}) \le \sup \zeta(\mathbb{Q}_{\eta d}) < \infty$, *where the infimum and supremum are taken over all* $\mathbb{Q}_{\eta d}$ *with* $d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}_{\eta d}) < \eta$.

PROOF. Inequalities (6.4), (6.7) and (6.8) give $0 < \inf \gamma(\mathbb{Q}_{\eta d}, \theta) \le \sup \gamma(\mathbb{Q}_{\eta d}, \theta) < \infty$, where the infimum and supremum are taken over all $\theta \in S^k$ and $\mathbb{Q}_{\eta d}$ with $d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}_{\eta d}) < \eta$. This together with W2 and (6.11) implies

$$\begin{aligned}(1 + \Delta(\mathbb{Q}_{\eta d}))/2 &= \int w\big(\zeta(\mathbb{Q}_{\eta d})O(\mathbb{Q}_{\eta d}, x)\big)\, d\mathbb{Q}_{\eta d}\\ &\le \int \min\Big\{1, c_{24}/\big(\zeta(\mathbb{Q}_{\eta d})\|x\|\big)^2\Big\}\, d\mathbb{Q}_{\eta d}\\ &\le \int \min\Big\{1, c_{24}/\big(\zeta(\mathbb{Q}_{\eta d})|\theta^T x|\big)^2\Big\}\, d\mathbb{Q}_{\eta d},\end{aligned}$$

for all $\theta \in S^k$. From this we deduce, with Theorem 6.2,

$$(1 + \Delta(\mathbb{Q}_d))/2 - \eta/2 \le \int \min\Big\{1, c_{25}/\big(\zeta(\mathbb{Q}_{\eta d})|\theta^T x| \exp(\eta)\big)^2\Big\}\, d\mathbb{Q}_d + \eta$$

and hence

$$\Delta(\mathbb{Q}_d) + \alpha(\eta) < \int \min\Big\{1, c_{25}/\big(\zeta(\mathbb{Q}_{\eta d})|\theta^T x| \exp(\eta)\big)^2\Big\}\, d\mathbb{Q}_d$$

with $\alpha(\eta) > 0$. This is only possible if the $\zeta(\mathbb{Q}_{\eta d})$ are bounded away from ∞.

In the other direction we argue as follows. From the definition of $\zeta(\mathbb{Q}_{\eta d})$,

$$\begin{aligned}(1 + \Delta(\mathbb{Q}_d))/2 + \eta/2 \ge (1 + \Delta(\mathbb{Q}_{\eta d}))/2 &= \int w\big(\zeta(\mathbb{Q}_{\eta d})O(\mathbb{Q}_{\eta d}, x)\big)\, d\mathbb{Q}_{\eta d}\\ &\ge \int w\big(c_{26}\zeta(\mathbb{Q}_{\eta d})\|x\|\big)\, d\mathbb{Q}_{\eta d},\end{aligned}$$

where we have used W2 and (6.11). This yields

$$(1 + \Delta(\mathbb{Q}_d))/2 + \eta/2 \ge \int w\Big(c_{27}\zeta(\mathbb{Q}_{\eta d}) \max_j |\theta_j^T x|\Big)\, d\mathbb{Q}_{\eta d},$$

where the θ_j's are orthogonal elements in S^k. Theorem 6.2 implies

$$(1 + \Delta(\mathbb{Q}_d))/2 + \eta/2 \ge \int w\Big(c_{28}\zeta(\mathbb{Q}_{\eta d}) \max_j |\theta_j^T x| \exp(\eta)\Big)\, d\mathbb{Q}_d - \eta$$

and hence

$$\int w\Big(c_{28}\zeta(\mathbb{Q}_{\eta d}) \max_j |\theta_j^T x| \exp(\eta)\Big)\, d\mathbb{Q}_d \le 1 - \alpha(\eta),$$

for some $\alpha(\eta) > 0$. This is only possible if the $\zeta(\mathbb{Q}_{\eta d})$ are bounded away from 0, proving the lemma. □

THEOREM 6.3. *For all* $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ *we have* $\varepsilon^*(T_{\mathrm{DS}}^D, \mathbb{Q}_d, d_{\mathfrak{Q}}) = (1 - \Delta(\mathbb{Q}_d))/3$.

PROOF. Suppose $d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}_{\eta d}) < \eta < (1 - \Delta(\mathbb{Q}_d))/3$. Then, for any $\theta \in S^k$,

$$\theta^T T_{\mathrm{DS}}^D(\mathbb{Q}_{\eta d})\theta = \int w\big(\zeta(\mathbb{Q}_{\eta d})O(\mathbb{Q}_{\eta d}, x)\big)|\theta^T x|^2\, d\mathbb{Q}_{\eta d} \le \int c_{29}\|x\|^{-2}\|x\|^2\, d\mathbb{Q}_{\eta d} \le c_{30}$$

by (6.4), (6.7), (6.8), (6.11), W2 and Lemma 6.2 Thus

(6.17) $$\sup_{Q_{\eta d}} \lambda_{\max}\big(T^D_{\mathrm{DS}}(\mathbb{Q}_{\eta d})\big) < \infty.$$

In the opposite direction we have

$$\begin{aligned}\theta^T T^D_{\mathrm{DS}}(\mathbb{Q}_{\eta d})\theta &= \int w\big(\zeta(\mathbb{Q}_{\eta d})O(\mathbb{Q}_{\eta d}, x)\big)|\theta^T x|^2\, d\mathbb{Q}_{\eta d}\\ &\geq r^2 \int_{|\theta^T x| \geq r} w\big(\zeta(\mathbb{Q}_{\eta d})O(\mathbb{Q}_{\eta d}, x)\big)\, d\mathbb{Q}_{\eta d}\\ &\geq r^2\big((1 + \delta(\mathbb{Q}_{\eta d}))/2 - \mathbb{Q}_{\eta d}(\{x\colon |\theta^T x| < r\})\big)\\ &\geq r^2\big((1 + \delta(\mathbb{Q}_d))/2 - \mathbb{Q}_d(\{x\colon |\theta^T x| < re^{\eta}\}) - 3\eta/2\big).\end{aligned}$$

As r tends to 0 the expression $(1 + \Delta(\mathbb{Q}_d))/2 - \mathbb{Q}_d(\{x\colon |\theta^T x| < re^{\eta}\}) - 3\eta/2$ tends to a value of at least $(1 - \Delta(\mathbb{Q}_d))/2 - 3\eta/2 > 0$ and hence $\inf_{Q_{\eta d}} \lambda_{\min}(T^D_{\mathrm{DS}}(\mathbb{Q}_{\eta d})) > 0$ which, in conjunction with (6.17), implies

(6.18) $$\varepsilon^*\big(T^D_{\mathrm{DS}}, \mathbb{Q}_d, d_{\mathfrak{Q}}\big) \geq (1 - \Delta(\mathbb{Q}_d))/3.$$

To demonstrate the opposite inequality, we define a design distribution $\mathbb{Q}_{\eta d}(R, \theta_1)$ as in the proof of Theorem 4.6 by

$$\begin{aligned}\mathbb{Q}_{\eta d}(R, \theta_1)(B) &= \eta\delta_{R\theta_1}(B) + \mathbb{Q}_d\big(B \cap \{x\colon \theta_0^T x = 0\}\big)\\ &\quad + (1 - \eta/(1 - \Delta(\mathbb{Q}_d)))\mathbb{Q}_d\big(B \cap \{x\colon \theta_0^T x \neq 0\}\big),\end{aligned}$$

where $(1 - \Delta(\mathbb{Q}_d))/3 < \eta < 1$, $\Delta(\mathbb{Q}_d) = \mathbb{Q}_d(\{x\colon \theta_0^T x = 0\})$, θ_1 and θ_0 both belong to S^k and $\theta_1^T\theta_0 = 0$. It is seen that

$$\begin{aligned}\mathbb{Q}_{\eta d}(R, \theta_1)(B) - \mathbb{Q}_d(B) &= \eta\delta_{R\theta_1}(B) - \eta/(1 - \Delta(\mathbb{Q}_d))\\ &\quad \times \mathbb{Q}_d\big(B \cap \{x\colon \theta_0^T x \neq 0\}\big),\end{aligned}$$

which implies $\sup_B|\mathbb{Q}_{\eta d}(R, \theta_1)(B) - \mathbb{Q}_d(B)| \leq \eta$ and hence $d_{\mathfrak{Q}}(\mathbb{Q}_d, \mathbb{Q}_{\eta d}(R, \theta_1)) \leq \eta$. Furthermore, we have $\mathbb{Q}_{\eta d}(R, \theta_1)(\{x\colon \theta_0^T x = 0\}) = \eta + \mathbb{Q}_d(\{x\colon \theta_0^T x = 0\}) = \eta + \Delta(\mathbb{Q}_d)$, giving $\delta(\mathbb{Q}_{\eta d}(R, \theta_1)) = \eta + \Delta(\mathbb{Q}_d)$.

From the definition of $\zeta(\mathbb{Q}_{\eta d}(R, \theta_1))$, we obtain

$$\begin{aligned}(1 + \Delta(\mathbb{Q}_d) + \eta)/2 &= \int w\big(\zeta(\mathbb{Q}_{\eta d}(R, \theta_1))O(\mathbb{Q}_{\eta d}(R, \theta_1), x)\big)\, d\mathbb{Q}_{\eta d}(R, \theta_1)(x)\\ &\leq \eta w\big(\zeta(\mathbb{Q}_{\eta d}(R, \theta_1))O(\mathbb{Q}_{\eta d}(R, \theta_1), R\theta_1)\big)\\ &\quad + \Delta(\mathbb{Q}_d) + (1 - \Delta(\mathbb{Q}_d) - \eta),\end{aligned}$$

which implies

$$\begin{aligned}\eta w\big(\zeta(\mathbb{Q}_{\eta d}(R, \theta_1))O(\mathbb{Q}_{\eta d}(R, \theta_1), R\theta_1)\big) &\geq (1 + \Delta(\mathbb{Q}_d) + \eta)/2 - \Delta(\mathbb{Q}_d)\\ &\quad - (1 - \Delta(\mathbb{Q}_d) - \eta)\\ &= 3\eta/2 - (1 - \Delta(\mathbb{Q}_d))/2 \geq \alpha(\eta),\end{aligned}$$

with $\alpha(\eta) > 0$ and independent of R. We deduce

$$\theta_1^T T_{\mathrm{DS}}^D(\mathbb{Q}_{\eta d}(R, \theta_1))\theta_1 \geq \eta w\big(\zeta(\mathbb{Q}_{\eta d}(R, \theta_1))O(\mathbb{Q}_{\eta d}(R, \theta_1), R\theta_1)\big)R^2,$$

which tends to ∞ as $R \to \infty$. Consequently, T_{DS}^D has broken down giving $\varepsilon^*(T_{\mathrm{DS}}^D, \mathbb{Q}_d, d_{\mathfrak{D}}) \leq (1 - \Delta(\mathbb{Q}_d))/3$ which together with (6.18) implies the statement of the theorem. □

6.3. *A Lipschitz dispersion functional.* The functional T_{DS}^D satisfies DP1, DP2 and DP3 of subsection 6.1. If one considers a sufficiently strong metric such as that defined by (6.15) but with $\mathfrak{C}$ replaced by the class of all symmetric convex sets with centre at the origin, then T_{DS}^D can be shown to satisfy a local Lipschitz condition of order 1 at each point $\mathbb{Q}$ of $\mathfrak{W}_d(\mathbb{R}^k)$. This class is not, however, a Vapnik–Cervonenkis class and the resulting supremum metric is strong in that empirical measures will not in general converge at the rate of $n^{-1/2}$ [Bolthausen (1978)]. We now show how the Donoho–Stahel functional may be altered to obtain a dispersion functional which satisfies DP1, DP2, DP3 and DP4′.

For each $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ we define

$$\mathfrak{D}(\mathbb{Q}_d) = \{x\colon O(\mathbb{Q}_d, x) \leq 1\}. \tag{6.19}$$

It is easily checked that $\mathfrak{D}(\mathbb{Q}_d)$ is convex and bounded for each $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$. From (6.4) we see that $\mathfrak{D}(\mathbb{Q}_d)$ has a nonempty interior and hence has positive k-dimensional Lebesgue measure. We define

$$T_{\mathfrak{D}}^D(\mathbb{Q}_d) = \int_{\mathfrak{D}(Q_d)} xx^T\, dm_k(x)/m_k(\mathfrak{D}(\mathbb{Q}_d)), \tag{6.20}$$

where m_k denotes k-dimensional Lebesgue measure.

THEOREM 6.4. *$T_{\mathfrak{D}}^D$: $\mathfrak{W}_d(\mathbb{R}^k) \to PDS(k)$ is well defined, linearly equivariant and has breakdown point $\varepsilon^*(T_{\mathfrak{D}}^D, \mathbb{Q}_d, d_{\mathfrak{H}}) = (1 - \Delta(\mathbb{Q}_d))/3$. Furthermore, for each $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ there exists an η, $0 < \eta < (1 - \Delta(\mathbb{Q}_d))/3$ such that $\|T_{\mathfrak{D}}^D(\mathbb{Q}_d) - T_{\mathfrak{D}}^D(\mathbb{Q}_d')\| = O(d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}_d'))$ uniformly for all $\mathbb{Q}_d'$ satisfying $d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}_d') < \eta$.*

PROOF. The first three claims of the theorem are easily checked. The fact that $\varepsilon^*(T_{\mathfrak{D}}^D, \mathbb{Q}_d, d_{\mathfrak{H}}) \leq (1 - \Delta(\mathbb{Q}_d))/3$ follows from (6.7) and (6.8) in the proof of Lemma 6.1. In the opposite direction we suppose $(1 - \Delta(\mathbb{Q}_d))/3 < \eta < 1 - \Delta(\mathbb{Q}_d)$ and consider the design distribution $\mathbb{Q}_{\eta d}(R, \theta_1)$ as in the proof of Theorem 6.3. We have

$$\int \chi\left(\frac{|x^T\theta|}{s}\right) d\mathbb{Q}_{\eta d}(R, \theta_1) \geq \eta\chi\left(\frac{R|\theta_1^T\theta|}{s}\right),$$

which together with $\Delta(\mathbb{Q}_{\eta d}(R, \theta_1)) = \Delta(\mathbb{Q}_d) + \eta$ implies that

$$\lim_{R\to\infty} \gamma\big(\mathbb{Q}_{\eta d}(R, \theta_1), \theta\big) = \infty \tag{6.21}$$

uniformly in all θ satisfying $|\theta_1^T\theta| \ge \alpha > 0$ for each $\alpha > 0$. On the other hand, we have

$$\int \chi\left(\frac{|x^T\theta|}{s}\right) d\mathbb{Q}_{\eta d}(R, \theta_1) \le \int \chi\left(\frac{|x^T\theta|}{s}\right) d\mathbb{Q}_d - \eta,$$

the right-hand side of which tends to at least $1 - \Delta(\mathbb{Q}_d) - \eta > (1 - \Delta(\mathbb{Q}_d) - \eta)/2$ as s tends to 0. It therefore follows that

$$\liminf_{R\to\infty} \inf_{\theta \in S^k} \gamma\big(\mathbb{Q}_{\eta d}(R, \theta_1), \theta\big) > 0$$

which together with (6.21) implies

$$\{x\colon |x^T\theta| \le c \text{ for all } \theta \in S^k \cap \theta_1^{\perp}\} \subset \limsup_{R\to\infty} \mathfrak{D}\big(\mathbb{Q}_{\eta d}(R, \theta_1)\big), \tag{6.22}$$

for some $c > 0$. We may deduce

$$\begin{aligned}
&\operatorname{tr}\Big(T_{\mathfrak{D}}^D\big(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1))\big)\Big)\\
&\quad = \int_{\mathfrak{D}(Q_{\eta d}(R, \theta_1))} \|x\|^2 \, dm_k(x)/m_k\big(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1))\big)\\
&\quad \ge r^2\Big(1 - m_k\big(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1) \cap \{x\colon \|x\| \le r\})\big)\Big)/m_k\big(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1))\big)\\
&\quad \ge r^2\Big(1 - c_{31} r^k/m_k\big(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1))\big)\Big)
\end{aligned}$$

and as, by (6.22), $\lim_{R\to\infty} m_k(\mathfrak{D}(\mathbb{Q}_{\eta d}(R, \theta_1))) = \infty$, we obtain

$$\limsup_{R\to\infty} \operatorname{tr}\Big(T_{\mathfrak{D}}^D\big(L(\mathbb{Q}_{\eta d}(R, \theta_1))\big)\Big) \ge r^2$$

for any $r > 0$. This implies that $T_{\mathfrak{D}}^D$ has broken down at $\mathbb{Q}_d$ and as this holds for any η satisfying $(1 - \Delta(\mathbb{Q}_d))/3 < \eta < 1 - \Delta(\mathbb{Q}_d)$ we may deduce $\varepsilon^*(T_{\mathfrak{D}}^D, \mathbb{Q}_d, d_{\mathfrak{H}}) = (1 - \Delta(\mathbb{Q}_d))/3$.

It remains to demonstrate that $T_{\mathfrak{D}}^D$ satisfies a local Lipschitz condition of order 1. From (6.6), (6.7) and (6.8) it follows that for all $\mathbb{Q}'_d$ with $d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}'_d)$ sufficiently small $(1 - c_{32} d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}'_d))\mathfrak{D}(\mathbb{Q}_d) \subset \mathfrak{D}(\mathbb{Q}'_d) \subset (1 + c_{32} d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}'_d))\mathfrak{D}(\mathbb{Q}_d)$ for some constant $c_{32} > 0$. On using $m_k(\tau B) = \tau^k m_k(B)$ for any Borel set B and any $\tau \ge 0$, we obtain $|m_k(\mathfrak{D}(\mathbb{Q}_d)) - m_k(\mathfrak{D}(\mathbb{Q}'_d))| = O(d_{\mathfrak{H}}(\mathbb{Q}_d, \mathbb{Q}'_d))$ from which the claim follows by elementary calculations. □

7. Efficiency, breakdown point and Lipschitz continuity.

7.1. *Efficiency and breakdown point.* It follows from the asymptotics of the Hampel–Rousseeuw least median of squares estimator [Kim and Pollard (1990), Davies (1990), and Rousseeuw (1984)] that it has zero asymptotic efficiency when the errors are normally distributed. This has led to a search for estimators which combine efficiency with a high breakdown point. Such estimators have been proposed by Jurečková and Portnoy (1987), Yohai (1987) and Yohai and Zamar (1988). However, the authors restrict attention to independently and identically distributed carriers with a finite second moment.

This effectively excludes arbitrarily large leverage points. If, however, arbitrarily large leverage points are allowed, Morgenthaler (1989) and Stefanski (1991) have claimed that all high breakdown regression functionals can have an arbitrarily small efficiency compared to the least squares estimator. The basic idea is the following. A high breakdown estimator must by its very nature ignore leverage points although it is such points which determine the efficiency of the least squares estimator. The argument given by Stefanski (1991) is restricted to regression functionals with a global breakdown point, that is, a breakdown point which is independent of the underlying distribution. It follows, however, from (1.1) for the finite sample breakdown point and from Theorem 3.1 for breakdown points in terms of metrics that this is only possible if the global breakdown point is respectively $1/n$ or 0. We are not able to give a general theorem on the relationship between an arbitrarily high breakdown estimator and its efficiency relative to that of the least squares estimator. For one thing it is not clear what exactly is meant by a high breakdown estimator. However, we can show that for empirical distributions with normal errors there exists a high breakdown estimator with an arbitrarily high efficiency compared with least squares. Our definition of efficiency is based on coverage probabilities. In subsection 7.2 we show that the idea of Morgenthaler and Stefanski can be made precise for locally uniformly Lipschitz continuous regression functionals.

We show that if the concept of efficiency is based on covering probabilities and not on variance, then it is possible to have high breakdown regression estimates with arbitrarily high efficiency independently of the design distribution. We consider a distribution $\mathbb{Q}$ of the form $\mathbb{Q} = \mathbb{Q}_d * \mathbb{Q}_e$, where $\mathbb{Q}_d$ is a design distribution with finite support, that is,

$$\mathbb{Q}_d = \frac{1}{n}\sum_1^n \delta_{x_j}. \tag{7.1}$$

An empirical version of $\mathbb{Q}$ will be denoted by $\hat{\mathbb{Q}}_n$ and is of the form

$$\hat{\mathbb{Q}}_n = \frac{1}{n}\sum_1^n \delta_{(x_j^T, y_j)^T},$$

where $y_j = x_j^T\beta + e_j$ and the $(e_j)_1^n$ are independently and identically distributed random variables with common distribution $\mathbb{Q}_e$ belonging to a scale family.

Let T_i^L, $i = 1, 2$, be the location part of linearly equivariant regression functionals and let α, $0 < \alpha < 1$, be fixed. Then, for any $\theta \in S^k$, we define $\Lambda_i(\alpha, \theta)$ by

$$\mathbb{P}\Big(\big|\theta^T\big(T_i^L\big(\hat{\mathbb{Q}}_n\big) - \beta\big)\big| \le \Lambda_i(\alpha, \theta)\Big) = \alpha.$$

The relative efficiency $\mathrm{eff}(T_1^L, T_2^L\colon \alpha)$ of T_1^L with respect to T_2^L is defined by

$$\mathrm{eff}\big(T_1^L, T_2^L\colon \alpha\big) = \inf_\theta \frac{\Lambda_2(\alpha, \theta)^2}{\Lambda_1(\alpha, \theta)^2}.$$

We now specialize to the case where $\mathbb{Q}_e$ is an $\mathfrak{N}(0, \sigma^2)$ distribution and T_2 is the least squares estimator T_{LS}. For any given α and η, $0 < \eta < 1$, we now construct a high breakdown estimator T_M such that $\mathrm{eff}(T_M^L, T_{\mathrm{LS}}^L, \alpha) \geq 1 - \eta$ uniformly over all $\mathbb{Q}_d$ of the form (7.1) with $\Delta(\mathbb{Q}_d) = k/n$. To do this, we consider the Hampel–Rousseeuw least median of squares estimator T_{LMS}. We denote the location part by $\hat{\beta}_{\mathrm{LMS}}$ and the scale part by $\hat{\sigma}_{\mathrm{LMS}}$. We require the following lemma.

LEMMA 7.1. *For any ρ, $0 < \rho < 1$, there exists a $\zeta > 0$ such that*

$$\mathbb{P}(\hat{\sigma}_{\mathrm{LMS}} \leq \zeta\sigma) \leq \rho$$

uniformly for all $\mathbb{Q}_d$ of the form (7.1) *with $\Delta(\mathbb{Q}_d) = k/n$ and $n \geq 2k + 2$.*

PROOF. It is known that there exist $k + 1$ design points $x_1^*, \dots, x_{k+1}^*$ and an appropriate choice of $\pm$ signs such that

$$y_j^* - x_j^{*T}\hat{\beta}_{\mathrm{LMS}} = \pm\hat{\sigma}_{\mathrm{LMS}}, \tag{7.2}$$

for $j = 1, \dots, k + 1$ [Steele and Steiger (1986)]. Furthermore, the solution $(\hat{\beta}_{\mathrm{LMS}}^T, \hat{\sigma}_{\mathrm{LMS}})^T$ of the set of equations (7.2) is unique. Consider now all equations of the form (7.2) for all choices of $k + 1$ design points and all choices of $\pm$ signs such that the solution is uniquely defined. If the design points are $x_{i_1}, \dots, x_{i_{k+1}}$, we denote the solution by $\hat{\beta}(i_1, \dots, i_{k+1})$. Then $\hat{\beta} = \hat{\beta}(i_1, \dots, i_{k+1})$ is a function of the errors $e_{i_1}, \dots, e_{i_{k+1}}$ only and hence independent of the remaining $n - k - 1$ errors. For any given design points $x_{l_1}, \dots, x_{l_m}$ with $m = [n/2] - k$ and $l_j \notin \{i_1, \dots, i_{j+1}\}$, $j = 1, \dots, m$, we have

$$\begin{aligned}
&\mathbb{P}\Big(|y_{l_j} - x_{l_j}^T\hat{\beta}| \leq \zeta\sigma,\ j = 1, \dots, l_m\Big)\\
&\quad\leq \mathbb{P}\Big(|e_{l_j} - x_{l_j}^T(\hat{\beta} - \beta)| \leq \zeta\sigma,\ j = 1, \dots, l_m\Big)\\
&\quad= \mathbb{E}\Big(\mathbb{P}\Big(|e_{l_j} - x_{l_j}^T(\hat{\beta} - \beta)| \leq \zeta\sigma,\ j = 1, \dots, l_m | e_{i_1}, \dots, e_{i_{k+1}}\Big)\Big)\\
&\quad= \mathbb{E}\left(\prod_{j=1}^m \mathbb{P}\Big(|e_{l_j} - x_{l_j}^T(\hat{\beta} - \beta)| \leq \zeta\sigma | e_{i_1}, \dots, e_{i_{k+1}}\Big)\right)\\
&\quad\leq (\Phi(\zeta) - \Phi(-\zeta))^m.
\end{aligned} \tag{7.3}$$

By the very definition of T_{LMS} there exist m design points apart from $x_1^*, \dots, x_{k+1}^*$ such that $|y - x^T\hat{\beta}_{\mathrm{LMS}}| \leq \hat{\sigma}_{\mathrm{LMS}}$. If now $\hat{\sigma} \leq \zeta\sigma$ it follows that there exist m design points such that $|y - x^T\hat{\beta}_{\mathrm{LMS}}| \leq \zeta\sigma$. On using (7.3) and summing over all possible choices of $x_1^*, \dots, x_{k+1}^*$ and $\pm$ signs, we see

$$\mathbb{P}(\hat{\sigma}_{\mathrm{LMS}} \leq \zeta\sigma) \leq 2^{k+1}\binom{n}{k+1} \cdot (\Phi(\zeta) - \Phi(-\zeta))^m. \tag{7.4}$$

As $m = [n/2] - k$ and $n \geq 2k + 2$ we may, by choosing ζ sufficiently small, arrange that $\mathbb{P}(\hat{\sigma}_{\mathrm{LMS}} \le \zeta\sigma) \le \rho$ for any given ρ and for all $\mathbb{Q}_d$ with $\delta(\mathbb{Q}_d) = k/n$, $n \ge 2k + 2$. This last fact follows from the independence of the inequality (7.4) from $\mathbb{Q}_d$. This proves the lemma. □

We now consider the least squares functionals T_{LMS} with location part $\hat{\beta}_{\mathrm{LS}}$ and scale part $\hat{\sigma}_{\mathrm{LS}}$. Given again any ρ, $0 < \rho < 1$, there exists a K such that

$$\mathbb{P}(\hat{\sigma}_{\mathrm{LS}} \ge K\sigma) \le \rho, \tag{7.5}$$

independently of $\mathbb{Q}_d$. From Lemma 7.1 and from (7.5), we deduce that for any ρ, $0 < \rho < 1$, there exists a γ such that

$$\mathbb{P}(\hat{\sigma}_{\mathrm{LS}} > \gamma\hat{\sigma}_{\mathrm{LMS}}) \le \rho. \tag{7.6}$$

We now define T_M by

$$T_M = \begin{cases} T_{\mathrm{LS}}, & \text{if } \hat{\sigma}_{\mathrm{LS}} \le \gamma\hat{\sigma}_{\mathrm{LMS}}, \\ T_{\mathrm{LMS}}, & \text{otherwise}. \end{cases} \tag{7.7}$$

We see that $\mathbb{P}(T_M = T_{\mathrm{LS}}) \ge 1 - \rho$ and hence, by choosing ρ sufficiently small, we have

$$\mathrm{eff}(T_M^L, T_{\mathrm{LS}}^L, \alpha) \ge 1 - \eta$$

and this holds uniformly for all models $\mathbb{Q}$ of the form described previously.

We now show that if T_{LMS} does not break down neither does T_M. Indeed, suppose T_M breaks down but not T_{LMS}. Then from the definition of T_M it follows that $\|\hat{\beta}_{\mathrm{LS}}\| \to \infty$ but $\hat{\sigma}_{\mathrm{LS}}$ remains bounded. As this is clearly not possible, we see that the breakdown point of T_m is at least that of T_{LMS}. Finally, as both T_{LS} and T_{LMS} are linearly equivariant and the switchover (7.7) is linearly invariant, it follows that T_M is linearly equivariant.

7.2. *Efficiency and Lipschitz continuity.* The example given in subsection 7.1 shows that high breakdown estimators may have an arbitrarily high efficiency with respect to the least squares estimator. The following argument shows that this is not the case for Lipschitz continuous functionals. We restrict ourselves to the case $k = 1$ and consider the regression distribution $\mathbb{Q} = \mathbb{Q}_d * \mathbb{Q}_e$ with $\mathbb{Q}_d = \delta_1$, $\mathbb{Q}_e = \mathfrak{N}(0,1)$ and, without loss of generality, location part $\beta = 0$. We write

$$\hat{\mathbb{Q}}_n = \frac{1}{n}\sum_1^n \delta_{(1, y_j)^T}, \tag{7.8}$$

with $(y_j)_1^n$ independently and identically distributed $\mathfrak{N}(0,1)$ random variables.

We denote by $\mathbb{Q}(r)$ the regression distribution $\mathbb{Q}(r) = \mathbb{Q}_d(r) * \mathbb{Q}_e$ with $\mathbb{Q}_d(r) = \delta_r$ and again with location part $\beta = 0$. We write

$$\hat{\mathbb{Q}}_n' = \frac{n-1}{n}\hat{\mathbb{Q}}_{n-1} + \frac{1}{n}\delta_{(r, y_n)^T}, \tag{7.9}$$

from which we may deduce

$$d_{\mathfrak{H}}\big(\hat{\mathbb{Q}}'_n, \hat{\mathbb{Q}}_{n-1}\big) = \frac{1}{n}. \tag{7.10}$$

It follows from standard empirical process theory that

$$d_{\mathfrak{H}}\big(\mathbb{Q}, \hat{\mathbb{Q}}_n\big) = O_p\bigg(\frac{1}{\sqrt{n}}\bigg). \tag{7.11}$$

Let T be a regression functional which is locally uniformly Lipschitz at $\mathbb{Q}$. By this we mean there exists an $\varepsilon > 0$ and a constant $c_{33} > 0$ such that

$$\|T(\mathbb{Q}') - T(\mathbb{Q}'')\| \le c_{33} d_{\mathfrak{H}}(\mathbb{Q}', \mathbb{Q}''), \tag{7.12}$$

for all $\mathbb{Q}'$ and $\mathbb{Q}''$ with $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}') \le \varepsilon$ and $d_{\mathfrak{H}}(\mathbb{Q}, \mathbb{Q}'') \le \varepsilon$.

We write $T(\hat{\mathbb{Q}}_n) = (\hat{\beta}_n, \hat{\sigma}_n)^T$ and $T(\hat{\mathbb{Q}}'_n) = (\hat{\beta}'_n, \hat{\sigma}'_n)^T$. Given α, $0 < \alpha < 1$, we define $L_n(T, \alpha)$ and $L'_n(T, \alpha)$ by $\mathbb{P}(|\hat{\beta}_n| \le L_n(T, \alpha)) = \alpha$ and $\mathbb{P}(|\hat{\beta}'_n| \le L'_n(T, \alpha)) = \alpha$. We suppose that T is not super efficient at $\mathbb{Q}$ so that

$$L_n(T, \alpha) \ge c_{34}/\sqrt{n}, \tag{7.13}$$

with $c_{34} > 0$.

It follows from (7.10) and (7.11) that $d_{\mathfrak{H}}(\mathbb{Q}, \hat{\mathbb{Q}}_n) = O_p(1/\sqrt{n})$ and $d_{\mathfrak{H}}(\mathbb{Q}, \hat{\mathbb{Q}}'_n) = O_p(1/\sqrt{n})$. Thus, with limiting probability 1, we deduce from (7.10) that

$$\|T\big(\hat{\mathbb{Q}}_n\big) - T\big(\hat{\mathbb{Q}}'_n\big)\| \le c_{35}/n.$$

This together with (7.13) shows that

$$L'_n(T, \alpha) \ge c_{36}/\sqrt{n}, \tag{7.14}$$

with limiting probability 1 independently of r. However, $\lim_{r \to \infty} \hat{\beta}_{\mathrm{LS}}(\hat{\mathbb{Q}}'_n) = 0$ for any fixed m which together with (7.14) shows that T has arbitrarily low efficiency when compared with the least squares estimator.

8. A Lipschitz functional for normal errors. As mentioned several times in this paper, the author has not succeeded in obtaining a global high breakdown regression functional which is locally Lipschitz. However, it is possible to obtain functionals which are Lipschitz at certain distributions. This is done by taking a reasonably smooth functional such as an S-estimator and then forming a k-step M-estimator to increase the smoothness. Such an approach was used in Davies (1992b) to obtain a Fréchet differentiable location and dispersion functional for certain elliptical distributions.

In order to obtain Lipschitz functionals, we require bounds for the difference of integrals of the form $|\int h\, d\mathbb{Q} - \int h\, d\mathbb{Q}'|$ in terms of a distance $d(\mathbb{Q}, \mathbb{Q}')$. We introduce the linearly invariant metric $d_{\mathfrak{P}}$ defined as in (6.15) but with

$$\mathfrak{C} = \big\{C : C = \big\{x : |(x - a_i)^T \theta_i| \le c_i, a_i \in \mathbb{R}^k, \theta_i \in S^k, c_i \in \mathbb{R}_+, 1 \le i \le k\big\}\big\}. \tag{8.1}$$

For any function $h\colon \mathbb{R}^k \to \mathbb{R}$, let $D_{1 \cdots k} h$ denote the partial derivative of order

k with respect to $x_1, \ldots, x_k$. We have, Theorem 3.3 of Davies (1992a), the following result.

THEOREM 8.1. *Let $h: \mathbb{R}^k \to \mathbb{R}$ have continuous partial derivatives of order k satisfying*

$$\text{(i)} \qquad h(x) = a \quad \text{for all } x \text{ with } \|x\| > R$$

and

$$\text{(ii)} \quad \int |D_{1\cdots k}h(x+y) - D_{1\cdots k}h(x)|\, dm_k < K\|y\| \quad \text{for all } y \text{ in } \mathbb{R}^k.$$

Then $|\int h\, d\mathbb{Q} - \int h\, d\mathbb{Q}'| \le c_{37} d_{\mathfrak{B}}(\mathbb{Q}, \mathbb{Q}')$ for some constant $c_{37} > 0$ which depends only on R and K.

The conditions on the function h can be weakened but only at the cost of using a stronger metric. As we can choose the function h, we prefer to impose stronger conditions on it rather than on the metric.

We consider M-estimators for β which are defined by a function ψ which satisfies the following conditions:

ML1: $\psi: \mathbb{R} \to [-1, 1]$ is asymmetric.
ML2: $\psi: \mathbb{R} \to [-1, 1]$ has a continuous $k + 2$ derivative $\psi^{(k+2)}$.

Let $\mathbb{Q} = \mathbb{Q}_d * \mathbb{Q}_e$, where $\mathbb{Q}_e$ has a density function f which satisfies F1–F3 and FR1–FR4 with $\rho(u) = \{|u| \le c_0\}$. This is the case if f has a continuous derivative $f^{(1)}$ at c_0 with $f^{(1)}(c_0) < 0$. For such a $\mathbb{Q}$ the Hampel–Rousseeuw least median of squares estimator T^G_{LMS} is well defined and we have $T^G_{\mathrm{LMS}}(\mathbb{Q}) = (\beta^T, \sigma)^T$.

Let w be a weight function which satisfies W1 and W2 as well as the following:

W3: $w: \mathbb{R}_+ \to [0, 1]$ has a compact support.
W4: $w: \mathbb{R}_+ \to [0, 1]$ has a continuous $k + 1$ derivative $w^{(k+1)}$.

Let $T^D_{\mathfrak{D}}$ be the global Lipschitz dispersion functional of Section 6. For any $\mathbb{Q}'_d$ we define $\xi(\mathbb{Q}'_d)$ by $\int w(\xi(\mathbb{Q}'_d) x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1} x)\, d\mathbb{Q}'_d = (1 + \Delta(\mathbb{Q}'_d))/2$. We consider the M-estimator T_M defined by

$$T_M(\mathbb{Q}') = \left\{ \left(b_M^T, s_M\right)^T : \left(b_M^T, s_M\right)^T \text{ satisfy (8.2) and (8.3) below} \right\},$$

$$\text{(8.2)} \qquad \int x w\left(\xi(\mathbb{Q}'_d) x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1} x\right) \psi\left(\frac{y - x^T b_M}{s_M}\right) d\mathbb{Q}' = 0,$$

$$\text{(8.3)} \qquad s_M = s_{\mathrm{HR}}, \quad \text{the dispersion part of } T^G_{\mathrm{HR}}.$$

The assumptions we have placed on ψ guarantee that T_M is defined in a $d_{\mathfrak{D}}$-neighbourhood of $\mathbb{Q}$. From F2, ML1 and Theorem 4.7, we have $(\beta^T, \sigma)^T \in T_M(\mathbb{Q})$. To define the one-step M-estimator, we require the following condition on ψ and f:

$$\text{FML} \qquad E(\psi^{(1)}) = \int \psi^{(1)}(u) f(u)\, du \neq 0.$$

Starting with the Hampel–Rousseeuw functional T^G_{LMS} with $T^G_{\mathrm{LMS}}(\mathbb{Q}') = (b^T_{\mathrm{LMS}}, s_{\mathrm{LMS}})^T$, we define the one-step M-estimator $T_{\mathrm{OS}}(\mathbb{Q}') = (b^T_{\mathrm{OS}}, s_{\mathrm{OS}})^T$ by

$$(8.4)\qquad \begin{aligned} b_{\mathrm{OS}} &= b_{\mathrm{LMS}} + s_{\mathrm{LMS}}\Sigma(\mathbb{Q}')^{-1} \\ &\quad \times \int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T b_{\mathrm{HR}}}{s_{\mathrm{HR}}}\right) d\mathbb{Q}', \\ s_{\mathrm{OS}} &= s_{\mathrm{LMS}}, \end{aligned}$$

where

$$\Sigma(\mathbb{Q}) = E(\psi^{(1)}) \int xx^T w\big(\lambda(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\, d\mathbb{Q}',$$

with $E(\psi^{(1)})$ as in FML. This is just one of many possibilities. In particular, one could also perform a one-step M-estimator for the scale parameter σ. We have not done this as we are mainly concerned with smoothness and the scale part of the Hampel–Rousseeuw functional is already Lipschitz.

In order to study the breakdown and smoothness properties of T_{OS}, we require the following lemma.

LEMMA 8.1. *If* $\eta < (1 - \Delta(\mathbb{Q}))/3$, *then*

(i) $$0 < \inf \xi(\mathbb{Q}'_\eta) \le \sup \xi(\mathbb{Q}'_\eta) < \infty,$$

(ii) $$0 < \inf \lambda_{\min}\big(\Sigma(\mathbb{Q}'_\eta)\big) \le \sup \lambda_{\max}\big(\Sigma(\mathbb{Q}'_\eta)\big) < \infty,$$

where in both cases the infimum and supremum are taken over all $\mathbb{Q}'_\eta$ *satisfying* $d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}') < \eta$.

PROOF. The proof of (i) follows the lines of the proof of Lemma 6.1. The proof of (ii) follows the lines of the proof of Theorem 6.3. The only difference is that in each case the function $O(\mathbb{Q}_{\eta d}, x)$ is to be replaced by $x^T T^D_{\mathfrak{D}}(\mathbb{Q}_{\eta d})^{-1}x$. □

THEOREM 8.2. T_{OS} *is defined for all* $\mathbb{Q}'$ *satisfying* $d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}') < (1 - \Delta(\mathbb{Q}))/3$ *and has breakdown point* $\varepsilon^*(T_{\mathrm{OS}}, \mathbb{Q}, d_{\mathfrak{D}}) = (1 - \Delta(\mathbb{Q}))/3$.

PROOF. This follows from Lemma 8.1 and Theorem 4.6. □

The smoothness of T_{OS} at $\mathbb{Q}$ is covered by the following result.

THEOREM 8.3. *For all* η, $0 < \eta < (1 - \Delta(\mathbb{Q}))/3$, *there exists a constant* c *such that* $\|b' - b\| + |s' - s| < c d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}')$ *for all* $\mathbb{Q}'$ *with* $d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}') < \eta$ *and all* $(b'^T, s')^T \in T_{\mathrm{OS}}(\mathbb{Q}')$.

PROOF. A Taylor expansion of (8.4) yields

$$b_{\mathrm{OS}} = b_{\mathrm{LMS}} + s_{\mathrm{LMS}}\Sigma(\mathbb{Q}')^{-1}\int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'$$
$$-\left(\Sigma(\mathbb{Q}')^{-1}\int xx^T w\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi^{(1)}\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'\right)$$
$$\times(b_{\mathrm{LMS}} - \beta) + O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}')),$$

where we have used Theorem 8.2, the fact that ψ and w both have a compact support and Theorem 4.7 to bound the quadratic term. As ψ and w have derivatives of order $k + 2$ and $k + 1$, respectively, we may deduce from Theorem 8.1 that

$$\left|\int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'\right.$$
$$\left.-\int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}\right|$$
$$= O(d_{\mathfrak{D}}(\mathbb{Q}, \mathbb{Q}'))$$

and as

$$\int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q} = 0$$

because of the symmetry of f and the asymmetry of ψ, we obtain

$$\left|\int xw\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'\right| = O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}'))$$

to give

$$b_{\mathrm{OS}} = b_{\mathrm{LMS}}$$
$$-\left(\Sigma(\mathbb{Q}')^{-1}\int xx^T w\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi^{(1)}\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'\right)$$
$$\times(b_{\mathrm{LMS}} - \beta) + O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}')).$$

As $|s_{\mathrm{LMS}} - \sigma| = O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}'))$ we obtain

$$\left|\int xx^T w\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi^{(1)}\left(\frac{y - x^T\beta}{s_{\mathrm{LMS}}}\right)d\mathbb{Q}'\right.$$
$$\left.-\int xx^T w\big(\xi(\mathbb{Q}'_d)x^T T^D_{\mathfrak{D}}(\mathbb{Q}'_d)^{-1}x\big)\psi^{(1)}\left(\frac{y - x^T\beta}{\sigma}\right)d\mathbb{Q}'\right|$$
$$= O(d_{\mathfrak{D}}(\mathbb{Q}, \mathbb{Q}')),$$

which, on using Theorem 8.1, the properties of ψ and w and the Lipschitz continuity of $T_{\mathfrak{D}}^T$, implies

$$\left| \int xx^T w\big(\xi(\mathbb{Q}_d')x^T T_{\mathfrak{D}}^D(\mathbb{Q}_d')^{-1}x\big)\psi^{(1)}\left(\frac{y - x^T\beta}{s_{\mathrm{HR}}}\right) d\mathbb{Q}' \right.$$

$$\left. - E(\psi^{(1)}) \int xx^T w\big(\xi(\mathbb{Q}_d)x^T T_{\mathfrak{D}}^D(\mathbb{Q}_d)^{-1}x\big)\, d\mathbb{Q} \right|$$

$$= O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}')).$$

The same argument gives $\|\Sigma(\mathbb{Q}') - \Sigma(\mathbb{Q})\| = O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}'))$ and hence finally $\|b_{\mathrm{OS}} - \beta\| = O(d_{\mathfrak{P}}(\mathbb{Q}, \mathbb{Q}'))$ as was to be shown. □

Finally, we consider the empirical behaviour of the one-step M-estimator T_{OS} defined previously. Let $(\mathbb{Q}_n)_1^\infty$ be a sequence of regression distributions of the form $\mathbb{Q}_n = \mathbb{Q}_{nd} * \mathbb{Q}_e$, where $\mathbb{Q}_{nd} = (1/n)\Sigma_1^n \delta_{x_j(n)}$, and let $(\hat{\mathbb{Q}}_n)_1^\infty$ be a corresponding sequence of empirical distributions with $\hat{\mathbb{Q}}_n = (1/n)\Sigma_1^n \delta_{(x_j(n)^T, y_j(n))^T}$, where the $((y_j(n) - x_j(n)^T\beta)/\sigma)_1^n$, $n = 1, 2, \dots,$ are independently and identically distributed with common distribution $\mathbb{Q}_e$. We have the following result.

THEOREM 8.4. *Let* $(\hat{\mathbb{Q}}_n)_1^\infty$ *and* $(\mathbb{Q}_n)_1^\infty$ *be as above. Then* $d_{\mathfrak{P}}(\hat{\mathbb{Q}}_n, \mathbb{Q}_n) = O_p(n^{-1/2})$, *where* O_p *denotes order in probability.*

PROOF. The class $\mathfrak{C}$ of sets defined by (8.1) is a Vapnik–Cervonenkis class and the reasoning of Lemma 3 of Davies (1990) shows that

$$\lim_{n \to \infty} sup_{C \in \mathfrak{C}} |\hat{\mathbb{Q}}_n(C) - \mathbb{Q}_n(C)| = 0$$

in probability. The maximal inequality [Kim and Pollard (1990) and Davies (1990)] implies

$$\sup_{C \in \mathfrak{C}} |\hat{\mathbb{Q}}_n(C) - \mathbb{Q}_n(C)| = O_p(n^{-1/2}).$$

As $d_{\mathfrak{P}}(\hat{\mathbb{Q}}_n, \mathbb{Q}_n) \le \sup_{C \in \mathfrak{C}} |\hat{\mathbb{Q}}_n(C) - \mathbb{Q}_n(C)|$, we obtain the statement of the theorem. □

Theorem 8.4 makes no assumptions about the design distribution apart from its being a fixed sequence. This is due to the use of the maximal inequality as given on page 1661 of Davies (1990). That it is applicable in the present situation is due to the fact that the class

$$\mathfrak{H}_n = \{h_{b,s} - h_{b',s'} : b, b' \in \mathbb{R}^k, s, s' \ge 0\},$$

where

$$h_{b,s}(e, x) = \{|e - x^T b| \le s\},$$

is a uniformly manageable class of bounded functions. Details are contained on pages 1661 and 1662 of Davies (1990).

We now consider the one-step M-estimator T_{OS} as described previously but with $\mathbb{Q}' = \hat{\mathbb{Q}}_n$. We assume that $\mathbb{Q}_e$ has a density function f which satisfies the conditions of Theorem 4.7 where the S-functional T_S is the Hampel–Rousseeuw functional T_{LMS} or T_{LMS}^G. Although we shall not require it we note that T_{LMS} is almost certainly uniquely defined at each $\hat{\mathbb{Q}}_n$ [Rousseeuw and Leroy (1987), page 206, and Steele and Steiger (1986)]. We have the following result.

THEOREM 8.5. *Suppose that there exists a design distribution* $\mathbb{Q}_d \in \mathfrak{W}_d(\mathbb{R}^k)$ *such that*

$$d_{\mathfrak{D}}(\mathbb{Q}_d, \mathbb{Q}_{nd}) = O(n^{-1/2}). \tag{8.5}$$

Then

$$\|T_{\mathrm{OS}}(\hat{\mathbb{Q}}'_n) - (\beta^T, \sigma)^T\| = O_p(n^{-1/2}). \tag{8.6}$$

PROOF. It follows from the assumptions placed on the ψ and f that $(\beta^T, \sigma)^T \in T_M(\mathbb{Q}_d * \mathbb{Q}_e)$. From Theorem 8.4 we have $d_{\mathfrak{P}}(\hat{\mathbb{Q}}_n, \mathbb{Q}_d * \mathbb{Q}_e) = O_p(n^{-1/2})$ which together with Theorem 8.3 implies (8.6), as was to be proved. □

We note that (8.5) holds if the design points $x_i(n)$, $1 \le i \le n$, are independently and identically distributed with distribution $\mathbb{Q}_d$. This is because the class of sets defined by (8.1) is a Vapnik–Cervonenkis class and the claim follows as in the proof of Theorem 8.4. The rate of convergence of T_{LMS} is estimated as $n^{-1/4}$ by Theorems 4.7, 8.4 and 8.5. Under more stringent conditions this can be improved to an exact order of $n^{-1/3}$ [Kim and Pollard (1990) and Davies (1990)], but in the present generality the rate $n^{-1/4}$ is probably the correct one.

Acknowledgments. The author would like to thank two referees and an Associate Editor for their detailed comments on the first three versions of this paper which led to improvements in all aspects, readability, correctness and precision.

REFERENCES

BEDNARSKI, T., CLARKE, B. R. and KOLKIEWICZ, W. (1991). Statistical expansions and locally uniform Fréchet differentiability. *J. Austral. Math. Soc. Ser. A* **50** 88–97.

BOLTHAUSEN, E. (1978). Weak convergence of an empirical process indexed by the convex subsets of I^2. *Z. Wahrsch. Verw. Gebiete* **43** 173–181.

CLARKE, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.* **11** 1196–1205.

DAVIES, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292.

DAVIES, L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.* **18** 1651–1675.

DAVIES, L. (1992a). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20** 1828–1843.

DAVIES, P. L. (1992b). An efficient Fréchet differentiable high breakdown multivariate location and dispersion estimator. *J. Multivariate Anal.* **40** 311–327.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann* (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, CA.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Dept. Statistics, Univ. California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

HAMPEL, F. R. (1975). Beyond location parameters: robust concepts and methods. *Proceedings of the 40th Session of the ISI* **46** 375–391.

HAMPEL, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proc. Statistical Computing Section* 59–64. Amer. Statist. Assoc., Washington, DC.

HAMPEL, F. R., ROUSSEEUW, P. J., RONCHETTI, E. M. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HETTMANSPERGER, T. P. and SHEATHER, S. J. (1992). A cautionary note on the method of least median squares. *Amer. Statist.* **46** 79–83.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

HUBER, P. J. (1983). Minimax aspects of bounded-influence regression (with discussion). *J. Amer. Statist. Assoc.* **78** 66–80.

JURECKOVA, J. and PORTNOY, S. (1987). Asymptotics for one-step M-estimators with application to combining efficiency and high breakdown point. *Comm. Statist. Theory Methods* **16** 2187–2199.

KERSTING, G. D. (1978). Die Geschwindigkeit der Glivenko–Cantelli-Konvergenz gemessen in der Prohorov-Metrik. *Math. Z.* **163** 65–102.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

KRASKER, W. S. (1980). Estimation in the linear regression models with disparate data points. *Econometrica* **48** 1333–1346.

KRASKER, W. S. and WELSCH, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

LOPUHAÄ, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683.

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248.

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MARONNA, R. A., BUSTOS, O. and YOHAI, V. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.) 91–116. Springer, New York.

MARONNA, R. A. and YOHAI, V. J. (1981). Asymptotic behaviour of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete* **58** 7–20.

MARONNA, R. A. and YOHAI, V. J. (1991). The breakdown point of simultaneous general M estimates of regression and scale. *J. Amer. Statist. Assoc.* **86** 699–703.

MORGENTHALER, S. (1989). Comment on Yohai and Zamar. *J. Amer. Statist. Assoc.* **84** 636.

NOLAN, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371.

OBENCHAIN, R. L. (1971). Multivariate procedures invariant under linear transformations. *Ann. Math. Statist.* **42** 1569–1578.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RIEDEL, M. (1989a). On the bias robustness in the location model. I. *Statistics* **20** 223–233.

RIEDEL, M. (1989b). On the bias robustness in the location model. II. *Statistics* **20** 235–246.

RIEDEL, M. (1991). Bias-robustness in parametric models generated by groups. *Statistics.* **22** 559–578.

RONCHETTI, E. and ROUSSEEUW, P. J. (1985). Change-of-variance sensitivities in regression analysis. *Z. Wahrsch. Verw. Gebiete* **68** 503–519.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

ROUSSEEUW, P. J. (1986). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Vol. B* (W. Grossmann, G. Pflug, I. Vincze and W. Wertz, eds.). Reidel, Dordrecht.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P. J. and YOHAI, V. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 256–272. Springer, New York.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85** 633–651.

SIMPSON, D. G., RUPPERT, D. and CARROLL, R. J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.* **87** 439–450.

STAHEL, W. A. (1981). Robuste Schätzungen: Infintesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. dissertation, Dept. Statistics, ETH, Zürich.

STEELE, J. M. and STEIGER, W. L. (1986). Algorithms and complexity for least median of squares regression. *Discrete Appl. Math.* **14** 93–100.

STEFANSKI, L. A. (1991). A note on high breakdown estimators. *Statist. Probab. Lett.* **11** 353–358.

YOHAI, V. J. (1987). High breakdown point and high efficiency robust estimators for regression. *Ann. Statist.* **15** 642–656.

YOHAI, V. J. and ZAMAR, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **83** 406–413.

FACHBEREICH MATHEMATIK UND INFORMATIK
UNIVERSITY OF ESSEN
D-4300 ESSEN
GERMANY