The Annals of Statistics
1993, Vol. 21, No. 1, 314–337

LOWER BOUNDS FOR CONTAMINATION BIAS: GLOBALLY MINIMAX VERSUS LOCALLY LINEAR ESTIMATION[1]

BY XUMING HE AND DOUGLAS G. SIMPSON

National University of Singapore and University of Illinois

We study how robust estimators can be in parametric families, obtaining a lower bound on the contamination bias of an estimator that holds for a wide class of parametric families. This lower bound includes as a special case the bound used to establish that the median is bias minimax among location equivariant estimators, and it is tight or nearly tight in a variety of other settings such as scale estimation, discrete exponential families and multiple linear regression. The minimum variation distance estimator has contamination bias within a dimension-free factor of this bound. A second lower bound applies to locally linear estimates and implies that such estimates cannot be bias minimax among all Fisher-consistent estimates in higher dimensions. In linear regression this class of estimates includes the familiar M-estimates, GM-estimates and S-estimates. In discrete exponential families, yet another lower bound implies that the "proportion of zeros" estimate has minimax bias if the median of the distribution is zero, a common situation in some fields. This bound also implies that the information-standardized sensitivity of every Fisher consistent estimate of the Poisson mean and of the Binomial proportion is unbounded.

1. Introduction. Huber (1964) established that the median is the most robust estimate of the center of a symmetric distribution in the following sense. Suppose $F_\theta(x) = F(x - \theta)$, where the distribution F is absolutely continuous and has a density f symmetric about zero and decreasing on $(0, \infty)$. Let T denote an estimating functional of θ, that is, T is a mapping from the space of distributions to the parameter space Θ. Define the bias function

$$b_T(\varepsilon; F_\theta) := \sup_G \left| T\left((1 - \varepsilon)F_\theta + \varepsilon G\right) - \theta \right|, \tag{1.1}$$

where $\varepsilon \in [0, 1]$. This is the maximum deviation of T from θ that can be induced by contamination of F_θ. Then the median functional $T(F) := F^{-1}(1/2)$ minimizes $b_T(\varepsilon; F_\theta)$ among location equivariant estimators. It is remarkable that this result holds for each ε, and no conditions beyond equivariance are imposed on the estimators.

In other settings, optimality results about contamination bias have usually restricted to smaller classes of estimators, and the earlier work focused primarily on infinitesimal point mass contaminations. Hampel (1974) considered general one-parameter models, and studied the trade-off between asymp-

Received May 1990; revised March 1992.

[1]Partially supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship and National Security Agency Grant NSA-MDA904-89-H-2011.

AMS 1991 *subject classifications*. Primary 62F35; secondary 62F10, 62G35.

Key words and phrases. Bias minimax, contamination model, discrete exponential family, invariance, linear regression, M-estimate, minimum distance estimate, scale model, sensitivity.

totic efficiency and local stability within the class of M-estimators, introducing the gross error sensitivity,

$$\gamma_T^{(\mathrm{GE})} := \sup_x \limsup_{\varepsilon \downarrow 0} \frac{|T((1-\varepsilon)F_\theta + \varepsilon\Delta_x) - \theta|}{\varepsilon}, \tag{1.2}$$

where Δ_x denotes a point mass at x. Hampel (1978), Krasker (1980), Krasker and Welsch (1982) and Ruppert (1985) considered extensions of Hampel's (1974) theory to linear regression, concentrating on GM-estimators (generalized M-estimators); see Hampel, Ronchetti, Rousseeuw and Stahel (1986) for a review. Stefanski, Carroll and Ruppert (1986) and Künsch, Stefanski and Carroll (1989) considered GM-estimates in generalized linear models. The emphasis in their work has been on constructing estimates with optimal asymptotic efficiency subject to a bound on the gross-error sensitivity.

Several recent results are more in the spirit of Huber's (1964) result concerning the bias optimality of the median. Hampel, Ronchetti, Rousseeuw and Stahel (1986) obtained minimal gross-error sensitivities for GM-estimates in regression. Martin and Zamar (1989) obtained the form of the M-estimate of scale with minimal bias, whereas Martin, Yohai and Zamar (1989) obtained the forms of S-estimators and GM-estimators with minimal bias for linear regression, in each case allowing full contamination neighborhoods as in (1.1). In a notable recent development, Maronna and Yohai (1991) constructed a regression equivariant functional with a high breakdown point and contamination sensitivity within a factor of 2 of the minimum among all regression equivariant functionals.

In the present paper we consider the general question of how robust estimators can be if one ignores efficiency considerations. For contamination neighborhoods the known bias optimality results in the location model [Huber (1964), (1981)] and linear regression [Martin, Yohai and Zamar (1989)] involved rather specialized derivations. One naturally wonders whether such results can be obtained more generally. We provide a partial answer to this question by presenting a global lower bound for contamination bias for parametric families of models. This unifying framework leads to tight or nearly tight lower bounds for contamination bias in a variety of settings, including location estimation, regression through the origin, scale models, discrete exponential families and multiple linear regression. Special cases include the bound of Huber [(1981), page 75] for location equivariant estimators and the bound of Maronna and Yohai (1991) for regression equivariant functionals. Techniques of Donoho and Liu (1988a) establish that in general the minimum variation distance functional has bias within a dimension-free factor of the lower bound; however, in particular examples it is often possible to construct a better estimator based on bias and efficiency considerations.

Local to the model the stability of an estimator is indicated by the contamination sensitivity:

$$\gamma_T^* := \limsup_{\varepsilon \downarrow 0} b_T(\varepsilon; F_\theta)/\varepsilon, \tag{1.3}$$

where $b_T(\varepsilon; F_\theta)$ is given in (1.1). Various authors have made the heuristic

connection between $\gamma_T^{(GE)}$ and γ_T^*; see, for example, Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 175]. One always has $\gamma_T^* \geq \gamma_T^{(GE)}$ if the same norm is used for both [He and Simpson (1992)]. Our global lower bound for bias implies a dimension-free lower bound for contamination sensitivity, and the minimum variation distance functional is within a dimension-free factor of this bound. However, for a large class of regular estimators we find that the contamination sensitivity must increase essentially like the square root of the dimension. Hence, estimators that are regular and have dimension-free bias function must have unbounded contamination sensitivities. Examples from linear regression include the S-estimators of Rousseeuw and Yohai (1984), the MM-estimators of Yohai (1987) and the τ-estimators of Yohai and Zamar (1988). By similar reasoning the minimum variation distance estimator cannot be regular in high dimensions.

We present a specialized bound for discrete exponential families with positive support. This bound is tight if the median of the model distribution is 0, which corresponds to a situation common in practice. For instance, Simpson (1987) presented count data exhibiting a substantial proportion of zeros. The bound is achieved by the "proportion of zeros" estimate, which therefore has minimax bias for each ε if the median of the model distribution is 0. In nonequivariant problems of this type Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 229] and He and Simpson (1992) discussed standardizing sensitivities by the root Fisher information to achieve a degree of invariance. The bound presented in Section 4.2 implies that the information-standardized sensitivity of *every* Fisher-consistent estimator of the mean in the Binomial or Poisson model has to be unbounded as a function of the parameter.

The rest of the paper is organized as follows. Section 2 presents the main results and some of their consequences. This section also introduces technical machinery that is used in subsequent discussions. Section 3 considers location estimation, regression through the origin and scale estimation, each of which is an equivariant problem. Section 4 is concerned with discrete exponential families, which do not have exact equivariance properties. Section 5 considers the trade-off between bias minimaxity and regular estimation in multiple linear regression. Some concluding remarks are made in Section 6. Technical proofs are given in the Appendix.

2. Lower bounds for contamination bias. In general one might like to consider neighborhoods other than the contamination type. Suppose $d(F, G)$ is a metric or a nonnegative discrepancy on the space of distribution functions such as the amount of contamination to reach G from F. A key property is that the sets $\{G\colon d(F, G) \leq \varepsilon\}$ are increasing in ε. Let $\rho(\cdot, \cdot)$ be a metric on the parameter space. The bias function

$$b_T(\varepsilon; F) := \sup_G \{\rho(T(G), T(F))\colon d(F, G) \leq \varepsilon\} \tag{2.1}$$

summarizes various aspects of the stability of T at F. Qualitative robustness of T with respect to $d(\cdot, \cdot)$ corresponds to continuity, that is, $b_T(0+; F) = 0$. Quantitative robustness refers to the magnitude of $b_T(\varepsilon; F)$ for nonzero ε. In

particular, the sensitivity

$$\gamma_T^* := \limsup_{\varepsilon \downarrow 0} b_T(\varepsilon; F_\theta)/\varepsilon$$

indicates the stability of the functional close to the model, whereas the breakdown point

$$\varepsilon_T^* := \inf\Big\{\varepsilon\colon b_T(\varepsilon; F_\theta) = \sup_\delta b_T(\delta; F_\theta)\Big\}$$

indicates how far from the model the functional becomes completely uninformative. Various authors have employed versions of the bias function, including Huber (1964, 1981), Donoho and Liu (1988a), Martin and Zamar (1989), Martin, Yohai and Zamar (1989) and He and Simpson (1992). Huber (1981) provided a general discussion. He, Simpson and Portnoy (1990) and He (1991) studied the robustness of test statistics on the scale inverse to that of the sup-bias in order to obtain comparisons invariant to one-to-one transformations of parameters.

2.1. *A global bound.* An estimating functional T is said to be Fisher consistent for a parameter $\theta \in \Theta$ if $T(F_\theta) \equiv \theta$. For such estimators, Donoho and Liu (1988a) introduced the gauge $b_0(\varepsilon) := \sup\{\rho(\theta, \eta)\colon d(F_\theta, F_\eta) \le \varepsilon\}$. Because of the restriction to a parametric subset of the neighborhood, $b_T(\varepsilon) \ge b_0(\varepsilon)$ for Fisher-consistent functionals.

We focus on contamination neighborhoods, which correspond to the discrepancy

$$\begin{aligned} d_c(P, Q) &:= \inf\{\varepsilon \ge 0\colon Q = (1-\varepsilon)P + \varepsilon R \text{ for some distribution } R\} \\ &= \inf\Big\{\varepsilon \ge 0\colon \sup_{\text{meas. } A} (1-\varepsilon)P(A) - Q(A) \le 0\Big\}. \end{aligned}$$

In general, the contamination gauge does not yield a tight lower bound for contamination bias, but a notable exception is presented in Section 4.2. As a technical tool we also use the variation norm

$$\begin{aligned} d_v(P, Q) &:= \sup_{\text{meas. } A} |P(A) - Q(A)| \\ &= \frac{1}{2}\int |p - q|\, d\lambda = \int (p - q)_+\, d\lambda, \end{aligned} \tag{2.2}$$

where λ is any σ-finite measure dominating P and Q, such as $(P + Q)/2$, and p and q are densities of P and Q with respect to λ. The second equality in (2.2) is Scheffe's theorem. It is not difficult to show that $d_c(P, Q) \ge d_v(P, Q)$, so a contamination neighborhood of size ε is contained in a variation neighborhood of size ε. A variation neighborhood need not be contained in a contamination neighborhood. However, the variation gauge

$$b_v(\varepsilon; F_\theta) := \sup\big\{\rho(\theta, \eta)\colon \eta \text{ such that } d_v(F_\theta, F_\eta) \le \varepsilon\big\}$$

appears in our lower bound for contamination bias.

We write the contamination bias as $b_T(\varepsilon; F_\theta) = \sup_{d_c(F_\theta, F) \le \varepsilon} \rho(T(F), \theta)$, which is valid even for functionals that are not Fisher consistent.

THEOREM 2.1. *Suppose $\{F_\theta\}$ is dominated by a σ-finite measure. If T is a functional mapping distributions to parameter values, then its contamination bias satisfies*

$$\sup_{\eta:\, \rho(\theta,\eta) \le b_v(\varepsilon/(1-\varepsilon);\, F_\theta)} b_T(\varepsilon; F_\eta) \ge \frac{1}{2} b_v\left(\frac{\varepsilon}{1-\varepsilon}; F_\theta\right). \tag{2.3}$$

REMARK 2.1 (Constant functionals). As an aid to understanding (2.3) consider the functional $T(F) \equiv \theta_0$ with the Euclidean metric for bias. In this case, $b_T(\varepsilon; F_\eta) = \|\theta_0 - \eta\|$. If $\theta = \theta_0$, then the left-hand side of (2.3) becomes simply $b_v(\varepsilon/(1-\varepsilon); F_\theta)$, twice the lower bound for estimating θ_0. Although the constant functional achieves the minimum bias of 0 at one point, the left-hand side of (2.3) is perhaps a more meaningful way to measure bias when θ is not known a priori.

REMARK 2.2 (Model breakdown point). Define the model breakdown point to be the smallest amount of contamination such that the estimator breaks down for some parameter value; formally, $\varepsilon_T^* := \inf\{\varepsilon\colon \sup_{\theta \in \Theta} b_T(\varepsilon; F_\theta) = \sup_{\eta \in \Theta} \sup_{\delta > 0} b_T(\delta; F_\eta)\}$. As the variation distance between two distributions can be at most 1, setting $\varepsilon = 1/2$ in (2.3) yields $\sup_{\eta \in \Theta} b_T(1/2; F_\eta) \ge (1/2) \sup_{\eta \in \Theta} \rho(\theta, \eta)$. Hence, we obtain the following.

COROLLARY 2.1. *If $\sup_{\theta, \eta \in \Theta} \rho(\theta, \eta) = \infty$, then the model breakdown point of an estimator can be at most $1/2$.*

REMARK 2.3 (Invariance). A simplification occurs in problems that have invariance properties. Rather than detailing general conditions for invariant estimation, we shall simply state that an invariance is present if there is a metric on the parameter space such that $b_v(\varepsilon; F_\theta)$ is independent of θ, and with the same metric there is a class of functionals, called equivariant functionals, such that $b_T(\varepsilon; F_\theta)$ is independent of θ. Section 3 provides several examples in which such invariance occurs. If an invariance is present and we restrict ourselves to equivariant functionals, then the supremum over η in (2.3) is unnecessary. Moreover, Corollary 2.1 implies that equivariant estimators can have breakdown point no larger than $1/2$. Special cases of the latter result have been derived previously in various settings.

2.2. *A bias-robust functional.* A surprising feature of the lower bound given in (2.3) is that it is dimension free. However, we shall exhibit a functional whose contamination bias is within a dimension-free factor of the bound. As the lower bound in (2.3) involves the variation distance, results of Donoho and Liu (1988) suggest consideration of the minimum variation

distance functional,

$$T(F) := \arg\min_{\theta \in \Theta} d_v(F, F_\theta). \tag{2.4}$$

In practice one would need to ensure that $\hat{F}$, the sample-based version of F, is not singular with respect to the model because otherwise $d_v(\hat{F}, F_\theta) \equiv 1$. When modeling a sample of n counts by, for example, a power series distribution (see Section 4), one might use the empirical distribution $\hat{F}(A) := n^{-1}\sum_{i=1}^{n} 1_A(X_i)$ as in Simpson (1987). However, in most settings it will be necessary to smooth the empirical distribution in some way such as convolving it with a continuous distribution as in kernel density estimation [Parzen (1962)].

Such concerns do not arise in connection with our formal results about the contamination bias of the minimum variation distance functional, because if F is in an ε-contamination neighborhood of F_θ it is also in an ε-variation neighborhood of F_θ. We use the following result due to Donoho and Liu (1988a).

PROPOSITION 2.1. *For the minimum variation distance functional* (2.4), $b_T(\varepsilon; F_\theta) \leq b_v(2\varepsilon; F_\theta)$.

Combining this with Theorem 2.1 we find that if an invariance is present so that $b_v(2\varepsilon; F_\theta)$ is independent of θ, then the bias of the minimum variation distance estimator is deficient by a factor of at most $2b_v(2\varepsilon)/b_v(\varepsilon/(1-\varepsilon))$ relative to the bound for equivariant functionals. As an interesting special case, if bias is measured by the parameterization invariant metric $d_v(F_T, F_\theta)$, then the deficiency factor is at most $4(1-\varepsilon)$. In other cases we evaluate the limits.

PROPOSITION 2.2. *Suppose* $b_v(\varepsilon)$ *is continuous on* $(0, 1/2)$, *and* $b_v(\varepsilon) = \varepsilon\gamma_v + o(\varepsilon)$ *as* $\varepsilon \downarrow 0$ *with* $\gamma_v > 0$. *Then the functional defined in* (2.4) *has*

$$\limsup_{\varepsilon \downarrow 0} \frac{2b_T(\varepsilon)}{b_v(\varepsilon/(1-\varepsilon))} \leq 4 \quad \textit{and} \quad \limsup_{\varepsilon \uparrow 1/2} \frac{2b_T(\varepsilon)}{b_v(\varepsilon/(1-\varepsilon))} \leq 2.$$

It follows that the minimum variation distance estimator has the best possible breakdown point and contamination bias no more than twice the minimum for equivariant functionals as the amount of contamination approaches $1/2$. Moreover, it has bounded bias sensitivity within a factor of 4 of best possible. However, this result serves primarily to establish that the lower bound in Theorem 2.1 is almost tight. The minimum variation distance estimator as presented poses serious computational challenges in multivariate settings, because it entails multivariate density estimation, multivariate numerical integration and nonlinear optimization. Donoho and Liu (1988b) observed certain pathological properties of the minimum variation distance estimate of location.

2.3. *Local considerations.* In general we assume that $\{F_\theta\}$ is dominated by a σ-finite measure λ such as Lebesgue or counting measure. In this section we assume further that Θ is an open subset of p-dimensional Euclidean space, and that bias is measured by the Euclidean distance, possibly after a parameter transformation. Let f_θ be a density of F_θ with respect to λ. We say the family $\{F_\theta\}$ is L_1 differentiable if there is a function u_θ with components in $L_1(f_\theta)$ such that

$$\int |f_{\theta+\delta} - f_\theta - \delta' u_\theta f_\theta| \, d\lambda = o(\|\delta\|). \tag{2.5}$$

The function u_θ in (2.5) is essentially the likelihood score function. If the model has this differentiability property, then (2.3) reduces to the following lower bound locally.

COROLLARY 2.2. *Suppose for each $\theta \in \Theta$ that F_θ has density f_θ with respect to a σ-finite measure λ. Suppose $\{F_\theta\}$ is L_1 differentiable. Then*

$$\limsup_{\varepsilon \downarrow 0} \sup_{\|\delta\| \le b_v(\varepsilon/(1-\varepsilon); F_\theta)} \frac{b_T(\varepsilon; F_{\theta+\delta})}{\varepsilon} \ge \gamma_v(\theta), \tag{2.6}$$

where

$$\gamma_v(\theta) = \left\{ \inf_{\|z\|=1} E_\theta |z' u_\theta(X)| \right\}^{-1}$$

If an invariance is present so that the bias of an equivariant estimator is invariant to θ, then the supremum in (2.6) can be removed; see Section 3 for examples. In noninvariant problems, the left side of (2.6) defines an extended notion of contamination sensitivity.

Suppose a functional T is defined at F and contaminations of F of the form $F_{\varepsilon,x} := (1-\varepsilon)F + \varepsilon\Delta_x$ if $\varepsilon \ge 0$ is sufficiently near zero, where Δ_x is the distribution assigning probability 1 to $\{x\}$. Hampel (1974) defined the *influence function*

$$IF(x; T, F) := \lim_{\varepsilon \downarrow 0} \{T(F_{\varepsilon,x}) - T(F)\}/\varepsilon$$

assuming this derivative exists. The influence function has been used extensively in the literature as a heuristic tool; see Hampel, Ronchetti, Rousseeuw and Stahel (1986) for a comprehensive treatment. The gross-error sensitivity $\gamma_T^{(\mathrm{GE})} = \sup_x \|IF(x; T, F)\|$ is often used to quantify the stability of a functional T. If the influence function fails to exist, the gross-error sensitivity can be defined without reference to the influence function as in (1.2). With the more general definition, He and Simpson (1992) observed that the gross-error sensitivity bounds the bias sensitivity below. We use this fact in the next section.

PROPOSITION 2.3. *For any functional* T, $\gamma_T^{(GE)} \leq \gamma_T^*$.

2.4. *Regular estimating functionals.* It turns out that a large class of regular estimators cannot have contamination bias within a finite dimension-free factor of the bound in (2.3). This class essentially consists of functionals that have influence functions, are Fisher consistent and are approximately linear within the parametric family. In view of the results for the minimum variation distance functional, it follows that there is a conflict between bias minimaxity and regular estimation in higher dimensions. We assume here that Θ is an open subset of p-dimensional Euclidean space.

DEFINITION 2.1. Suppose $\{F_\theta; \theta \in \Theta\}$ is an L_1-differentiable family of distributions with finite Fisher information. A functional T is *locally linear* if it has the following properties:

(i) For each finite x and $\theta \in \Theta$, T has an influence function $\psi_\theta(x) := IF(x; T, F_\theta)$ such that $E_\theta[\psi_\theta(X)] = 0$ and $E_\theta\|\psi_\theta(X)\|^2 < \infty$.

(ii) For each $\theta \in \Theta$, and as $\|\delta\| \to 0$,

$$T(F_{\theta+\delta}) - T(F_\theta) = \int \psi_\theta \, d(F_{\theta+\delta} - F_\theta) + o(\|\delta\|). \tag{2.7}$$

REMARK 2.4. A functional T is Fréchet differentiable at F_θ with respect to the variation distance if ψ_θ is bounded and $T(F) - T(F_\theta) = \int \psi_\theta \, d(F - F_\theta) + o(d_v(F_\theta, F))$ holds for any sequence of distributions F such that $d_v(F_\theta, F) \to 0$. Fréchet differentiable functionals are locally linear, because the L_1 differentiability of $\{F_\theta\}$ implies $d_v(F_\theta, F_{\theta+\theta}) = O(\|\delta\|)$. If T is Fisher consistent, the left side of (2.7) is simply δ.

REMARK 2.5. Variation distances are larger than Prohorov, Lévy and Kolmogorov–Smirnov distances, the latter two being associated with the root-n convergence of the empirical distribution function; see Huber [(1981), pages 34–39]. Hence, estimators that are Fréchet differentiable with respect to these distances are also locally linear.

REMARK 2.6. Hadamard or compact differentiability is a weaker condition than Fréchet differentiability. Under slightly different conditions on $\{F_\theta\}$, Fernholz [(1983), page 117] established that Hadamard differentiable functionals satisfy conditions (i) and (ii) of Definition 2.1.

EXAMPLE. If F_θ has a finite mean vector $\mu(\theta)$ and a finite covariance matrix, then the mean functional $T(F) := E_F(X)$ is a locally linear estimate of $\mu(\theta)$ with influence function $x - \mu(\theta)$ and remainder 0 in (2.7). The mean is not Fréchet differentiable unless the sample space is bounded.

THEOREM 2.2. *Assume Θ is an open subset of p-dimensional Euclidean space. Suppose $\{F_\theta\}$ is L_1 differentiable with finite Fisher information. If T is Fisher consistent and locally linear, then $\gamma_T^*(\theta) \geq p/E_\theta\|u_\theta(X)\|$.*

REMARK 2.7. In the proof it is shown that if T is Fisher consistent with a bounded influence function, and if $\{F_\theta\}$ is L_1 differentiable, then (2.7) is equivalent to the requirement that the influence function satisfy $E_\theta[\psi_\theta(X)u_\theta(X)'] = I_p$. This condition is usually straightforward to check.

Theorem 2.2 does not require a supremum over the parameter space, in contrast to Corollary 2.2, because of its restriction to a class of regular functionals. Both yield the same bound if $p = 1$.

EXAMPLE. Consider estimation of the mean in the p-dimensional normal model with covariance I_p. In this case $E_0\|X\| = E|W_p|$, where W_p^2 is chi-square with p degrees of freedom, and $p/E\|X\| = \sqrt{2}\,\Gamma(1/2(p+2))/\Gamma(1/2(p+1)) \geq p^{1/2}$. On the other hand, for any unit vector z, $E_0|z'X| = (2/\pi)^{1/2}$, so the lower bound of Corollary 2.2 is $(\pi/2)^{1/2}$ regardless of the dimension.

3. Some equivariant examples.

3.1. *Location of a symmetric distribution.* Given any distribution function F one can define a location-shift family of distributions $\{F_\theta\}$ by setting $F_\theta(x) := F(x - \theta)$. A functional T is location equivariant if it satisfies $T(F_\theta) = T(F) + \theta$ for each θ and each F for which T is defined.

Suppose F is a continuous distribution function on the real line with density f symmetric about 0 and radially decreasing, which is true if F is Gaussian, Cauchy, double exponential and so on. Define the location-shift family with density $f_\theta(x) := f(x - \theta)$, where f is taken to be continuous at 0. As an application of Theorem 2.1, we obtain Huber's (1964) result that for each ε, the median is bias minimax among location equivariant estimators of θ. We also verify that the median is locally linear in the sense of Definition 2.1.

First observe that if T is equivariant, then $b_T(\varepsilon; F_\theta) = b_T(\varepsilon; F_0)$, the reason being that $T((1-\varepsilon)F_\theta + \varepsilon G) - \theta = T((1-\varepsilon)F_0 + \varepsilon G_{-\theta})$, and $G_{-\theta}(x) = G(x + \theta)$ still ranges over all contaminating distributions. Moreover, the variation distance is invariant, that is, $d_v(F_\theta, F_{\theta+\delta}) = d_v(F_0, F_\delta)$. Hence, by Theorem 2.1 an equivariant functional satisfies

$$b_T(\varepsilon; F_\theta) = b_T(\varepsilon; F_0) \geq \frac{1}{2} b_v\left(\frac{\varepsilon}{1-\varepsilon}; F_0\right). \tag{3.1}$$

In order to evaluate the lower bound in (3.1) use the symmetry of f and (2.2) to calculate

$$d_v(F_{\theta/2}, F_{-\theta/2}) = \int_{-\infty}^{0} \left\{ f\left(x + \frac{1}{2}|\theta|\right) - f\left(x - \frac{1}{2}|\theta|\right) \right\} dx$$
$$= 2F\left(\frac{|\theta|}{2}\right) - 1. \tag{3.2}$$

It follows that $b_v(\varepsilon) = 2F^{-1}(1 + (1/2)\varepsilon)$, and $(1/2)b_v((\varepsilon/(1 - \varepsilon); F_0) = F^{-1}(2^{-1}(1 - \varepsilon)^{-1})$. This lower bound, which applies to all location equivariant functionals, is the same as the contamination bias of the median computed by Huber [(1981), page 75]; the bound is tight.

Under the assumptions on F, the median is Fisher consistent for θ. Moreover, it has the influence function $\psi_\theta(x) = (1/2)\text{sign}(x - \theta)/f(0)$; see Huber [(1981), page 57]. A direct calculation establishes that (2.7) holds:

$$\int \psi_0(x)\{f(x - \delta) - f(x)\}\, dx = \frac{1 - 2F(\delta)}{2f(0)} = \delta + o(|\delta|),$$

so the median is locally linear.

3.2. *Regression through the origin.* Let F be a bivariate distribution function for jointly distributed random variables X and U. One can define a linear regression family of distributions via transformations of the form $F_\theta(x, u) := F(x, u - x\theta)$. Observe that F_θ is the distribution of $(X, Y) := (X, X\theta + U)$. The usual linear model assumes X and U are independent, and U has a continuous distribution symmetric about 0. We consider functionals for estimating θ.

Using the transformation family defined above a functional T is regression equivariant if $T(F_\theta) = T(F_0) + \theta$ for each real θ. This is the functional version of the sample-based definition given by Rousseeuw and Leroy [(1987), page 116]. To remove the dependence of the bias on the scale of X we use the parameter distance $\rho(T, \theta) := |T - \theta|/s(F_X)$, where s is a scale equivariant functional and F_X is the marginal distribution of X. Hence, we can assume that F_X has unit scale. By reasoning similar to that in Section 3.1, the contamination bias is invariant in θ.

Observations at $X = 0$ provide no information about θ, so we assume that $\{X = 0\}$ has probability 0 under F. Given a bivariate distribution for X and Y, let F_R denote the distribution of $R := Y/X$. Under the model, θ is a location parameter for R, which has a continuous distribution symmetric about θ. Hence, the median ratio $F_R^{-1}(1/2)$ has the smallest contamination bias among location equivariant functionals of F_R. Such functionals are regression equivariant, but not all regression equivariant functionals are functionals of the ratio distribution. Consider the ordinary least squares estimate. However, Martin, Yohai and Zamar (1989) reported that the median ratio minimizes the contamination bias over the broader class of regression equivariant functionals. We verify this result via Theorem 2.1.

Using Huber's [(1981), page 75] calculation, the median ratio has contamination bias $F_R^{-1}(2^{-1}(1-\varepsilon)^{-1})$, where F_R is the distribution of the ratio for F_0. For positive r,

$$F_R(r) = P_0(X > 0, Y \le Xr) + P_0(X < 0, Y \ge Xr)$$

$$= \int_0^\infty F_U(|xr|)\, dF_X(x) + \int_{-\infty}^0 \{1 - F_U(-|xr|)\}\, dF_X(x) = E_0[F_U(|Xr|)]$$

using the symmetry of F_U. The bias $b = b(\varepsilon)$ of the median ratio therefore solves $E_0[F_U(|Xb|)] = 2^{-1}(1-\varepsilon)^{-1}$

In order to calculate the lower bound given by Theorem 2.1 we solve for b_0 in the equation $d_v(F_0, F_{2b_0}) = \varepsilon/(1-\varepsilon)$. Let f_U denote the density for F_U. Then, using (3.2),

$$d_v(F_0, F_{2b_0}) = \frac{1}{2}\iint |f_U(u + 2xb_0) - f_U(u)|\, du\, dF_X(x)$$

$$= 2\int F_U(|xb_0|)\, dF_X(x) - 1.$$

Upon rearranging we find that b_0 satisfies $E_0[F_U(|Xb_0|)] = 2^{-1}(1-\varepsilon)^{-1}$. Hence, $b = b_0$, and the median ratio achieves the minimum contamination bias for equivariant functionals in regression through the origin.

3.3. *Scale of a positive random variable.* Starting with F as in Section 3.1, one can define a scale family of distributions by setting $F_\theta(x) := F(x/\theta)$. If F has density f symmetric about 0, then the absolute values of the observations are sufficient for θ. Hence, we shall assume that the unit model F has positive support. Martin and Zamar (1989) considered robust estimation of θ, obtaining the form of the min-max bias estimate among M functionals of the form

$$S(\chi, F) = \inf\left\{s : \int_0^\infty \chi(x/s)\, dF(x) \le b\right\},$$

where $\chi(0) = 0$, and χ is nondecreasing on $[0, \infty)$ with at most a finite number of discontinuities. The min-max estimate depends on ε, but a scaled median functional $S(F) := cF^{-1}(1/2)$ provides comparable performance.

A functional T is scale equivariant if $T(F_\theta) = \theta T(F_1)$ for each $\theta > 0$ and F such that $T(F)$ is defined. We shall obtain lower bounds for all scale equivariant functionals. In order to measure bias invariantly we employ the distance $\rho(T, \theta) := |\log(T/\theta)|$. In other words, we evaluate the estimate of the log-scale parameter. In this metric, $b_T(\varepsilon; F_\theta) = b_T(\varepsilon; F_1)$ for each equivariant functional T. The variation distance $d_v(F_\theta, F_{\tau\theta})$ and parameter distance $|\log(\tau\theta) - \log(\theta)|$ are both invariant to θ, so $b_v(\varepsilon; F_\theta)$ is invariant as well. Theorem 2.1 implies that a scale equivariant functional satisfies

$$b_T(\varepsilon; F_\theta) = b_T(\varepsilon; F_1) \ge \frac{1}{2} b_v\left(\frac{\varepsilon}{1-\varepsilon}; F_1\right). \tag{3.3}$$

TABLE 1
Contamination biases and lower bounds for log-scale estimation in the exponential model

ε	Scaled median	Min-Max M	Lower bound
0.10	0.165	0.159	0.1516
0.20	0.389	0.359	0.3466
0.30	0.723	0.632	0.6195
0.40	1.335	1.080	1.0779
0.45	1.984	1.530	1.5233

Suppose $\{F_\theta\}$ has monotone likelihood ratio in X. Then, using (2.2),

$$d_v(F_1, F_\theta) = \int_0^\infty \left(f(x) - \frac{1}{\theta} f\left(\frac{x}{\theta}\right) \right)_+ dx = |F_1(a(\theta)) - F_\theta(a(\theta))|,$$

where $a(\theta)$ solves $f_1(a(\theta)) = f_\theta(a(\theta))$. To compute the lower bound b_0 in (3.3) one would solve for θ_ε in the equation $|F_1(a(\theta_\varepsilon)) - F_{\theta_\varepsilon}(a(\theta_\varepsilon))| = \varepsilon/(1 - \varepsilon)$, and set $b_0 := 1/2|\log(\theta_\varepsilon)|$.

As an example, consider the exponential scale model with $f(x) = e^{-x}$. In this case $a(\theta) = \theta(\theta - 1)^{-1}\log(\theta)$. The lower bound in (3.3) becomes $-(1/2)\log(\theta_\varepsilon)$, where θ_ε solves

$$\theta^\theta(1 - \theta)^{(1-\theta)} = \left(\frac{\varepsilon}{1 - \varepsilon}\right)^{(1-\theta)} \tag{3.4}$$

Taking logarithms and applying the Newton–Raphson algorithm yields the iteration sequence

$$\theta^{(k+1)} := \left(1 - \frac{\log(\theta^{(k)})}{\log(1 - \theta^{(k)}) + \operatorname{logit}(\varepsilon)}\right)^{-1}$$

for solving (3.4). The computed bounds for several choices of ε are given in Table 1. For comparison we show the bias values reported by Martin and Zamar (1989) for the scaled median and min-max bias M functional. It is interesting to note that their min-max functional is very close to minimax among all scale equivariant functions.

4. A class of noninvariant models: Power series distributions. Suppose F_θ is a discrete distribution on the set of nonnegative integers with the density

$$f_\theta(x) = a(x)\theta^x/c(\theta), \qquad x = 0, 1, 2, \ldots, v, \theta \in \Theta. \tag{4.1}$$

Here v may be finite or infinite. The parameter set in (4.1) is the interval of positive θ for which $c(\theta) := \sum_{x=0}^{v} a(x)\theta^x$ is convergent. This is a discrete exponential family, also known as a power series model [Johnson and Kotz

(1968)]. Examples include the following:

Binomial$(v, \theta/(1+\theta))$	$c(\theta) = (1+\theta)^v$;
Poisson(θ)	$c(\theta) = e^\theta$;
Negative Binomial	$c(\theta) = (1-\theta)^{-s} (s > 0)$;
Logarithmic Series	$c(\theta) = -\log(1-\theta)$.

We consider estimates of θ and functions of θ such as the mean $\mu(\theta) = \theta\dot{c}(\theta)/c(\theta)$ and the proportion $p(\theta) := \mu(\theta)/v$ if v is finite. Unlike the models of the preceding section, power series distributions lack exact invariance properties. However, they satisfy our regularity conditions.

PROPOSITION 4.1. *For a power series model, $\theta \mapsto f_\theta$ is L_1 differentiable and F_θ has finite Fisher information if θ is in the interior of Θ.*

The mean functional, which corresponds to maximum likelihood estimation of $\mu(\theta)$, has contamination bias $b_T(\varepsilon) = \varepsilon \cdot \max\{\mu(\theta), v - \mu(\theta)\} \geq v\varepsilon/2$. If v is infinite as in the Poisson model, then the mean is nonrobust with $b_T(\varepsilon) = \infty$ for $\varepsilon > 0$. If v is finite as in the binomial case, then the mean functional is robust, but its contamination sensitivity can be improved.

4.1. *Fisher consistent M-estimates.* Given a random sample $X_1, \ldots, X_n$, the maximum likelihood estimator of θ solves

$$n^{-1} \sum_{i=1}^{n} \left\{ X_i - \mu\left(\hat{\theta}_{\mathrm{ML}}\right) \right\} = 0, \tag{4.2}$$

so $\mu(\hat{\theta}_{\mathrm{ML}}) = n^{-1}\sum_{i=1}^{n} X_i$. Hampel's (1974) Fisher-consistent M-estimators generalize (4.2) as follows:

$$\begin{aligned} &\text{solve} && E_{\hat{\theta}_M}[\psi_\kappa(X - \beta)] = 0, \\ &\text{subject to} && n^{-1} \sum_{i=1}^{n} \psi_\kappa(X_i - \beta) = 0, \end{aligned} \tag{4.3}$$

where $\psi_\kappa(u) := \min\{\kappa, \max(-\kappa, u)\}$, and κ may depend on θ. Setting $\kappa \equiv \infty$ in (4.3) yields (4.2). The functionals corresponding to maximum likelihood estimates and Hampel's M-estimates replace the empirical averaging in (4.2) and (4.3) by expectations with respect to F. Hampel (1974) showed that the estimator given by (4.3) minimizes asymptotic variance within the class of Fisher-consistent M-estimators subject to the bound on the influence function. Simpson, Carroll and Ruppert (1987) studied the large sample theory and recommended to replace ψ_κ by a differentiable function when fitting discrete models. Stefanski, Carroll and Ruppert (1986) and Künsch, Stefanski and Carroll (1989) discussed extensions of (4.3) for generalized linear models.

We first consider M-estimates that take the form of (4.3), except with the piecewise linear ψ_κ replaced by $\psi_\kappa(u) := \kappa\psi(u/\kappa)$, where ψ satisfies the follow-

ing conditions:

(i) ψ is monotone increasing and odd, that is, $\psi(-u) = -\psi(u)$ for all u.
(ii) $\psi(\infty) = 1$.
(iii) ψ has a derivative $\dot{\psi}$ such that $0 < \dot{\psi}(u) \leq 1$ for all real u.

An example is $\psi(u) = \tanh(u)$. Denote the M-estimator with tuning constant κ by T_κ.

PROPOSITION 4.2. *Suppose θ is in the interior of Θ in the power series family. Then for each positive, finite κ, T_κ is Fréchet differentiable in the variation norm, and it has contamination sensitivity $\gamma_\kappa^*(\theta) := |E_\theta \psi((X - \beta)/\kappa) u_\theta(X)|^{-1}$, where β solves $E_\theta[\psi_\kappa(X - \beta)] = 0$. Moreover, $\gamma_\kappa^*(\theta)$ is increasing in κ.*

REMARK 4.1 (The least sensitive M-estimate). The smallest contamination sensitivity is obtained by letting $\kappa \to 0$, which yields the limit $\gamma_0^* := \theta/E_\theta|X - x_0|$, where x_0 is the smallest integer $k \in \{0, 1, \ldots, v\}$ such that $F_\theta(k) \geq 1/2$. In general, an M-estimate can have sensitivity arbitrarily close to γ_0^*. If $x_0 = 0$, then γ_0 is achieved; see Section 4.2.

In the present setting the minimum sensitivity among M-estimators is within a factor of 2 of the bound given by Corollary 2.2.

PROPOSITION 4.3. *In a power series family $\gamma_0^* \leq 2/E_\theta|u_\theta(X)|$.*

REMARK 4.2 (Binomial p estimation). The maximum likelihood estimate of the binomial $p(\theta) = \theta/(1 + \theta)$ corresponds to the empirical proportion $T(F) := E_F[X]/v$, which $\gamma_T^* = \max(p, 1 - p)$. For large v in the Binomial(v, p) model we have $\gamma_0^*(p) \sim v^{-1/2}\{2\pi p(1 - p)\}^{1/2}$. Hence, the sensitivity of the empirical proportion can be improved at least by a factor of order $v^{1/2}$ for moderate values of p.

4.2. *A bias bound for small θ.* Calculation shows that if the median is zero and κ is sufficiently small, then (4.3) reduces to the proportion-of-zeros estimate of θ, which corresponds to the functional T defined by

$$\frac{a(0)}{c(T(F))} = f(0), \tag{4.4}$$

where $f(0)$ is the probability at 0 under F. This estimate achieves the minimum sensitivity for M-estimators if the median of F_θ is 0. We now show that in fact it has optimal contamination sensitivity among all Fisher consistent functionals, and optimal contamination bias with respect to a specialized parameter discrepancy, if the median of F_θ is 0. Bishop, Feinberg and Holland [(1975), page 506] briefly considered the proportion-of-zeros estimate of the Poisson mean.

THEOREM 4.1. *If T is Fisher consistent for θ in the power series model, and if $m(t)$ is monotone increasing with derivative $m(t)$ for $t > 0$, then*

$$(4.5) \qquad b_{\log c(T)}(\varepsilon) \geq -\log(1-\varepsilon) \quad \text{and} \quad \gamma^*_{m(T)} \geq \dot{m}(\theta)c(\theta)/\dot{c}(\theta).$$

The proportion of zeros functional achieves the equalities in (4.5) *if* $c(\theta) \leq 2a(0)$.

REMARK 4.3 (Binomial p estimation continued). If $p \leq 1 - (1/2)^{1/v}$ the proportion-of-zeros functional has $\gamma^*_T = (1-p)/v$, which improves on the sensitivity of the empirical proportion by a factor of v.

REMARK 4.4 (Information standardized sensitivity). Hampel, Rochetti, Rousseeuw and Stahel (1986) defined the information-standardized sensitivity $J^{1/2}(\theta)\gamma^*_T = \gamma^*_{J^{1/2}(\theta)T}$, where $J(\theta)$ is the Fisher information. It is invariant to one-to-one differentiable transformations of θ. By (4.5), the information-standardized sensitivity of every Fisher-consistent estimator of either the Poisson mean or the binomial p has

$$\gamma^*_{J^{1/2}(\theta)T} \geq J^{1/2}(\theta)\frac{c(\theta)}{\dot{c}(\theta)} = \frac{\left\{E_\theta\left[X-\mu(\theta)\right]^2\right\}^{1/2}}{\mu(\theta)},$$

which blows up as $\theta \to 0$. In these cases, Fisher-consistent estimators with uniformly bounded information-standardized sensitivities do not exist.

4.3. *Comparison of bounds.* For power series families as a whole we have established that the minimum sensitivity for M-estimates is within a factor of 2 of the minimum for Fisher-consistent functionals. However, M-estimates can be much closer to the minimum. Information-standardized sensitivities versus the binomial p for $v = 5$ and $v = 10$ are shown in Figure 1(a) and (b). Also shown are the sensitivity bounds of (2.6) and (4.5). These are clearly complementary. The least sensitive M-estimator achieves the tighter of the two bounds if $x_0 = 0$ or $x_0 = vp$, and it is never very far off. The situation for the Poisson model is similar. In fact, whenever the mean of the Poisson or binomial distribution is an integer it is equal to the median [Kass and Buhrman (1980)], in which case γ^*_0 achieves the lower bound of (2.6).

5. Multiple linear regression. The conflict between bias minimax and locally linear estimation does not arise in the one-dimensional estimation problems of Sections 3 and 4. This section considers multiple linear regression, in which the conflict helps explain why certain high breakdown-point estimators such as S-estimators necessarily have unbounded contamination sensitivities. Moreover, as another application of Theorem 2.1, we obtain the bias bound for regression equivariant functionals reported by Maronna and Yohai (1991).

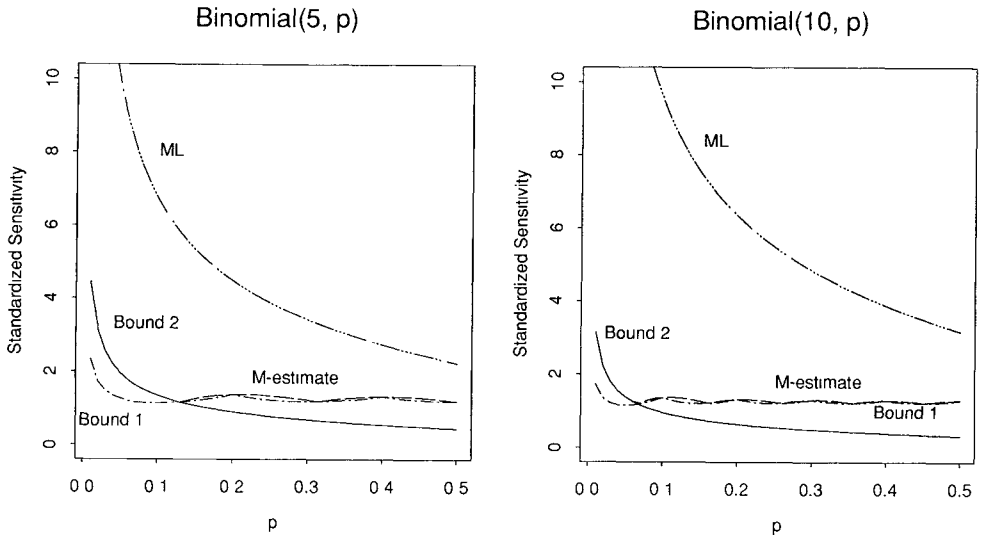

FIG. 1. *Information-standardized sensitivities for Binomial models*: *Bound* 1 *is from* (2.6) *and Bound* 2 *is from* (4.5).

Suppose F is the distribution function for a random vector $(X', U)'$, where X is p-dimensional and U is a scalar. Define the multiple linear regression family $F_\theta(x, u) := F_0(x, u - x'\theta)$, where the parameter vector θ takes values in p-dimensional Euclidean space. F_θ is the distribution of $(X', Y)' := (X', X'\theta + U)'$. We assume that, under the model, X and U are independent, and U has a continuous distribution symmetric about zero.

Regression equivariant functionals satisfy the relation $T(F_\theta) = T(F_0) + \theta$ for each p-vector θ. If we were to employ the ordinary Euclidean distance for the parameter vector, the resulting bias function would be invariant to θ; however, it would be affected by nonsingular transformations of X. We therefore employ the distance $\rho(T, \theta) := \{(T - \theta)'C^{-1}(F_X)(T - \theta)\}^{1/2}$, where C is a scatter functional that satisfies $C(F_{AX}) = AC(F_X)A'$ for each nonsingular $p \times p$ matrix A. The resulting invariance implies that we can assume $C(F_X) = I_p$. Martin, Yohai and Zamar (1989) and Maronna and Yohai (1991) employed the same strategy in studying robust functionals for multiple linear regression.

5.1. *Lower bounds for bias and sensitivity.* We first specialize the lower bound of Theorem 2.1 to the present setting. We need to solve for b_0 in

$$\inf_{\|z\|=1} d_v(F_0, F_{2b_0z}) = \frac{\varepsilon}{1 - \varepsilon}.$$

If we assume further that X has a spherically symmetric distribution, then the problem simplifies because then $d_v(F_0, F_{2bz})$ is invariant to the direction z.

In this case, the computations parallel those for regression through the origin. Let $z = e_1 := (1, 0, \ldots, 0)'$. Using (3.2) we obtain $d_v(F_0, F_{2b_0e_1}) = 2\int F_U(|X_1b_0|)\,dF_X(x) - 1$, so b_0 satisfies $E_0[F_U(|X_1b_0|)] = 2^{-1}(1-\varepsilon)^{-1}$. Solving for b_0 yields the bound given by Maronna and Yohai (1991) for regression equivariant functionals when X has a spherically symmetric distribution. If F_{X_1} and F_U are both standard normal, the bound simplifies to give $b_T(\varepsilon) \geq \tan(\pi\varepsilon/(2(1-\varepsilon)))$ and $\gamma_T^* \geq \pi/2$. Maronna and Yohai (1991) constructed a projection based functional that is equivariant, has a high breakdown point and is within twice the lower bound for contamination sensitivity.

To compute the local bound implied by Corollary 2.2 we evaluate γ_v in (2.6). Assume F_U has a differentiable density $f_U(u) > 0$. Under the linear model, the score function has the form $u_\theta(x, y) = \phi(y - x'\theta)x$, where $\phi(u) = -\dot{f}_U(u)/f_U(u)$. Hence,

$$\gamma_v := \{E_U|\phi(U)|E_X|X_0|\}^{-1},$$

where $X_0 := z_0'X$ and z_0 is the unit vector minimizing $E_X|z'X|$. We can simply set $X_0 = X_1$ if X is spherically symmetric. Locally linear regression functionals are subject to the sensitivity bound of Theorem 2.2, which in the present setting becomes $\gamma_T^* \geq p\{E_U|\phi(U)|E\|X\|\}^{-1}$.

EXAMPLE. Suppose X has a spherical standard normal distribution, and U is standard normal. Then the contamination sensitivity of an equivariant functional is at least $\gamma_v = \pi/2$. The sensitivity of a locally linear functional is at least $\sqrt{\pi}\,\Gamma((1/2)(p+2))/\Gamma((1/2)(p+1)) \geq (p\pi/2)^{1/2}$. The Maronna–Yohai functional has contamination sensitivity at most π, so it cannot be locally linear in dimensions higher than 5. Maronna and Yohai (1991) argued that the corresponding estimator is root-n consistent, but has a non-Gaussian limiting distribution.

5.2. *GM-estimates.* A *GM*-estimating functional $T(F)$ solves

$$E_F\eta(X, Y - X'\theta)X = 0$$

for some function $\eta(x, r)$. We adopt the regularity conditions on η given in Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 315], including the assumption that η is increasing in each argument so the estimate is Fisher consistent. The influence function of the *GM*-estimate at F_θ is given by

$$\psi_\theta(x, y) = \eta(x, y - x'\theta)M^{-1}x,$$

where $M = E_\theta\eta'(X, U)XX'$. Suppose the density function of U is continuously differentiable. A standard integration by parts argument shows that $E_\theta\psi_\theta(X, Y)u_\theta(X, Y)' = I_p$. By Remark 2.7, the *GM*-estimate functionals are locally linear.

A lower bound for the gross-error sensitivity of *GM*-estimators is given in Hampel, Ronchetti, Rousseeuw and Stahel [(1986), page 318]. It is the same as the bound implied by Theorem 2.2 in the regression setting. The bound is achieved by the *GM*-estimate with $\eta(x, r) = \operatorname{sgn}(r)x/\|x\|$ if X has a spheri-

cally symmetric distribution. Martin, Yohai and Zamar (1989) further showed that it has the optimal contamination bias among *GM*-estimators. Theorem 2.2 shows that this *GM*-estimate has the smallest contamination sensitivity among all locally linear estimators.

5.3. *S-estimates.* We next consider a class of functionals with contamination bias independent of the dimension. Rousseeuw and Yohai (1984) defined an S-estimator to be a vector $T(F)$ that minimizes the scale functional $s(t; F) > 0$ defined by $E_F\rho((Y - X't)/s(t; F)) = b$, where $t \in R^p$ and $0 < b < \sup \rho$. They assumed the following conditions for ρ:

(i) ρ is symmetric, has a continuous derivative ψ and $\rho(0) = 0$.

(ii) There exists a finite constant c such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

Under these conditions, S-estimators satisfy the first-order conditions of *GM*-estimators and have influence functions [Rousseeuw and Yohai (1984) and Lopuhaä (1989)], but their influence functions are unbounded in x and nonmonotone in the residual. The scale minimization is needed to ensure that a Fisher-consistent solution of the estimating equation is selected. S-estimators are locally linear, so their contamination sensitivities are subject to the dimension dependent lower bound of Theorem 2.2. Martin, Yohai and Zamar (1989) observed that the S-estimators have *dimension-free* bias functions. In view of the lower bound, S-estimates must have infinite contamination sensitivities. This fact is well known, but Theorem 2.2 shows that it is related to the general conflict between bias minimaxity and local linear estimation.

6. Further remarks. Focusing on the stability of the estimating functionals, we derived lower bounds for contamination bias that are tight or nearly tight across a broad range of parametric estimation problems. However, the estimates that achieve the bounds might not be suitable in practice, either because they sacrifice too much efficiency to achieve the optimal contamination bias, or they are insufficiently smooth to allow reliable inferences. Restricting to a class of locally linear functionals, we derived a tight lower bound on the contamination sensitivity. It is intriguing that such estimates cannot achieve the global bound for bias in higher dimensions. Estimates that improve on the sensitivity bound for locally linear estimates cannot be Fréchet differentiable, and their behavior close to the model may not be as predictable as one would like. See Donoho and Liu (1988a) for an example involving the minimum variation distance estimate. The most familiar classes of estimates such as M-estimates of location and scale, regression *GM*-estimates and S-estimates are locally linear, and outside this class one can expect limiting distributions to be non-Gaussian, as in the Maronna–Yohai (1991) example, and large sample inferences to be complicated.

The primary use of lower bounds for contamination bias is to provide a standard against which to measure the functional stability of compromise

estimators that take account of other measures of performance such as efficiency and stability of inference. The bounds we have established also enable us to extend many bias minimaxity results from a conveniently defined class of estimators to a broader class, as well as to understand the conflict between bias minimaxity and local linearity in higher dimensions.

Note added in proof. After final acceptance of this manuscript for publication it came to our attention that Reidel (1991) obtained independently a lower bound for contamination bias in parametric models generated by groups and showed that the bound is tight. Our Theorem 2.1 specializes to the same bound for equivariant functionals in group models.

7. Proofs.

PROOF OF THEOREM 2.1. Fix θ. Given $\eta \in \Theta$ set $\delta = \delta(\eta) := d_v(F_\theta, F_\eta)/(1 + d_v(F_\theta, F_\eta))$, so that

$$d_v(F_\theta, F_\eta) = \frac{\delta}{1-\delta}. \tag{7.1}$$

Let F_θ and f_η be densities of F_θ and F_η with respect to the dominating measure λ, and set

$$g := \frac{1-\delta}{\delta}(f_\eta - f_\theta)_+, \quad \text{and} \quad h := \frac{1-\delta}{\delta}(f_\eta - f_\theta)_-.$$

Using (2.2), $\int (f_\eta - f_\theta)_+ \, d\lambda = \int (f_\eta - f_\theta)_- \, d\lambda = d_v(F_\theta, F_\eta)$, which, along with (7.1), implies that both g and h are probability densities. Now $(1-\delta)f_\theta + \delta g = (1-\delta)f_\eta + \delta h$, so any functional T must satisfy $\rho(\theta, \eta) \le b_T(\delta(\eta); F_\theta) + b_T(\delta(\eta); F_\eta)$. Taking the supremum with respect to η, and observing that $d_v(F_\theta, F_\eta) \le c$ implies $\delta(\eta) \le c/(1+c)$, we have

$$\begin{aligned} b_v(\varepsilon/(1-\varepsilon); F_\theta) &\le \sup_{d_v(F_\theta, F_\eta) \le \varepsilon/(1-\varepsilon)} \left\{ b_T(\delta(\eta); F_\theta) + b_T(\delta(\eta); F_\eta) \right\} \\ &\le b_T(\varepsilon; F_\theta) + \sup_{d_v(F_\theta, F_\eta) \le \varepsilon/(1-\varepsilon)} b_T(\varepsilon; F_\eta) \\ &\le 2 \sup_{d_v(F_\theta, F_\eta) \le \varepsilon/(1-\varepsilon)} b_T(\varepsilon; F_\eta). \end{aligned}$$

The result follows because $d_v(F_\theta, F_\eta) \le c$ implies $\rho(\theta, \eta) \le b_v(c; F_\theta)$. □

PROOF OF PROPOSITION 2.2. This is immediate from Proposition 2.1 and Theorem 2.1. □

PROOF OF COROLLARY 2.2. We need to approximate the right-hand side of (2.3) for small ε. Using (2.2),

$$\begin{aligned} b_v(\varepsilon; F_\theta) &= \sup_b \left\{ b: \inf_{\|z\|=1} \int |f_{\theta+bz} - f_\theta| \, d\lambda \le 2\varepsilon \right\} \\ &= \sup_b \left\{ b: \inf_{\|z\|=1} b \int |z' u_\theta| f_\theta \, d\lambda + o(b) \le 2\varepsilon \right\} \\ &= \sup_b \{ b: b + o(b) \le 2\varepsilon \gamma_v(\theta) \} = 2\gamma_v(\theta)\varepsilon + o(\varepsilon). \end{aligned}$$ □

PROOF OF THEOREM 2.2. Let $\psi_\theta(x)$ be the influence function for T at F_θ. By Proposition 2.3 we have $\gamma_T^*(\theta) \ge \gamma_T^{(\mathrm{GE})} = \sup_x \|\psi_\theta(x)\|$, so it is sufficient to establish the bound for $\gamma_T^{(\mathrm{GE})}$. If $\gamma_T^{(\mathrm{GE})} = \infty$, we are done. So assume henceforth that ψ_θ is bounded.

By assumption, T is Fisher consistent and satisfies (2.7), so

$$\delta = \int \psi_\theta (f_{\theta+\delta} - f_\theta) \, d\lambda + o(\|\delta\|). \tag{7.2}$$

Moreover, as $\{F_\theta\}$ is L_1 differentiable,

$$\int \psi_\theta (f_{\theta+\delta} - f_\theta) \, d\lambda = \int \psi_\theta u_\theta' \, \delta f_\theta \, d\lambda + R_\theta(\delta), \tag{7.3}$$

where $\|R_\theta(\delta)\| \le \gamma_T^{(\mathrm{GE})} \int |f_{\theta+\delta} - f_\theta - \delta' u_\theta f_\theta| \, d\lambda = o(\|\delta\|)$. Equations (7.2) and (7.3) hold regardless of the direction of δ, so $E_\theta[\psi_\theta u_\theta'] = \int \psi_\theta u_\theta' f_\theta \, d\lambda = I_p$.

The remainder of the derivation is standard from the theory of influence functions:

$$p = \operatorname{tr} E_\theta[\psi_\theta u_\theta'] = E_\theta[u_\theta' \psi_\theta] \le E_\theta \|u_\theta\| \, \|\psi_\theta\| \le \gamma_T^{(\mathrm{GE})} E_\theta \|u_\theta\|.$$ □

PROOF OF PROPOSITION 4.1. Let $\mu(\theta)$ and $\sigma^2(\theta)$ be the mean and variance of the distribution F_θ. If the power series $\sum_{x=0}^{v} a(x)\theta^x$ is convergent, then both $\mu(\theta)$ and $\sigma^2(\theta)$ are finite and continuously differentiable. Moreover, as $f_\theta'(x)/f_\theta(x) = (x - \mu(\theta))/\theta$, the Fisher information $I(\theta) = E_\theta(f_\theta'(X)/f_\theta(X))^2$ is finite.

Given $\theta > 0$, consider sufficiently small $|\delta|$ such that $\sum a(x)(\theta + \xi)^x$ is convergent for all $|\xi| \le |\delta|$. By second-order Taylor expansion of $f_{\theta+\delta}(x)$ at $\delta = 0$,

$$\sum |f_{\theta+\delta}(x) - f_\theta(x) - \delta u_\theta(x) f_\theta(x)| \le \delta^2 \sum \sup_{|\xi| \le |\delta|} |\ddot{f}_{\theta+\xi}(x)|.$$

It is therefore sufficient to establish that the sum on the right is bounded as $\delta \to 0$. Direct calculation gives

$$\begin{aligned} |\ddot{f}_\theta(x)| &= \theta^{-2} f_\theta(x) |(x - \mu(\theta))^2 - 1 - \theta\mu'(\theta)| \\ &\le \theta^{-2} f_\theta(x) \{1 + \theta\mu'(\theta) + \mu^2(\theta) + 2x\mu(\theta) + x^2\}. \end{aligned}$$

For $x = 0, 1, 2, \ldots$, the function $\theta \mapsto \theta^x$ is monotone, so $\sup_{|\xi| \le |\delta|} f_{\theta+\xi}(x) \le f_{\theta+|\delta|}(x)c(\theta + |\delta|)/c(\theta - |\delta|)$. The result now follows from the continuity of $c(\theta)$, $\mu(\theta)$, $\mu'(\theta)$ and $\sigma^2(\theta)$. □

PROOF OF PROPOSITION 4.2. *Monotonicity of* $\beta(\theta)$. We first show that β is strictly increasing in θ. Let $\dot{\beta}$ be the derivative of β with respect to θ. By the definition of β,

$$0 = E_\theta \psi\left(\frac{X-\beta}{\kappa}\right) = \frac{d}{d\theta} E_\theta \psi\left(\frac{X-\beta}{\kappa}\right)$$

$$= \theta^{-1} E_\theta \psi\left(\frac{X-\beta}{\kappa}\right)(X-\mu) - \kappa^{-1}\dot{\beta} E_\theta \psi'\left(\frac{X-\beta}{\kappa}\right).$$

The assumptions on ψ_1 imply $E_\theta \psi'\{\kappa^{-1}(X-\beta)\} > 0$ and

$$E_\theta \psi\left(\frac{X-\beta}{\kappa}\right)(X-\mu) = E_\theta \psi\left(\frac{X-\beta}{\kappa}\right)(X-\beta) > 0,$$

so $\dot{\beta}$ is positive.

Continuity of T. Suppose $\liminf_{d_v(F_\theta, G)\to 0} T_\kappa(G) < \theta - \delta$ for some $\delta > 0$. As $E_\theta \psi\{\kappa^{-1}(X - \beta(\theta - \delta))\} > E_\theta \psi\{\kappa^{-1}(X - \beta(\theta))\} = 0$, there exists G such that $d_v(F_\theta, G) \le (1/4) E_\theta \psi\{\kappa^{-1}(X - \beta(\theta - \delta))\}$ and $T_\kappa(G) \le \theta - \delta$. Then $\beta(T) \le \beta(\theta - \delta) < \beta(\theta)$, and

$$0 = E_G \psi\left(\frac{X-\beta(T)}{\kappa}\right) \ge E_G \psi\left(\frac{X-\beta(\theta-\delta)}{\kappa}\right)$$

$$\ge E_\theta \psi\left(\frac{X-\beta(\theta-\delta)}{\kappa}\right) - 2d_v(F_\theta, G) \ge \frac{1}{2} E_\theta \psi\left(\frac{X-\beta(\theta-\delta)}{\kappa}\right),$$

a contradiction. Hence, $\liminf_{d_v(F_\theta, G)\to 0} T_\kappa(G) \ge \theta$. A similar argument shows that $\limsup_{d_v(F_\theta, G)\to 0} T_\kappa(G) \le \theta$, so T_κ is continuous.

Fréchet differentiability of T. Use the shorthand $T = T_\kappa(G)$ and $\psi_\eta(x) = \psi\{\kappa^{-1}(x - \beta(\eta))\}$. By definition of T and Fisher consistency, $\int \psi_T \, dG = \int \psi_T \, dF_T = \int \psi_\theta \, dF_\theta = 0$. It follows that

$$\int \psi_\theta \, d(F_\theta - G) = \int (\psi_T - \psi_\theta) \, d(G - F_T) + \int \psi_\theta \, d(F_\theta - F_T). \tag{7.4}$$

The first term on the right in (7.4) is dominated in absolute value by

$$\|\psi_T - \psi_\theta\|_\infty \cdot 2\big(d_v(F_\theta, G) + d_v(F_\theta, F_T)\big) = o\big(d_v(F_\theta, G)\big) + o(|T - \theta|).$$

The second term on the right in (7.4) is $(\theta - T)\int \psi_\theta u_\theta \, dF_\theta + o(|T - \theta|)$. Rearranging yields

$$T - \theta = \big(E_\theta[\psi_\theta u_\theta]\big)^{-1} \int \psi_\theta \, d(G - F_\theta) + o\big(d_v(F_\theta, G)\big) + o(|T - \theta|),$$

which, as T is continuous, implies T is Fréchet differentiable.

Monotonicity of γ_κ. To establish that $\gamma_\kappa^*(\theta)$ is increasing in κ, we show that the derivative of

$$E_\theta\psi\left(\frac{X-\beta}{\kappa}\right)(X-\mu(\theta)) = E_\theta\psi\left(\frac{X-\beta}{\kappa}\right)(X-\beta)$$

is negative. Let $\dot\beta$ be the derivative of β with respect to κ. Differentiating the relation

$$E_\theta\psi\left(\frac{X-\beta}{\kappa}\right) = 0$$

yields

$$\dot\beta = -\kappa^{-1}E_\theta\psi_1'\left(\frac{X-\beta}{\kappa}\right)\frac{X-\beta}{E_\theta\psi'((X-\beta)/\kappa)}.$$

Hence,

$$\begin{aligned}\frac{d}{d\kappa}E_\theta\psi\left(\frac{X-\beta}{\kappa}\right)(X-\beta) &= -\kappa^{-2}E_\theta\psi'\left(\frac{X-\beta}{\kappa}\right)\{(X-\beta)^2+\kappa\dot\beta(X-\beta)\}\\ &= \kappa^{-2}\left[\frac{\{E_\theta\psi'((X-\beta)/\kappa)(X-\beta)\}^2}{E_\theta\psi'((X-\beta)/\kappa)}\right.\\ &\qquad\left.-E_\theta\psi'\left(\frac{X-\beta}{\kappa}\right)(X-\beta)^2\right]\le 0,\end{aligned}$$

because $\psi' > 0$ and by the Cauchy–Schwarz inequality. □

PROOF OF PROPOSITION 4.3. For the power series family $E_\theta|u_\theta(X)| = \theta^{-1}E_\theta|X-\mu(\theta)|$. Hence, it suffices to prove that $E_\theta|X-\mu(\theta)| \le 2E_\theta|X-x_0|$. However, this is almost immediate: $E_\theta|X-\mu(\theta)| \le E_\theta|X-x_0| + |x_0-\mu(\theta)|$ and $|x_0-\mu(\theta)| = |E_\theta(X-x_0)| \le E_\theta|X-x_0|$. □

PROOF OF THEOREM 4.1. *Bias bound.* Fix θ. Define η by

$$c(\eta) = c(\theta)/(1-\varepsilon). \tag{7.5}$$

Observe that $\eta \ge \theta$, because c is increasing on Θ. Define $h(x)$ by $(1-\varepsilon)f_\theta(x) + \varepsilon h(x) = f_\eta(x)$. Clearly h must sum to one. It is also nonnegative because, using (7.5),

$$h(x) = \frac{1-\varepsilon}{\varepsilon}\left\{\left(\frac{\eta}{\theta}\right)^x - 1\right\}\frac{a(x)\theta^x}{c(\theta)} \ge 0.$$

If T is Fisher consistent, then $T((1-\varepsilon)f_\theta + \varepsilon h) = \eta$. Hence, $b_{\log c(T)}(\varepsilon) \ge \log\{c(\eta)/c(\theta)\} = -\log(1-\varepsilon)$.

Sensitivity bound. For small ε we have $-\log(1-\varepsilon) = \varepsilon + o(\varepsilon)$, and so

$$\varepsilon + o(\varepsilon) = \log\left\{\frac{c(\eta)}{c(\theta)}\right\} = \frac{\dot{c}(\theta)}{c(\theta)}(\eta - \theta) + o(\eta - \theta). \tag{7.6}$$

Moreover, if m is monotone increasing and differentiable,

$$b_{m(T)}(\varepsilon) \geq m(\eta) - m(\theta) = \dot{m}(\theta)(\eta - \theta) + o(\eta - \theta). \tag{7.7}$$

Combining (7.6) and (7.7) yields the bound for $\gamma^*_{m(T)}$ in (4.5).

Proportion of zeros functional. For any contaminating distribution H, let h_0 be the probability mass of 0. Let $T_\varepsilon = T((1-\varepsilon)F_\theta + \varepsilon H)$. Using (4.4),

$$\begin{aligned}\log c(T_\varepsilon) &= \log a(0) - \log\left\{(1-\varepsilon)\frac{a(0)}{c(\theta)} + \varepsilon h_0\right\} \\ &= \log c(\theta) - \log\left\{1 - \varepsilon + \varepsilon h_0 \frac{c(\theta)}{a(0)}\right\}.\end{aligned}$$

This has extremes at $h_0 = 0$ and $h_0 = 1$, so

$$b_{\log c(T)}(\varepsilon) = \max\left\{-\log(1-\varepsilon), \log\left(1 + \varepsilon\left(\frac{c(\theta)}{a(0)} - 1\right)\right)\right\}. \tag{7.8}$$

The maximum in (7.8) is $-\log(1-\varepsilon)$ if

$$c(\theta) \leq \frac{2-\varepsilon}{1-\varepsilon} a(0). \tag{7.9}$$

As $(2-\varepsilon)/(1-\varepsilon)$ is increasing in ε, (7.9) holds for each $\varepsilon > 0$ if $c(\theta) \leq 2a(0)$. □

Acknowledgments. The authors thank an Associate Editor and referees for constructive comments. A referee suggested we consider the scale model.

REFERENCES

BISHOP, Y., FEINBERG, S. and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. MIT Press.

DONOHO, D. L. and LIU, R. C. (1988a). The 'automatic' robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586.

DONOHO, D. L. and LIU, R. C. (1988b). Pathologies of some minimum distance functionals. *Ann. Statist.* **16** 587–608.

FERNHOLZ, L. T. (1983). *von Mises Calculus for Statistical Functionals. Lecture Notes in Statist.* **19**. Springer, New York.

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. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *ASA 1978 Proceedings of the Statistical Computing Section* 59–64, Amer. Statist. Assoc., Washington, D.C..

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

HE, X. (1991). A local breakdown property of robust tests in linear regression. *J. Multivariate Anal.* **38** 294–305.

HE, X. and SIMPSON, D. G. (1992). Robust direction estimation. *Ann. Statist.* **20** 351–369.

HE, X., SIMPSON, D. G. and PORTNOY, S. (1990). Breakdown robustness of tests. *J. Amer. Statist. Assoc.* **85** 446–452.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

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

JOHNSON, N. L. and KOTZ, S. (1968). *Discrete Distributions*. Houghton Mifflin, Boston.

KASS, R. and BUHRMAN, J. M. (1980). Mean, median and mode in binomial distributions. *Statist. Neerlandica* **34** 13–18.

KRASKER, W. S. (1980). Estimation in 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.

KÜNSCH, H. R., STEFANSKI, L. A. and CARROLL, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.* **84** 460–466.

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

MARONNA, R. A. and YOHAI, V. J. (1991). Recent results on bias-robust regression estimates. In *Directions in Robust Statistics and Diagnostics, Part I* (W. Stahel and S. Weisberg, eds.) 221–232. Springer, New York.

MARTIN, R. D. and ZAMAR, R. H. (1989). Asymptotically min-max bias robust M-estimates of scale for positive random variables. *J. Amer. Statist. Assoc.* **84** 494–501.

MARTIN, R. D., YOHAI, V. J. and ZAMAR, R. H. (1989). Min-max bias robust regression. *Ann. Statist.* **17** 1608–1630.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

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

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. In *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle and R. D. Martin, eds.). Springer, New York.

RUPPERT, D. (1985). On the bounded-influence regression estimator of Krasker and Welsch. *J. Amer. Statist. Assoc.* **80** 205–208.

SIMPSON, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82** 802–807.

SIMPSON, D. G., CARROLL, R. J. and RUPPERT, D. (1987). M-estimation for discrete data: asymptotic distribution theory and implications. *Ann. Statist.* **15** 657–669.

STEFANSKI, L. A., CARROLL, R. J. and RUPPERT, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **73** 413–425.

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates 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.

DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 0511

DEPARTMENT OF STATISTICS
101 ILLINI HALL
UNIVERSITY OF ILLINOIS
725 SOUTH WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820