for some (gaussian?) noise sequence ω_t. Thus the unobservable AR process Z_t, provides the "success" probability for S_t after transformation. Generalisations to higher order processes, transforms other than the log-odds, and time-varying parameters (ϕ_t rather than simply ϕ), are evident. The outlier model (1) can now be extended to this binary series by a minor extension of (2) to

$$Z_t = X_t + \delta_t,$$
$$X_t = \phi X_{t-1} + \omega_t,$$

incorporating changes via ω_t series, patchy outliers, and, now, observational outliers through appropriate models for the δ_t series. The only point of significant difference between this model and (1) is that the sampling model is now Bernoulli, rather than gaussian, which leads to a slightly different view of the way in which observational outliers are generated. A closely related, but structurally quite different, class of models for binary series provides for dynamic evolution of transition probabilities in Markov chains. The first order case, for example, has a basic model for $P(S_t = 1|\pi_t)$ as above, but, instead of the continuous process model for the log-odds probability Z_t in (2), a discrete version

$$Z_t = \theta_t + \phi_t S_{t-1} + \omega_t, \tag{3}$$

where θ_t and ϕ_t are time-varying process parameters and ω_t, as usual, process evolution noise. Concerning outlier models, a basic problem arises with (3) in that the observations S_t are fed back into the process model, so a little more thought is required in modelling pure observational outliers. Perhaps the authors have some comments on such problems.

REFERENCES

WEST, M. (1986). Bayesian model monitoring. *J. Roy. Statist. Soc. Ser. B* **48** 70–78.

WEST, M. and HARRISON, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. To appear in *J. Amer. Statist. Assoc.*

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Assoc.* **80** 73–97.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY CV4 7AL
ENGLAND

REJOINDER

R. DOUGLAS MARTIN AND VICTOR J. YOHAI

University of Washington and University of Buenos Aires
and CEMA, Buenos Aires

The discussants have provided us with more than ample food for thought concerning a myriad of issues related to our work on influence functionals for time series. Leading issues include the following: (1) Relationships and differences between **ICH** and **IF**. Aspects of this include (a) the model-dependent feature of **IF**, and (b) the desirability or lack thereof of averaging inherent in the definition of **IF**; (2) data-oriented measures of influence; (3) generalizations and/or modifications of the **IF** to cover prediction, spectral estimates, nonstationarity, testing, etc.; (4) lack of consistency and bias control; (5) robust model selection; and (6) innovations outlier models, intervention, adaptivity, etc.

Relationships and differences between ICH and IF. *Künsch* has provided a transparent heuristic formula which displays the relationship between **ICH** and **IF** for time series, and allows one to see clearly why a difficulty can arise when **T** does not depend only on a finite dimensional marginal measure: namely, boundedness of $\tilde{\boldsymbol{\Psi}}$ does not necessarily yield boundedness of **IF**. This is an enlightening viewpoint, which gets at a problem special to the time-series setting in short order. Also, *Künsch's* suggestion to split Theorem 4.2 up in the manner of his Theorem 4.1′ is a useful and welcome contribution. It is indeed desirable to state a main result in a form which is as free as possible from model-specific assumptions, and therefore facilitates a wider range of applications of the **IF**.

Künsch's main concerns with our approach are that **IF** involves (too much) averaging (of **ICH**), and that the **IF** and **GES** are determined by contamination models which are too specific and narrow.

The basis of *Künsch*'s complaint about model specificity is evidently his view that one seldom knows the type of contamination in advance. However, it is our experience that on the contrary, for many time series arising in practice the investigator does indeed have a pretty good idea of whether patchy or isolated outliers are to be encountered (and perhaps a little detail is available concerning qualitative patch shape, but not too much else.) Indeed, *Wegman* has indicated the viability of the forms already included in (2.2) for a variety of real-world applications, where either patchy or isolated outliers are expected. In the case of radar glint noise, for example (Figure 14 of Martin and Thomson, 1982), the outliers consist of spikes having a moderately consistent shape, with random amplitude and separations which are approximately independent exponential random variables. In target tracking contexts, the amplitudes will be negligible and hence there will be no outliers at far range situations, whereas the amplitudes will be large and the outliers will be quite potent at close ranges. Aside from the varying amplitudes and separation times, the structure of the outlier model is relatively constant.

Given that one can in many cases be relatively confident of outlier type, the model dependent aspect of **IF** is highly desirable. In such cases, **IF** and **GES** calculations, and optimality results over "narrow" classes can help indicate what type of estimate is preferred with regard to its infinitesimal bias control. Furthermore, in case different specific types of outliers are considered likely to appear, one may use the corresponding **GES** criterion to select the estimate instead of the **IF**.

Of course when one is indeed relatively ignorant of the type of outliers to be encountered, then optimality over narrow classes is indeed of little use, and

Künsch's suggestions concerning the choice of **P** in Definition 6.1 may be appropriate. Furthermore, under complete ignorance one might be happy with the conservative/pessimistic approach of optimally bounding **GESH** $\triangleq \sup|\mathbf{ICH}|$ in the time series case as Künsch (1984) has done.

With regard to the pure replacement (PR) versus additive outliers (AO) issue, *Künsch*'s PR calculations show that the Huber function can be preferred over the bisquare function (which is opposite to the result of our AO calculation). However, it is our opinion that the PR model is seldom appropriate in practice. We are not thereby suggesting that AO is always appropriate. It is just that the value of a time series at an outlier position will usually contain some shadow or vestige of the core process, and in these cases AO will often be a better approximation than PR. Furthermore, AO will often be quite a good approximation—this is true, for example, in the glint noise example cited above, and it is certainly true in those situations for which intervention analysis (Box and Tiao, 1975) is appropriate. Also, it appears feasible to construct statistics which discriminate between PR and AO.

In any event, we feel that the understanding gained by the calculation of **IF**'s for different estimators at different contamination models yields insight concerning the interplay between different estimators and different contamination models which is useful in its own right. Many more such calculations remain to be carried out. Such calculations can help resolve the kind of question raised by *Franke* and *Hannan*: How much does **IF** depend on different contamination models having qualitatively similar sample paths? We pursue this question with respect to their particular model (1).

Although model (1) does not quite fit into the general model (2.2), it does fit into the following slight generalization:

$$y_i^\gamma = (1 - z_i^\gamma)x_i + z_i^\gamma w_i^\gamma, \tag{2.2'}$$

where now the contaminating process w_i^γ has a distribution depending on γ. In order to get model (1) we can take z_i^γ as in (2.4) with $\gamma = kp$ and $\tilde{z}_i^\gamma = g(\varepsilon_i)$ where $g(t) = 1$ for $t \neq 0$ and $g(0) = 0$, and then set $w_i^\gamma = x_i + \sum_{j=0}^{k-1}\beta_j\varepsilon_{i-j}$. Since the distribution of the ε_i's depends on p and therefore on γ, we need (2.2′) instead of (2.2).

The definition (4.5) of **IF** can be extended without modifications to the more general model (2.2′). In the case of model (1) with x_i an AR(1) process we get for GM estimates

$$\begin{aligned}\mathrm{IF}\big(\{\mu_w^\gamma\}, T_{\mathrm{GM}}, \{\mu_y^\gamma\}\big) = \sum_{j=1}^{k-1} E\eta\big(u_1 + (\beta_j - \phi\beta_{j-1})\varepsilon_0, h(\phi)\phi(x_0 + \beta_{j-1}\varepsilon_0)\big)\\ + E\eta\big(u_1 - \phi\beta_j\varepsilon_0, h(\phi)(x_0 + \beta_j\varepsilon_0)\big),\end{aligned}$$

where $h(\phi) = (1 - \phi^2)^{1/2}$. In the case of $\beta_0 = \cdots = \beta_{j-1} = 1$, we get

$$\begin{aligned}\mathrm{IF}\big(\{\mu_w^\gamma\}, T_{\mathrm{GM}}, \{\mu_j^\gamma\}\big) = (k-1)E\eta\big(u_1 + (1 - \phi)\varepsilon_0, h(\phi)(x_0 + \varepsilon_0)\big)\\ + E\eta\big(u_1 - \phi\varepsilon_0, h(\phi)(x_0 + \varepsilon_0)\big),\end{aligned}$$

which is very close to our formula (5.8) for patchy additive outliers. This is reassuring, since, as observed by *Franke* and *Hannan*, the qualitative behavior of model (1) will be very similar to our case of patchy additive outliers, provided that all the β_i have the same sign.

The issue of whether or not **IF** involves too much averaging is a most basic one. It is primarily when the number of outliers is small, which corresponds on average to γn small, that the averaging-based **IF** is unlikely to provide a good indication of the influence of outliers on an estimate in finite samples. Such situations are indeed troublesome, for neither sup|**ICH**| nor **IF** is likely to give a uniformly accurate assessment across different configurations of outliers. On the other hand, such situations are precisely where totally data-oriented measures of influence for time series, such as that suggested by *Brillinger*, may come into their own. Of course when the number of outliers is moderate to large, one must use an appropriate averaging in order that **IF** adequately reflect the influence of the outliers. Analysis of the data at hand, aided by any reasonably robust method, will often provide useful guidance here.

Data-oriented measures of influence. *Brillinger*'s suggestions concerning "leave-one-out" diagnostics/influence measures for time series are highly appropriate, both for their own sake and for the complementary nature the techniques have relative to **ICH** and **IF**. The "leave-one-out" approach is tailor-made for *Künsch*'s "single outlier in a series of length *n*." This natural data-oriented approach for time series has been neglected for so long, in spite of Brillinger's (1966) proposal, only because the method is fairly computing intensive, and (as *Brillinger* points out) good algorithms for handling missing data problems in time series have been developed only relatively recently (see *Brillinger*'s references). There are, however, some issues concerning leave-one-out diagnostics for time series which should be mentioned.

First of all, there is a clear *smearing* effect associated with the influence of isolated outliers in the leave-one-out approach. This effect is evident in *Brillinger*'s Figure 3: Adjacent to each "large" peak there are one or two values of roughly half the local amplitude of the peak. The reason for this behavior is inherent in the Gaussian maximum-likelihood leave-one-out technique, and it can conceivably give a false indication of a single dominant outlier when in fact there are two outliers separated by a single good point.

Another, perhaps more serious difficulty, is that leave-one-out diagnostics can fail to give an indication of problems when the outlier is of a "k-in-a-row" patch form. This is a special form of what has been called the "masking" problem in the regression diagnostics literature. In the regression setting the masking problem has been relatively ignored due to the computational burden required to check for masking in unstructured problems, namely order $\binom{n}{k}$ when k well-masked outliers are present. However, in the structured time-series setting we can easily detect masking due to a single patch of length k in order n by computation of "leave-k-out" diagnostics whereby $y_{i+1}, y_{i+2}, \dots, y_{i+k}$ are deleted and a Gaussian missing-data MLE is used to fit the model for $i = 0, 1, \dots, n - k$.

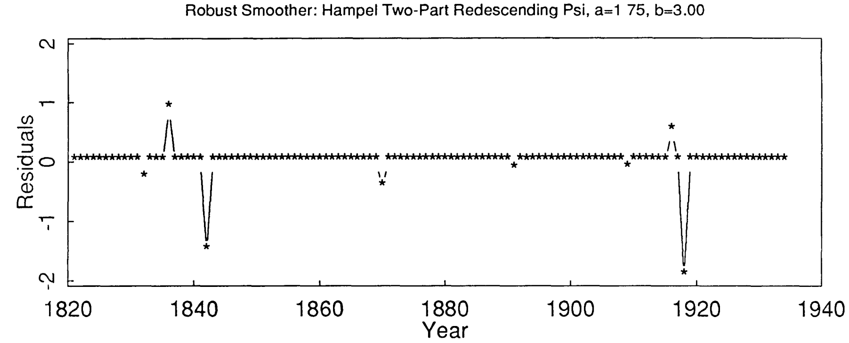

FIG. 1. *Residuals $y_t - \hat{x}_t$ for Canadian lynx data, where $\hat{x}_t$ is smoother-cleaned value based on robust ARMA* (3, 3) *fit.*

Of course one cannot completely solve the masking problem with any degree of computational ease. The existence of more than one patch, or more than one isolated outlier, or both, can result in masking which can only be completely dealt with by an order of computational complexity which approaches that of unstructured regression problems. It is possible that some data sets may contain too many configurations of outliers to effectively cope with. Fortunately, experience indicates that: (i) complete masking does not occur with great frequency, even with multiple patches and isolated outliers, and (ii) iterative interpolation of the most influential data points will often reveal other influential points which are initially masked.

Details concerning some of the various claims made above are provided in Bruce and Martin (1986).

In answer to Brillinger's question to us: There exist good robust techniques for fitting ARIMA models to data with many outliers, based on robust filter-cleaners or smoother-cleaners (see for example Martin, 1981, or Martin, Samarov and Vandaele, 1983), and these techniques along with their robust residuals diagnostics should always be used (along with other methodologies, including leave-k-out diagnostics) when one is not absolutely sure that the data is outlier free. These residuals diagnostics generally give a much clearer view of outlier structure than leave-k-out diagnostics, and take considerably less computational time (our current version of "leave-k-out" is still too slow to be really pleasant on a PC). Figure 1 shows the residuals $y_t - \hat{x}_t$ for a robust fit of an ARMA(3, 3) model to the Canadian lynx data. Since $\hat{x}_t = y_t$ for "good" data points, most of the residuals are zero. The nonzero residuals indicate nearly the same "suspect" data points as those revealed by Brillinger's plot.

In summary: while "leave-k-out" diagnostics appear to have a useful role in time series analysis, procedures based on robust filter- or smoother-cleaners would be preferred if just one of the two techniques were to be used.

Generalizations of IF. The issue of generalizations and/or applications of the **IF** to problems such as prediction, testing, spectral estimation and long-memory processes have been touched upon by *Franke* and *Hannan*, *Robinson*, *Tsay*, and *Poor*. Although Sections 8.1 and 8.2 make small contributions to **IF**'s for spectral estimates and tests, there remain a number of questions to be pursued. At the moment we are able to respond to some of the specific questions raised by these discussants.

Consider the case of an ARMA-type spectral estimate. Let $S(\mathbf{T}(\mu_y^\gamma)) = S(f;\mathbf{T}(\mu_y^\gamma))$ denote the asymptotic functional representation of an ARMA-type spectral estimate, where

$$\mathbf{T}(\mu_y^\gamma) = \left(\phi_1(\mu_y^\gamma),\dots,\phi_p(\mu_y^\gamma),\theta_1(\mu_y^\gamma),\dots,\theta_q(\mu_y^\gamma),s^2(\mu_y^\gamma)\right)$$

with $\mathbf{T}(\mu_x) = \boldsymbol{\alpha}$ the parameters of an x_i process ARMA model and $S(\boldsymbol{\alpha})$ the corresponding spectral density. Then it is straightforward to calculate a pointwise influence functional $\mathbf{IF}_S(\mu_w, f, \mathbf{T})$ for S. Use of the chain rule gives:

$$\mathbf{IF}_S(\mu_w, f, \mathbf{T}) = \left[\frac{dS(\mathbf{T}(\mu_y^\gamma))}{d\gamma}\right]_{\gamma=0} = \left[\frac{dS(\mathbf{T})}{d\mathbf{T}}\right]_{\mathbf{T}=\boldsymbol{\alpha}} \cdot \left[\frac{d\mathbf{T}(\mu_y^\gamma)}{d\gamma}\right]_{\gamma=0},$$

and the first factor of the right-hand side does not depend on $\mathbf{T}$ (but does depend upon frequency f). Therefore in order to compare the influence curve of two estimators $S(\mathbf{T}_1)$ and $S(\mathbf{T}_2)$ of $S(\boldsymbol{\alpha})$, we only need to compare the influence curves of $\mathbf{T}_1$ and $\mathbf{T}_2$. Hence, in answer to *Franke* and *Hannan*, robustness of ARMA model parameter estimates determined by **IF** properties is inherited by an ARMA-type spectral estimate, albeit with a frequency dependent weighting factor. Essentially the same is true if one computes $\mathbf{IF}_{\log S}$ for $\log S$.

Of course, $S(\mathbf{T})$ may not be a good estimate of the functional S in the case where x_t does not conform to a parametric ARMA model, but the main issue in such cases is parametric approximation rather than robustness of $\mathbf{T}$.

Unfortunately, nonparametric robust spectral density estimates, such as those involving robust prewhitening described in Kleiner, Martin, and Thomson (1979), are much more complicated, and simulation will be required to determine an $\mathbf{IF}_S$ or $\mathbf{IF}_{\log S}$ in such cases.

Franke and *Hannan* note that the computed $\mathbf{IF}_S(\zeta, f)$ in Section 8.2 does not depend on frequency. The reason for this is that the contamination is white noise, along with the averaging involved in the **IF** (see the earlier comments by both *Künsch* and ourselves). To get a measure of influence which will show the frequency-dependent effects associated with specific outlier patterns, one can either take a data-oriented approach as suggested by *Brillinger* and discussed above, or pursue the **IF** approach with an appropriate contamination model. In the data-oriented approach one could use the leave-k-out technique to fit a good AR or ARMA model and interpolate at the deleted points—the difference between spectrum estimates based on the original data and those based on "leave-k-out and interpolate" modifications of the data will show frequency-dependent influence (however, this data-oriented approach may often involve an embarrassing amount of computation). In the **IF** approach, μ_w might be selected so as to generate outliers resembling those seen in the data, e.g., in the waveguide of Kleiner, Martin, and Thomson (1979), the outliers might come in pairs having a fixed separation, but with random separation from pair to pair—the resulting $\mathbf{IF}_S(\zeta, f)$ will depend upon f. (Incidentally, such possibilities suggest how **IF** may give a useful indication of how an estimate reacts to a specific configuration of outliers in the data at hand.)

Franke and *Hannan* have also raised the question of how stable the **IF** is when the nominal model is known only approximately. Here stability is equivalent to asking for some kind of continuity of **IF** with respect to μ_x. If we use the weak topology, continuity of **IF** is closely related to robustness (Hampel, 1971; Boente, Fraiman, and Yohai, 1982). According to (4.2) and (4.6)

$$\mathbf{IF}\left(\mu_w, \mathbf{T}, \{\mu_y^\gamma\}\right) = -\mathbf{C}^{-1} \lim_{\gamma \to 0} \frac{E(\tilde{\boldsymbol{\Psi}}(\mathbf{y}_1^\gamma, \mathbf{t}_0))}{\gamma},$$

with $\mathbf{C}$ given by (4.2′). If $\tilde{\boldsymbol{\Psi}}$ is bounded and depends only on a finite number k of coordinates, then the second factor on the right-hand side will depend continuously (with respect to the weak topology) on the corresponding k-marginal distribution of the nominal process x_t. Similar results occur when $\tilde{\boldsymbol{\Psi}}$ depends on an infinite number of coordinates, but this dependence decreases quickly enough, e.g., as with GM and RA estimates. However, typically the behavior of $\mathbf{C}$ will be different. For example, in the case of GM and RA estimates, $\mathbf{C}$ depends on the first moment of x_i when the η function is of the Mallows type and on the second moment of x_t when it is of the Hampel type. Thus in order to have continuity of **IF** with respect to μ_x at a nominal model $\mu_{x,0}$ one may have to use a metric which implies closeness of the moments for μ_x and $\mu_{x,0}$.

Künsch, *Poor* and *Tsay* all raise more or less directly a quite important question: How does one deal with nonstationarity in the central model μ_x, in deviations from the central model, or in both?

In order to be as general as possible, one might proceed as follows. Let $\{T_n\}$ be a sequence of univariate estimators indexed by sample size n, and let T_n^γ denote the value of T_n for the contaminated process y_i^γ. The arc $\{\mu_y^\gamma\}$ may now be nonstationary by virtue of one or more of the measures $\mu_x, \mu_w, \{\mu_z^\gamma\}$ being nonstationary. Then define the (absolute-value) influence functional for nonstationary processes as

$$\mathrm{IF}_{\mathrm{ns}}\left(\mu_w, \{T_n\}, \{\mu_y^\gamma\}\right) = \lim_{\gamma \to 0} \frac{1}{\gamma} E_{\mu_y^\gamma}\left(\limsup_{n \to \infty} |T_n^\gamma - T_n^0|\right).$$

Of course for many cases of interest, including the nonstationarity examples presented by *Künsch* and *Poor*, one will have $T_n^\gamma \to T(\mu_y^\gamma)$ and $T_n^0 \to T(\mu_y^0)$ almost surely. In such cases we have $\limsup_{n \to \infty} |T_n^\gamma - T_n^0| = |T(\mu_y^\gamma) - T(\mu_y^0)|$, the expectation is superfluous, and $\mathrm{IF}_{\mathrm{ns}} = |\mathrm{IF}|$. Correspondingly, given a family $\mathbf{P}$ of nonstationary arcs $\{\mu_y^\gamma\}$ we define the gross-error sensitivity:

$$\mathrm{GES}(\mathbf{P}, \{T_n\}) = \sup_{\{\mu_y^\gamma\} \in \mathbf{P}} \mathrm{IF}\left(\mu_w, \{T_n\}, \{\mu_y^\gamma\}\right).$$

Similar definitions may be given for the multivariate case.

For some types of contamination, nonstationary process $\{w_t\}$ are not more harmful than stationary $\{w_0\}$ in terms of **GES**. For example, consider the AR(1) model with independent and additive outliers, i.e., $w_i = x_i + v_i$, and suppose that η satisfies (A1)–(A4), along with

(A5) $\eta(u, v)$ is monotone in each variable and $uv > 0$ implies $\eta(u, v) \geq 0$.

Then we can prove that the GES for GM and RA estimates when $\{v_i\}$ is allowed to be nonstationary is the same as when it is restricted to being stationary, and is given by Theorem 5.1 of Martin and Yohai (1984b). Further study is needed to determine the extent to which similar results may be true for other models and different types of contamination.

Concerning robust tests, *Robinson*'s suggestion regarding robustified score tests is a good one which has been recently pursued (Basawa, Huggins, and Staudte, 1985). In fact, the general area of formal inference for robust procedures is in need of more attention not only in the time-series setting, but also in the more classical contexts such as linear regression, etc. However, the construction of useful finite-sample tests and confidence intervals has proved difficult enough in the non-time-series setting, and the problem can hardly be any easier for the time-series setting.

Franke and *Hannan*, and *Tsay* are quite correct in pointing out that one should consider the purpose of the analysis when defining influence functionals. Thus if one is concerned about prediction, then one should use an appropriate influence functional IF_p for prediction.

If one is willing to focus on prediction based on the "good" data x_ι, then the following definition would be suitable. Consider the autoregression context: Let $\phi(\mu_y^\gamma)$ denote the functional representation of the parameter estimates, and let $\mathbf{x}_\iota^T = (x_{i-1}, \dots, x_{\iota-p})$; suppose we use the linear predictor $\hat{x}_\iota(\mu_y^\gamma) = \mathbf{x}_\iota^T \phi(\mu_y^\gamma)$. Then

$$\mathrm{IF}_p = \frac{d}{d\gamma} \hat{x}_\iota(\mu_y^\gamma)|_{\gamma=0} = \mathbf{x}_i^T \mathbf{IF}$$

where $\mathbf{IF} = \mathbf{IF}(\mu_w)$ is the influence functional for $\phi(\mu_y^\gamma)$. It may be convenient to use the square root of the average squared value of IF_p:

$$\mathrm{AIF}_p(\mu_w) = \left(E_{\mu_x} \mathrm{IF}_p^2\right)^{1/2} = \left(\mathbf{IF}^T \mathbf{C}_x \mathbf{IF}\right)^{1/2},$$

where $\mathbf{C}_x$ is the $p \times p$ covariance matrix of x_ι. It is easy to check that $\gamma^2 \mathrm{AIF}_p^2(\mu_w)$ is the squared-bias component of prediction mean-squared-error for small γ: $\sigma_{\mathrm{MSE}}^2 \cong \sigma_\varepsilon^2 + \gamma^2 \mathrm{AIF}_p^2(\mu_w)$.

However, one would be considerably more interested in a measure of influence for robust predictors which reflects the effect of outliers in $(\mathbf{y}_i^\gamma)^T = (y_{i-1}^\gamma, \dots, y_{i-p}^\gamma)$ used as predictor variables, as well as the effect of outliers on the parameter estimates. Correspondingly, we expect a robust predictor of the core value x_i to have the nonlinear asymptotic form

$$\hat{x}_i = g\left(\mathbf{y}_{i-1}^\gamma, \phi(\mu_y^\gamma)\right),$$

where $\mathbf{y}_{i-1}^\gamma = (y_{i-1}^\gamma, y_{i-2}^\gamma, \dots)$. Predictors based on joint robust filter-cleaners and AM-type estimation (Martin, 1981; Martin and Yohai, 1985) will have such a form. Then one might define

$$\mathrm{IF}_p = \frac{\partial}{\partial\gamma} E_{\mu_y^\gamma} g^2\left(\mathbf{y}_{i-1}^\gamma, \phi(\mu_y^\gamma)\right)|_{\gamma=0}.$$

Poor is interested in the possible use of **IF**'s in connection with long-memory processes. Since long-memory processes of the type mentioned by *Poor* do not result in asymptotic bias for most parameter estimates, the **IF** is not a useful tool for assessing the influence of such long-memory processes. For parameter estimation problems where the rate of convergence can be maintained in the face of variance inflation due to long-memory contamination (the notable case where this is not so being that of the sample mean), perhaps an analogue of the change-of-variance curve CVC (Hampel, Rousseeuw, and Ronchetti, 1981; Hampel et al., 1986) would be a useful tool.

Lack on consistency and bias control. *Robinson* has made a number of interesting comments and suggestions having to do with the issues of asymptotic bias and the second-order structure of time-series contamination models. First of all we should recall that the spirit of robustness is that of doing well *near* a parametric model (see Huber, 1981; Hampel et al., 1986). In terms of contamination models "near" means not too large a fraction γ of contamination, but the contamination can be arbitrarily bad when it occurs. Obviously quite small γ in our (2.2) can give rise to m_y and $c_y(j)$'s that are quite far from the m_x and $c_x(j)$'s in *Robinson* (1)–(4). On the other hand, when γ is small, the measures μ_x and μ_y^γ will be close in metrics which are suitable for robustness in the time-series setting (see for example Boente, Fraiman, and Yohai, 1982). For this reason the second-order viewpoint is not too appealing.

With a view toward asymptotics one can of course put down a richer, more accurate class of models, perhaps from a second-order point of view as suggested by *Robinson*, and then estimate everything in sight. However, some caveats are in order. In the first place, the fact that we may have some knowledge of what type of outliers may occur does not exclude the possibility that other types may occur which we do not anticipate, and hence one may find it difficult to specify a sufficiently rich model. Furthermore, we have not run into many situations where the sample size is sufficient to render estimation of a rich outlier model a practically realizable goal.

On the other hand there do seem to be many applications where the sample size is nonetheless sufficiently large that squared bias will be the dominating component of mean-squared error. In such situations bias control is a dominant robustness consideration. Hampel's approach of optimally bounding the influence curve, pursued by Künsch (1984) in the autoregression context, takes a significant step toward obtaining analytic results with regard to bias control. However, one must remember the **ICH** and **IF** are infinitesimal in nature, as are optimality results based on them. Global robustness results are also highly desirable.

To date the main focus of global robustness has been on the breakdown point (see Hampel, 1971; Huber, 1981; and Hampel et al., 1986, for definitions). Indeed, the problem of constructing (and computing) high breakdown point estimates has been a lively area of research in recent years (see for example Rousseeuw and Yohai, 1984; Hampel et al., 1986; Yohai, 1985; Yohai and Zamar, 1985). High breakdown point estimates having high efficiency may well provide the preferred approach in areas such as robust regression.

On the other hand, global bias optimality results have received little attention. One approach to global bias optimality is to define an optimal bias robust estimate as one which minimizes the maximum asymptotic bias for a given fraction of contamination γ. In spite of Huber's (1964, 1981) proof that the sample median has this property (see also Section 2.7 of Hampel et al., 1986), the approach has been essentially neglected. From recent results (Martin and Zamar, 1985; Zamar, 1985) it appears that min–max bias robust estimates, both with and without an efficiency constraint at the Gaussian model, can be obtained in situations such as estimating location, scale, and regression parameters with independent observations. It is hoped that one can obtain similar analytical bias robust solutions in the time-series setting. Again, the issue of how large a class $\mathbf{P}_\gamma$ of contaminating measures one should use will arise. Both relatively narrow and quite broad classes should be considered, in correspondence with a practitioner's state of knowledge.

We concur with *Robinson* that the identification problem should be taken seriously, but our emphasis in this area would be somewhat different than his, as reflected in the following comments on robust model selection.

Robust model selection. *Franke* and *Hannan*, *Robinson* and *Tsay* all raise the issue of the interplay between model fitting and robustness, and implicitly this raises the issue of robust model selection. This is a thorny issue concerning which there is a notable lack of understanding, even in the ordinary regression setting.

As *Robinson* has aptly pointed out, an (arbitrarily small) PR- or AO-type contamination results in a more complicated model. A pure autoregression becomes an ARMA model, and an ARMA model becomes an ARMA model with a higher-order moving average component, etc. Similar effects will be caused by almost any kind of contamination. The basic point is that in arbitrarily small neighborhoods of an ARMA (p_0, q_0) Gaussian model there will be non-Gaussian ARMA $(p q)$ models with p, q arbitrarily large—and as *Robinson* points out, the covariances of the ARMA (p, q) model may be quite far from those of the ARMA (p_0, q_0) model.

As a consequence, one cannot, for example, expect any robust procedure to asymptotically fit a finite-order autoregression to a contaminated time series y_i^γ in which x_i is AR(p_0). However, GM, RA and (probably) AM estimates are qualitatively robust in the AR case, i.e., a small fraction of contamination γ will produce only small biases (a proof for GM estimates is given in Boente, Fraiman, and Yohai, 1982). Thus, for such estimates most of the estimated AR coefficients will be small, and one expects that a good AR (p) fit can be made with p close to p_0. Correspondingly, one expects to obtain a quite reasonable identification of the order if a good robust order selection rule is used.

We propose that a robust model selection rule be constructed in the following way. Let $s_n(p, q) = s_n(p, q, \hat{\alpha})$ be a robust measure of scale of the prediction errors. Here n denotes sample size, and $\hat{\alpha}$ is a robust estimate of the parameters of an ARMA (p, q) model, with $p \leq P_n$, $q \leq Q_n$, and P_n, Q_n nondecreasing in n. Robustness of both $\hat{\alpha}$ and s_n are needed. Then choose $\hat{p}_n$, $\hat{q}_n$ to minimize

$$\mathrm{RMOD}_n(p, q) = s_n(p, q)(1 + K_n),$$

where K_n is a penalty term for overfitting. For a specific proposal using pure autoregressive fits, see Martin (1981), where a robustified AIC-type statistic was proposed and the efficacy of its use illustrated by example.

Let $\mathrm{RMOD}(\mu_y^\gamma, \bar{P}, \bar{Q})$ be the asymptotic value of $\mathrm{RMOD}_n(\hat{p}_n, \hat{q}_n)$ when the observations are $y_i^\gamma \sim \mu_y^\gamma$ and $\bar{P} = \lim P_n$, $\bar{Q} = \lim Q_n$, $\bar{P}, \bar{Q}$ finite or infinite, and $p_0 \leq \bar{P}$, $q_0 \leq \bar{Q}$ where $\mu_{x,0}$ is for an ARMA (p_0, q_0) model. The basic requirement is that $\mathrm{RMOD}(\cdot, \bar{P}, \bar{Q})$ be continuous at $\mu_{x,0}$ and preferably also continuous at all μ_x in a neighborhood of $\mu_{x,0}$ (where it is possible that $p_0 > \bar{P}$, $q_0 > \bar{Q}$). In fact it is desirable to be somewhat more nonparametric. Simply assume that we agree to fit with ARMA (p_n, q_n) models, but that $\mu_{x,0}$ is an arbitrary stationary Gaussian measure. The basic robustness property of $\mathrm{RMOD}(\cdot, \bar{P}, \bar{Q})$ should still be the same.

It is also natural to require "Fisher" consistency in the sense that $\mathrm{RMOD}(\mu_{x,0}, \bar{P}, \bar{Q}) = \sigma_\varepsilon^2$, where σ_ε^2 is the innovations variance of the ARMA (p_0, q_0) model. One might then try to establish consistency of $\hat{p}_n, \hat{q}_n$ at $\mu_{x,0}$. (Because small biases are inescapable—unless one wants to be super-adaptive—one cannot expect consistency of any model selection rule except at the nominal model $\mu_{x,0}$.) It would also be desirable to establish optimality properties at $\mu_{x,0}$ (see, for example, Shibata, 1980; Härdle, 1985).

Innovations outliers, intervention analysis, adaptivity, etc. *Künsch* points out that (2.2) is not sufficiently general to include innovations outlier models, which is certainly true. However, we regard this as relatively unimportant for the following reasons. First of all, though heavy-tailed symmetric innovations distributions will produce outliers (of highly structured form), such distributions will not result in asymptotic bias. Even asymmetric innovations distributions will not result in asymptotic bias for GM and RA estimates of AR models, provided an intercept term is included in the model (this will even be true for ARMA models, but the RA and GM estimates will no longer be robust without "truncation"—see Bustos and Yohai, 1986). Secondly, innovations outliers are often good in the sense that they are "good" leverage points which result in increased precision for estimates of the parameters $\phi_1, \ldots, \phi_p$, as has been pointed out in earlier literature. *Poor*'s system identification problem provides an interesting contrary case since the innovations are replaced by measured system inputs u_i which may be observed with contamination errors.

Of course, innovations outliers represent just one of several kinds of deviations from a nominal Gaussian model which are often substantially different in character from the kinds of contamination-type deviations we have focussed on. Level shifts, changes in trend, a variety of other "shaped" changes, and time-varying parameters are among the problems *West* and *Miller* and *Lee* are concerned with. These kinds of behavior certainly occur with some frequency in economic time series, and in other subject-matter areas as well. It is clear that the use of intervention analysis/structured dummy variables often gives good results in those situations where shaped changes in a time series are attributable to known causes.

Wegman expresses his concern with the possibility that a robust method might mask an effect which would be well accommodated by intervention analysis. We consider, on the contrary, that robust estimates have two distinct and useful roles in conjunction with intervention analysis. The first role occurs in those cases where there is enough knowledge to specify an intervention form. In such cases one in general still has no assurance that outliers will not cause problems. This can be dealt with by adapting AM, GM, or RA estimates to intervention models. The second role occurs in those situations where one overlooks the possibility of intervention modeling, or where one is rather uncertain about what intervention shape to use. A robust estimate will produce large residuals in locations where an intervention should be applied (the AM estimate/robust filter- or smoother-cleaner approach may be preferred in this case). These residuals will help suggest the form of the intervention and therefore enable its incorporation into the model. If, instead, a nonrobust procedure is used, the parameters of the core model can be severely biased in an effort to explain the overlooked intervention effects. As a result, examination of the prediction residuals may not reveal the need for an intervention.

Of course the adaptive and Bayesian techniques proposed by *West* have substantial appeal. We would emphasize that for forecasting purposes, one of the most crucial needs is for a methodology which can assess whether or not unusual behavior near the end of a series is passing in nature, or represents real changes in the structure of the process (e.g., are the last few points additive outliers or innovations outliers). A Bayesian approach is quite natural and appealing. However, one difficulty is clearly paramount even when a user is able to specify good priors: There will be relatively few data points with which to estimate the unusual new structure, and hence even short term forecasts based on such changes may not be very good. One must give an honest assessment of this to the user. In general we would both push the Bayesian approach hard, and also force the user to carefully evaluate multiple forecast options (including the associated models and uncertainty). Perhaps this is the kind of thing *West* has in mind—however, his 1986 references were unavailable to us.

We do question *West*'s almost total rejection of stationarity for economic time series—this runs against the grain of a considerable amount of experience according to which specialized adjustment for nonstationarity and structured effects results in a stationary core process—and it is the parameters of this core process which determine the confidence intervals for short-term forecasts. Also, one must be careful that adaptivity does not become superadaptivity with little precision or confidence in the model—the data can certainly be fitted too well (see Los, 1985).

West is correct in saying that omnibus robust methods will in *some* circumstances tend to oversmooth the data, and this is an issue which one certainly must pay attention to. It should be noted, however, that a good *smoother*-cleaner (Martin, 1979) handles isolated outliers or short patches nicely (just as does a good *filter*-cleaner, e.g., as in Martin, Samarov, and Vandaele, 1983) while at the same time making rapid transitions (not oversmoothing.) at level shifts (where a *filter*-cleaner may result in smearing of the shift). Furthermore, we would never recommend blind use of an "omnibus" method to the exclusion of other reasonable procedures.

In fact, one needs a variety of methodologies, robust and otherwise, at one's disposal, and the good data analyst follows Tukey's dictum of multiple analysis. Among the methodologies we should have in hand are those which combine robustness with other features. For example, in answer to one of *West*'s questions: The extension of robust methods, particularly AM-type estimates (see Martin and Yohai, 1985), to cover estimation of the fixed parameters in the dynamic/time-varying parameter problem seems quite feasible. Also, we do not see any problem in applying the current **IF** concept to estimation of the fixed parameters in dynamic models with time varying parameters.

West raises some very interesting questions about outliers and binary time-series models, about which we have not thought very much (but are stimulated to do so very soon).

Some of the questions raised by *Miller* and *Lee* have been covered by our preceding discussion, and we shall respond to a few others. It is true that we should assume joint stationarity of $(x_\iota, w_i, z_i^\gamma)$. Note, however, our comments above concerning **IF's** for nonstationary processes. With regard to assumptions on the estimator sequence $\{\mathbf{T}_n\}$, consistency is the main requirement. The important point with regard to the domain of **T** is that $\mathbf{T}(\mu_y^\gamma)$ be well defined by (3.3). The fact that **ICH** is defined for measures that are not stationary and ergodic is quite consistent with the fact that, in general, the directional derivative of **T** in the direction determined by δ_y is not the same as the derivative **IF** along a stationary arc.

While it is true that outliers with no assignable causes may deserve to have full weight, it is equally true that they may deserve to be downweighted. One must distinguish between downweighting in estimating structural parameters and downweighting for estimating error variances and for forecasting. There is usually little harm in downweighting for estimating structural parameters—at most some efficiency is lost. Forecasting is quite another matter, which we have commented upon above. With regard to error-variance estimates: it is true that a robust residuals scale estimate can result in a considerably smaller estimate of variance of future observations than the usual sum-of-squares estimate. However, how much reliability will one put on the latter type of estimate when it is influenced quite heavily by a very small number of observations?

With regard to the above issues, the recognition of PR- and AO-type behavior versus innovations outliers- (IO) like behavior is relevant. *Miller and Lee* correctly note the difficulty of assessing AO structure using conventional methods, and the poor quality of ARMA(1, 1) least-squares fits in such situations. The latter point is hardly surprising since the ARMA(1, 1) structure may often be determined by just a small fraction of observations, and only a very large sample size would then result in good estimates via a least-squares/second-order fit. Robust estimates on the other hand have at least some useful role in both supplying good parameter estimates and checking for IO versus AO structure. Some evidence concerning the latter point is provided by Martin and Zeh (1977).

We would also address an attitude which permeates the *Miller and Lee* viewpoint—and which is held by others, particularly those who concentrate on analysis of economic time series. Namely, "the analyst knows enough about his data to provide sufficient model structure to accommodate all conceivable problems, including contamination by outliers" (take away a competent statistical modeling capability, and this becomes a rather antistatistical attitude). Thus robust techniques are not needed, they probably are not to be trusted anyway, and they certainly are not completely developed. True, some analyst will do quite well without robust techniques most of the time. However, our experience is that many will do not so well much of the time, and this group will often benefit from the availability of good robust techniques to aid their analysis. Even the first group will sometimes benefit from the use of robust procedures by virtue of discovering the difficulties in the data more quickly.

Although the last sentence of *Miller and Lee* makes a valid point, it also reveals a certain myopia concerning the nature of the universe of time series. This universe is incredibly large and diverse, and the "other events" which impact economic time series would be regarded as highly specialized by a radar engineer (who might be concerned with real-time problems) or an oceanographer, for example. We can think of many users who would be quite delighted to have available robust estimates which promise *only* good estimates of parameters of core processes.

Finally, we certainly do not believe that robust procedures are a be-all and end-all. They are simply an often useful statistical tool which should be on the shelf with other standard statistical methods for the user to choose. Intervention analysis and other modeling techniques, such as those of *West*, and robust estimation for time series are methodologies which can and should live happily next to one another, and as such they will be mutually complimentary.

Donoho's comments. Although not a formal discussant, *Dave Donoho* has raised a number of very interesting questions concerning our paper. We respond to some of them here. The first has to do with the relationship between **IF** and a "Hampel" influence curve $\mathbf{IC} = \mathbf{IC}_{\mu_x}(\nu)$ defined as the directional derivative of $\mathbf{T}(\mu)$ at μ_x in the direction of stationary ergodic measures ν (rather than in the direction of nonstationary point masses δ_y as in (4.1). With $\mu_y = (1-\gamma)\mu_x + \gamma\nu$, we have

$$\mathbf{IC}_{\mu_x}(\nu) = \lim_{\gamma \to 0} \frac{\mathbf{T}(\mu_\gamma) - \mathbf{T}(\mu_x)}{\gamma}.$$

Let Λ_{μ_x} be the linear functional (defined on the set of signed measures for which **ICH** is integrable) given by

$$\Lambda_{\mu_x}(\nu) = \int \mathbf{ICH}(\mathbf{y}_1, \mathbf{t}_0)\, d\nu.$$

Then $\Lambda_{\mu_x}(\nu) = \mathbf{IC}_{\mu_x}(\nu)$, and Theorem 4.1 states that

$$\mathbf{IF}\big(\mu_w, \mathbf{T}, \{\mu_y^\gamma\}\big) = \Lambda_{\mu_x}(\nu_y^* - \mu_x) = \Lambda_{\mu_x}(\nu_y^*),$$

where

$$\nu_y^* = \left[\frac{\partial \mu_y^\gamma}{\partial \gamma}\right]_{\gamma=0} + \mu_x,$$

and $\Lambda_{\mu_x}(\mu_x) = 0$. The tangent line to the arc μ_y^γ at $\gamma = 0$ is given by

$$\nu_y^\gamma = (1 - \gamma)\mu_x + \gamma\nu_y^*.$$

Although ν_y^* is a signed measure, it is not necessarily a probability measure and it need not be bounded. Thus **IF** does not in general have any heuristic interpretation as $\mathbf{IC}_{\mu_x}(\nu) = \Lambda_{\mu_x}(\nu)$ with ν such that $\mu_\gamma = (1 - \gamma)\nu_x + \gamma\nu$ corresponds to a mixture of processes. Furthermore, we do not see any simple way to determine $\Lambda_{\mu_x}(\nu_y^*)$ using the values of $\Lambda_{\mu_x}(\nu)$ with ν ranging over the class of stationary ergodic measures.

Nonetheless a "long-patch" interpretation suggested by *Donoho* is correct: If ν is an ergodic stationary probability measure, and $\mu_y^{\gamma, k}$ is the probability measure corresponding to a process $y_i^{\gamma, k}$ defined by observing patches of length k of a contaminating process with measure ν a fraction of time γ, and observing the nominal process x_i the rest of the time, then under regularity conditions

$$\Lambda_{\mu_x}(\nu) = \lim_{k \to \infty} \mathbf{IF}\left(\nu, \mathbf{T}, \left\{\mu_y^{\gamma, k}\right\}\right).$$

Donoho has also suggested that it would be interesting to determine **GES**'s using the largest natural class **P** of measures in (6.1), which would be the class of all possible arcs $\{\mu_y^\gamma\}$ corresponding to processes defined by (2.1) and (2.2). The corresponding "least-favorable" measure would appear to be of considerable interest. It follows from our comments above that the **GES** in question is given by

$$\overline{\mathbf{GES}} = \sup_{\nu_y^* \in \mathbf{P}^*} |\Lambda_{\mu_x}(\nu_y^*)|,$$

where $\mathbf{P}^*$ is the family of all stationary ergodic signed measure ν_y^* as specified above. Unfortunately, it appears at the moment to be difficult to compute the least favorable measure $\bar{\nu}_y^*$. It seems likely that the least favorable arc $\{\bar{\mu}_y^\gamma\}$ will correspond to a process y_i^γ with z_i^γ depending on the nominal process x_i, the reason being that $\overline{\mathbf{GES}}$ should be attained by placing outliers where they will be most harmful and this would depend on x_i. Thus the least favorable arc may correspond to a rather complicated process. Perhaps some real effort here will nonetheless pay off.

One other question raised by *Donoho* was "What does a fixed finite patch length, say $k = 20$, mean, when the sample size goes to infinity?" Let's focus briefly on GM estimates of order p autoregressions for simplicity, where the estimate has a "span" of $p + 1$. Then the length k of the patch relative to the span $p + 1$ determines the proportion of end effects of the patch relative to the estimate (an end effect occurs when the patch does not cover the entire span), and thus k should clearly affect **IF**.

In the general case one also expects **IF** to depend upon k, and furthermore patch length effects have their own asymptotics (see Theorems 5.2(ii) and Corollary 5.3 of Martin and Yohai, 1984b): patch length asymptotics have set in when k is such that

$$\frac{1}{k}\sum_{i=1}^{k} \tilde{\psi}(\mathbf{w}_i^1, \mathbf{x}_0, \mathbf{t}_0) \approx E\tilde{\psi}(\mathbf{w}_1, \mathbf{t}_0).$$

This depends on: (i) How fast $\tilde{\psi}(\mathbf{w}_i^1, \mathbf{x}_0, \mathbf{t}_0)$ is approximated by $\tilde{\psi}(\mathbf{w}_i, \mathbf{t}_0)$, and

(ii) how fast the ergodic theorem holds for $k^{-1}\Sigma_{i=1}^{k}\tilde{\psi}(\mathbf{w}_\iota, \mathbf{t}_0)$. Factor (i) depends upon the patch length and effective span of $\tilde{\psi}$.

Vote of thanks. Borrowing on a nice tradition of the Royal Statistical Society, we offer our vote of thanks to the discussants.

Acknowledgment. The authors wish to thank Adrian Raftery for some useful comments made during the preparation of this rejoinder.

REFERENCES

Basawa, I. V., Huggins, R. M. and Staudte, R. C. (1985). Robust tests for time series with an application to first-order autoregressions. *Biometrika* **72** 559–572.

Boente, G., Fraiman, R. and Yohai, V. J. (1982). Qualitative robustness for general stochastic processes. Technical Report No. 26, Dept. Statist., Univ. of Washington, Seattle.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environment problems. *J. Amer. Statist. Assoc.* **70** 70–79.

Bruce, A. and Martin, R. D. (1986). Leave-k-out diagnostics for time series. Technical Report, Dept. Statist., Univ. of Washington, Seattle, in preparation.

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

Hampel, F. R., Rousseeuw, P. J. and Ronchetti, E. (1981). The change-of-variance curve and optimal redescending M-estimators. *J. Amer. Statist. Assoc.* **76** 643–648.

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.

Härdle, W. (1985). An efficient selection of regression variables when error distribution is incorrectly specified. Mimeo Series No. 1582, Dept. Statist., Univ. of North Carolina.

Los, C. A. (1985). Discussion contribution to paper by West, Harrison and Migon. *J. Amer. Statist. Assoc.* **80** 92–93.

Martin, R. D. (1979). Approximate conditional-mean type smoothers and interpolators. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.) 117–143. Springer, Berlin.

Martin, R. D., Samarov, A. and Vandaele, W. (1983). Robust methods for ARIMA models. In *Applied Time Series Analysis of Economic Data.* (A. Zellner, ed.) 153–169. Economic Research Report ER-5, Bureau of the Census, Washington.

Martin, R. D. and Zamar, R. (1985). Efficient min-max bias M-estimates of location and scale. Technical Report No. 72, Dept. Statist., Univ. of Washington, Seattle.

Martin, R. D. and Zeh, J. E. (1977). Determining the character of time series outliers. *Proc. ASA Bus. Econ. Statist. Sec.* 818–823. Amer. Statist. Assoc., Washington.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle and D. Martin, eds.) 256–272. Springer, New York.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

Yohai, V. J. (1985). High breakdown point and high efficiency robust estimates for regression. Technical Report No. 66, Dept. Statist., Univ. of Washington, Seattle.

Yohai, V. J. and Zamar, R. H. (1985). High breakdown point estimates of regression by means of minimization of an efficient scale. Technical Report No. 81, Dept. Statist., Univ. of Washington, Seattle.

Zamar, R. H. (1985). Robust estimation for the errors in variables model. Ph.D. dissertation, Dept. Statist., Univ. of Washington, Seattle.

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195

Department of Mathematics
University of Buenos Aires
Cuidad Universitaria, Pabellon 1
1428 Buenos Aires
Argentina