The Annals of Statistics
1984, Vol. 12, No. 3, 843–863

INFINITESIMAL ROBUSTNESS FOR AUTOREGRESSIVE PROCESSES[1]

BY H. KÜNSCH[1]

University of Tokyo and ETH Zürich

We define the influence function and construct optimal robust estimators for autoregressive processes. We show that the asymptotic bias caused by small contaminations of the marginals can be written as the integral of a certain function with respect to the contamination. This function is called the influence function. It is unique only up to an equivalence relation, but there is a natural unique version which describes the limiting influence of an additional observation given the previous observations. Moreover, with this version the asymptotic variance at the true model can be expressed in a simple form. Optimal robust estimators minimize this asymptotic variance under a constraint on the influence function. As in the i.i.d. case, they are found by truncating a multiple of the influence function of the maximum likelihood estimator.

0. Introduction. Up to now, two main approaches in robust statistics have emerged, the minimax approach of Huber presented in his book (1981) and the infinitesimal approach of Hampel, see e.g. the handouts of Hampel et al. (1982). However, these theories deal almost exclusively with independent observations. The robust estimators for time series proposed so far in the literature (see e.g. Kleiner et al., 1979) are based mainly on heuristic ideas. Papantoni-Kazakos and Gray (1979) and Boente et al. (1982) investigated qualitative robustness for time series, but this is not sufficient for constructing good robust procedures. Here we define and obtain optimal robust estimators for autoregressive processes by generalizing the infinitesimal approach.

The basic tool in this approach is the influence function of Hampel (1968, 1974). It contains the information on both the asymptotic bias and the asymptotic variance. In order to obtain such an influence function in our situation, we study the infinitesimal asymptotic bias of a contamination for estimators which depend only on some finite dimensional empirical marginal. We show that this bias can be written as the integral of a function with respect to the contamination (Theorem 1.1). This function is unique only up to an equivalence relation (Theorem 1.2), but there is a unique version which has the interpretation as the influence of an additional observation given the previous values. Namely, such an influence should have—for the clean process—conditional expectation zero given the past. There is exactly one version which satisfies this condition (Theorem 1.3). Moreover we can express the asymptotic variance at the model with this version in a simple form (Formula (1.23)).

Received February 1983; revised December 1983.

[1] Research supported in part by a scholarship from the Japan Society for the Promotion of Science.

AMS 1980 *subject classification.* Primary 62F35; secondary 62M10.

Key words and phrases. Robustness for time series, influence function, optimal robust estimators, contaminations of time series, M-estimators, autoregressive processes, asymptotic properties.

In Chapter 2 we show that, subject to a bound on this version of the influence function, the trace of the asymptotic variance is minimized by a Huberized maximum likelihood estimator. The bound used here is in general stronger than a bound on the infinitesimal asymptotic bias (see Example 2.7), but the use of this stronger bound is justified by a resistance point of view, cf. Section 1.5. The location of the observations and the scale of the innovations are estimated without affecting the precision and robustness of the main parameter. In the Gaussian case, the optimal robust estimator is explicitly given in Formulas (2.12)–(2.14). It is similar to the Hampel–Krasker–Welsch form of bounded influence regression. An example of observed data with outliers is included, and there the robust estimators perform much better than maximum likelihood.

1. The influence function for autoregressive processes. We consider in this paper the usual autoregressive model of order p(AR(p) for short):

$$(X_i - \eta) = \sum_{k=1}^{p} \beta_k (X_{i-k} - \eta) + U_i \quad \text{with } U_i \text{ i.i.d.} \tag{1.1}$$

We assume that the innovations U_i are distributed according to $\sigma^{-1}h(x/\sigma)\,dx$ where h is a known probability density on $\mathbb{R}$ satisfying $\int xh(x)\,dx = 0$, $\int x^2h(x)\,dx = 1$. $\beta_1, \cdots, \beta_p$, η and σ are functions of an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^q$ ($q \leq p + 2$). Furthermore we assume that for all θ the usual stationarity condition is satisfied, i.e. all roots of $z^p - \sum_{k=1}^{p} \beta_k z^{p-k}$ have absolute value less than one.

1.1 *The maximum likelihood estimator and other estimators defined by functionals.* By $f^n(x_1, \cdots, x_n; \theta)$ we denote the joint density of $X_1, \cdots, X_n$ where X_i is as in (1.1). By the Markov property of an AR-process we obtain for all $n > p$

$$f^n(x_1, \cdots, x_n; \theta) = f^p(x_1, \cdots, x_p; \theta) \prod_{j=p+1}^{n} h((x_j^* - \Sigma\beta_k x_{j-k}^*)/\sigma)\sigma^{p-n} \quad \text{with } x_i^* = x_i - \eta. \tag{1.2}$$

Using the notations

$$\lambda^n(x_1, \cdots, x_n; \theta) = \frac{\partial}{\partial\theta} \log f^n(x_1, \cdots, x_n; \theta) \tag{1.3}$$

and

$$\kappa(x_1, \cdots, x_{p+1}; \theta) = \frac{\partial}{\partial\theta} \log(h((x_{p+1}^* - \Sigma\beta_k x_{p+1-k}^*)/\sigma)/\sigma) \tag{1.4}$$

we have from (1.2)

$$\lambda^n(x_1, \cdots, x_n; \theta) = \lambda^p(x_1, \cdots, x_p; \theta) + \sum_{j=1}^{n-p} \kappa(x_j, \cdots, x_{j+p}; \theta). \tag{1.5}$$

In the sequel we will consider λ^n and κ as column vectors. From (1.5) we see that the maximum likelihood estimator for θ based on $x_1, \cdots, x_n$ is up to a boundary term the solution of the following set of equations

$$\sum_{j=1}^{n-p} \kappa(x_j, \cdots, x_{j+p}; \hat{\theta}_n) = 0. \tag{1.6}$$

But it is well known that the maximum likelihood estimator is in general very sensitive to outliers and other irregularities in the data. In this paper we look for a more robust estimator in a certain class. One possible class consists of estimators which can be written as a solution of the following equations

$$\sum_{j=1}^{n-m+1} \psi(x_j, \cdots, x_{j+m-1}; \hat{\theta}_n) = 0 \tag{1.7}$$

for some fixed but arbitrary $m \in \mathbb{N}$ and ψ: $\mathbb{R}^m \times \Theta \to \mathbb{R}^q$. We call such an estimator an *M-estimator*. The GM-estimators of Kleiner et al. (1979) and the Φ-estimators of Bustos (1982) are special cases of M-estimators in our terminology.

A more general class of estimators is obtained in the following way. For observations $x_1, \cdots, x_n$ we define the empirical m-dimensional marginal distribution $\rho(x, n)^m$ $(m < n)$ by

$$\rho(x, n)^m = n^{-1} \sum_{i=1}^{n} \delta_{(x_i, \cdots, x_{i+m-1})} \quad \text{where} \quad x_i = x_{i-n} \quad \text{for} \quad i > n. \tag{1.8}$$

Here δ_x is the point mass at $x \in \mathbb{R}^m$. The periodic continuation of the observations is just a technical device in order that $\rho(x, n)^m \in \mathcal{M}^m_{\text{stat}} = \{m\text{-dimensional marginals of stationary processes}\}$. We consider then estimators $\hat{\theta}_n$ which can be written as a functional evaluated at $\rho(x, n)^m$ for some m independent of n:

$$\hat{\theta}_n(x_1, \cdots, x_n) = T(\rho(x, n)^m) \quad \text{where} \quad T: \mathcal{M}^m_{\text{stat}} \to \Theta. \tag{1.9}$$

Sometimes it is necessary to restrict T to a certain subset of $\mathcal{M}^m_{\text{stat}}$, e.g. if second moments are needed. The modifications in such a case are straightforward. For $m = 1$ this setup is the usual one in the i.i.d. case. Choosing as T the functional defined by $T(\mu^m) = \theta$ iff $\int \psi(x_1, \cdots, x_m; \theta)\mu^m(dx) = 0$, we see that our M-estimators belong to this class after an asymptotically negligible modification at the boundary.

1.2 *The bias caused by contaminations.* In time series there are many different types of contamination which makes robustness more difficult than in the i.i.d. case. At least one should consider innovative and additive outliers (see e.g. Kleiner et al., 1979), and we should distinguish between single outliers separated by good data and outliers occurring in patches. Fixed or random size of the patches gives rise to many variations; and instead of adding outliers to the good data we could multiply the good data by outliers or replace them completely. Another type of contamination which is quite different affects only the dependence structure, but has no atypically large observations, e.g. a Gaussian model with a spectral density different from an AR-model. In all the types above, one can introduce a parametric family of contaminated observations $(Y_i^\varepsilon)_{\varepsilon \geq 0}$ such that for $\varepsilon = 0$ we have the model (1.1) and ε is related to the percentage and/or the size of the outliers.

If the observations (x_i) come from any stationary ergodic process, then $\rho(x, n)^m$ converges by the ergodic theorem weakly to the m-dimensional marginal μ^m almost surely as n goes to infinity. So if T is continuous in the weak topology, then $\hat{\theta}_n$ defined by (1.9) converges to $T(\mu^m)$. We therefore ask that T is *Fisher*

continuous, i.e.

$$T(\mu_\theta^m) = \theta \quad \text{for all} \quad \theta \tag{1.10}$$

where μ_θ^m is the m-dimensional marginal of (1.1). In a contamination model with marginals $\mu^{m,\varepsilon}$, the asymptotic bias is then

$$b(\varepsilon) = T(\mu^{m,\varepsilon}) - \theta = T(\mu^{m,\varepsilon}) - T(\mu^{m,0}).$$

In order to see how quickly $b(\varepsilon)$ goes to zero we take the derivative

$$b'(\varepsilon) = \lim \varepsilon^{-1}(T(\mu^{m,\varepsilon}) - T(\mu^{m,0})).$$

In order to make computations easier, we will approximate the arc $\mu^{m,\varepsilon}$ in $\mathcal{M}_{\text{stat}}^m$ by a segment $\tilde{\mu}^{m,\varepsilon} = (1 - \varepsilon)\mu_\theta^m + \varepsilon\nu^m$, $\nu^m \in \mathcal{M}_{\text{stat}}^m$. It is very plausible that this will give the same results if the arc $\mu^{m,\varepsilon}$ is smooth enough, but since $\mathcal{M}_{\text{stat}}^m$ is not a linear space, some care is needed. In all cases where we have a $(1 - \varepsilon)$-percentage of clean data, it is easily seen that for any m

$$\mu^{m,\varepsilon} = (1 - \varepsilon c)\mu_\theta^m + c\varepsilon\nu^m + o(\varepsilon) \tag{1.11}$$

where $c = \lim \varepsilon^{-1}$ Prob [at least one outlier in a block of length m] and ν^m depends on both the distribution of the outliers and μ_θ^m. In addition, it can be shown that we obtain $c = 1$ and any $\nu^m \in \mathcal{M}_{\text{stat}}^m$ if we choose outliers in long patches which replace the good data. If (1.11) holds, it is obvious that we may replace $\mu^{m,\varepsilon}$ by a segment. At the end of the next section we will show that the same is true for quite general contaminations.

1.3 *The definition of the influence function.* Based on the considerations of the previous sections, we are going to study now

$$T'(\theta, \nu^m) = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1}(T((1 - \varepsilon)\mu_\theta^m + \varepsilon\nu^m) - T(\mu_\theta^m)) \quad (\nu^m \in \mathcal{M}_{\text{stat}}^m). \tag{1.12}$$

In the case where T belongs to an M-estimator, we obtain by the usual Taylor series approximation that under some regularity condition

$$T'(\theta, \nu^m) = M^{-1} \int \psi(x_1, \cdots, x_m; \theta)\nu^m(dx_1, \cdots, dx_m) \quad \text{where}$$
$$M = \int \psi(x, \theta)\lambda^m(x, \theta)^t\mu_\theta^m(dx). \tag{1.13}$$

The next theorem shows that in general $T'(\theta, \nu^m)$ is given by a kernel $t(x_1, \cdots, x_m; \theta)$

$$T'(\theta, \nu^m) = \int t(x_1, \cdots, x_m; \theta)\nu^m(dx_1, \cdots, dx_m).$$

Note that $T'(\theta, \nu^m)$ will be *affine*, i.e. $T'(\theta, \alpha\nu_1^m + (1 - \alpha)\nu_2^m) = \alpha T'(\theta, \nu_1^m) + (1 - \alpha)T'(\theta, \nu_2^m)$. Therefore we can apply the following result to each component of $T'(\theta, \nu^m)$.

THEOREM 1.1. *A functional $L: \mathcal{M}_{\text{stat}}^m \to \mathbb{R}$ is of the form $L(\nu^m) = \int t(x)\nu^m(dx)$ with t bounded and continuous iff L is affine and weakly continuous.*

The proof is given in Chapter 3. For $m = 1$ it is trivial, just put $t(x) = L(\delta_x)$. But for $m \geq 2$, $\delta_x \notin \mathcal{M}^m_{\text{stat}}$, and the convex set $\mathcal{M}^m_{\text{stat}}$ has a more complicated structure than for $m = 1$. Namely, two different extremal points are no longer mutually singular as the example $m = 2$, $\nu_1 = (\delta_{(x,y)} + \delta_{(y,x)})/2$ and $\nu_2 = (\delta_{(x,y)} + \delta_{(y,z)} + \delta_{(z,x)})/3$ shows, and it seems difficult to determine all extremal points. For these reasons we cannot define the kernel by taking derivatives of T in the direction of suitable ν's.

Another difficulty with $m \geq 2$ comes from the fact that the kernel is not unique:

THEOREM 1.2. $\int t(x)\nu^m(dx) = 0$ $\forall \nu^m \in \mathcal{M}^m_{\text{stat}}$ *iff* $t(x_1, \cdots, x_m) = g(x_1, \cdots, x_{m-1}) - g(x_2, \cdots, x_m)$ *with an arbitrary* g.

For the proof see Chapter 3.

We therefore call any function $IC_T(x, \theta)$: $\mathbb{R}^m \times \Theta \to \mathbb{R}^q$ such that

$$T'(\theta, \nu^m) = \int IC_T(x, \theta)\nu^m(dx) \quad \forall \nu^m \in \mathcal{M}^m_{\text{stat}} \tag{1.14}$$

an *influence function of* T *at* θ. So the influence function is in fact a whole equivalence class of all functions

$$IC_T(x_1, \cdots, x_m; \theta) + g(x_1, \cdots, x_{m-1}; \theta) - g(x_2, \cdots, x_m; \theta),$$

g: $\mathbb{R}^{m-1} \times \Theta \to \mathbb{R}^q$ arbitrary.

By definition any influence function satisfies

$$\int IC_T(x, \theta)\mu^m_\theta(dx) = 0. \tag{1.15}$$

Moreover, recalling that the segment $(1 - \varepsilon)\mu^m_\theta + \varepsilon\nu^m$ is just an approximation for any $\tilde{\nu}$ near μ^m_θ, we conclude

$$T(\nu^m) \approx \theta + \int IC_T(x, \theta)\nu^m(dx) \quad \text{if} \quad \nu^m \text{ is near } \mu^m_\theta. \tag{1.16}$$

From the Fisher-consistency and (1.16) we obtain

$$\begin{aligned} \tilde{\theta} - \theta = T(\mu^m_{\tilde{\theta}}) - T(\mu^m_\theta) &\approx \int IC_T(x, \theta)f^m(x, \tilde{\theta})\, dx_1 \cdots dx_m \\ &= \int IC_T(x, \theta)(f^m(x, \tilde{\theta}) - f^m(x, \theta))\, dx_1 \cdots dx_m. \end{aligned}$$

Hence by differentiating with respect to θ:

$$\begin{aligned} Id &= \int IC_T(x, \theta)\left(\frac{\partial}{\partial\theta} f^m(x, \theta)\right)^t dx_1 \cdots dx_m \\ &= \int IC_T(x, \theta)\lambda^m(x, \theta)^t\mu^m_\theta(dx). \end{aligned} \tag{1.17}$$

For M-estimators (1.17) is an obvious consequence from (1.13). On the other

hand it follows from (1.17) that any estimator of the type (1.9) which has an influence function can be replaced by an M-estimator with the same IC: we just take $IC_T(x, \theta)$ as our ψ.

As a second application of (1.16) we consider the asymptotic bias in a general contamination model $\mu^{m,\varepsilon}$.

$$b(\varepsilon) = T(\mu^{m,\varepsilon}) - T(\mu^{m,0}) \approx \int IC_T(x, \theta)\mu^{m,\varepsilon}(dx)$$

$$= \int IC_T(x, \theta)[\mu^{m,\varepsilon}(dx) - \mu^{m,0}(dx)].$$

So if a finite signed measure $\dot{\mu}^m$ exists such that $\varepsilon^{-1}[\mu^{m,\varepsilon} - \mu^{m,0}]$ converges weakly to $\dot{\mu}^m$, then

$$b' = \int IC_T(x, \theta)\dot{\mu}^m(dx). \tag{1.18}$$

1.4 *Asymptotic variance and a particular version of the influence function.* In order to simplify notations in what follows, we introduce the shift-operator S: $\mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ and its iterations. They are defined by

$$(S^n x)_i = x_{i+n} (n = 0, 1, \cdots). \tag{1.19}$$

If our observations follow the model (1.1), then $\rho(x, n)^m$ is near μ_θ^m for n large, so by (1.16)

$$\hat{\theta}_n - \theta = T(\rho(x, n)^m) - \theta \approx n^{-1} \sum_{i=0}^{n-m-1} IC_T(S^i x, \theta). \tag{1.20}$$

This suggests that for clean observations $n^{1/2}(\hat{\theta}_n - \theta)$ will be asymptotically normal with mean zero and variance-covariance matrix

$$\begin{aligned} C(T, \theta) &= E_\theta[IC_T(x, \theta)IC_T(x, \theta)^t] \\ &\quad + \sum_{i=1}^{\infty} E_\theta[IC_T(x, \theta)IC_T(S^i x, \theta)^t + IC_T(S^i x, \theta)IC_T(x, \theta)^t]. \end{aligned} \tag{1.21}$$

Rigorous conditions and proofs for this to be true are given in Bustos (1982). For the definition and construction of optimal robust estimators, we will avoid these problems by minimizing $C(T, \theta)$ regardless if asymptotic normality holds or not.

Formula (1.21) is independent of the particular choice of the influence function, but one may ask if there is a version for which it becomes simpler. We will denote by $IC_T^{\text{cond}}(x, \theta)$ (cond for conditional mean zero) any version of the influence function which satisfies

$$\int IC_T^{\text{cond}}(x_1, \cdots, x_m; \theta)\mu_\theta(dx_m \mid x_{m-1}, \cdots, x_{m-p}) = 0 \tag{1.22}$$

for all $x_1, \cdots, x_{m-1}$. By (1.22) $IC_T^{\text{cond}}(S^i x, \theta)$ and $IC_T^{\text{cond}}(S^j x, \theta)$ are uncorrelated for $i \neq j$, so we obtain from (1.21)

$$C(T, \theta) = \int IC_T^{\text{cond}}(x, \theta)IC_T^{\text{cond}}(x, \theta)^t \mu_\theta^m(dx), \tag{1.23}$$

i.e. the same formula as in the i.i.d. case. The following result shows that IC_T^{cond} always exists if $m > p$ (which is no restriction) and that it is unique.

THEOREM 1.3. *Let μ denote the distribution of an* AR(p)*-process* (1.1). *If* f: $\mathbb{R}^m \to \mathbb{R}$, $m > p$, *is continuous,* $\sup |f(x)|/(1 + |x|) < \infty$ *and* $\int f(x)\mu^m(dx) = 0$, *then there exists a continuous function* g: $\mathbb{R}^{m-1} \to \mathbb{R}$ *with* $\sup |g(x)|/(1 + |x|) < \infty$ *and* $\int (f(x_1, \cdots, x_m) + g(x_1, \cdots, x_{m-1}) - g(x_2, \cdots, x_m))\mu(dx_m | x_{m-1}, \cdots, x_{m-p}) = 0$ *for all* $x_1, \cdots, x_{m-1}$. g *is unique up to an additive constant.*

The proof is given in Chapter 3.

IC_T^{cond} is uncorrelated with any function depending only on $x_1, \cdots, x_{m-1}$. In particular, from (1.5) and (1.17)

$$\begin{aligned} &\int IC_T^{\text{cond}}(x, \theta)\kappa(S^{m-p-1}x, \theta)^t\mu_\theta^m(dx) \\ &= \int IC_T^{\text{cond}}(x, \theta)\lambda^m(x, \theta)^t\mu_\theta^m(dx) = Id. \end{aligned} \tag{1.24}$$

In the case of the approximate maximum likelihood estimator (1.6) it is easily checked that

$$IC_T^{\text{cond}}(x, \theta) = I(\theta)^{-1}\kappa(x, \theta) \tag{1.25}$$

where $I(\theta)$ is the *Fisher information*

$$I(\theta) = \int \kappa(x, \theta)\kappa(x, \theta)^t\mu_\theta^{p+1}(dx). \tag{1.26}$$

So for the maximum likelihood estimator the obvious choice for the influence function leads to the version IC_T^{cond}, and we get the well known result $C(\text{MLE}, \theta) = I(\theta)^{-1}$.

IC_T^{cond} has the intuitive interpretation of the influence of an observation at x_m given the $m - 1$ previous values. Namely, by (1.22) its conditional expectation is zero for the clean process and by (1.20) $\hat{\theta}_n - \theta$ is approximately the average of all influences. This interpretation together with the simplicity of formula (1.23) and the result in the case of maximum likelihood suggests that IC_T^{cond} is the natural choice for the influence function.

1.5 *The gross error sensitivity.* We have seen in Sections 1.2 and 1.3 that our influence function contains all information on the infinitesimal asymptotic bias b'. If one is interested in the value of b' for a particular contamination, one has of course to integrate over the relevant $\nu^m \in \mathcal{M}_{\text{stat}}^m$, but in most cases there are no problems in finding ν^m and doing the calculations. Another question is how big b' can be at most. For any contamination with a $(1 - \varepsilon)$-percentage of clean data $|b'|$ is by (1.11) less than or equal to $c\gamma$ where $\gamma = \sup\{|\int IC_T(x, \theta)\nu^m(dx)|, \nu^m \in \mathcal{M}_{\text{stat}}^m\}$ and c is the same as in (1.11). Moreover by the remark following (1.11) $|b'|$ is arbitrarily close to γ if the outliers occur in long patches. This suggests calling $\gamma = \gamma(T, \theta)$ the gross-error-sensitivity of the estimator T at the parameter θ. One might object that outliers in long patches are not so frequent, but I think that at least for a first analysis of the data, safety against all possible contaminations is desirable and in the spirit of robustness. Taking the sup of $|b'|$ only for a certain class of contaminations is dangerous

because for the calculation of b', specific properties of the outlier model are used which are hard to check (e.g. independence of the outliers and the clean data). Moreover, as we shall see in Chapter 2, bounding γ does not result in huge efficiency losses for the clean data. So from the point of view of precision, there also is no need to use methods which protect only against certain kinds of contamination.

There is also an argument in the opposite direction, namely that a bound on γ is not enough. If we want to see how sensitive an estimate is to small changes in the sample (the so-called resistance point of view, cf. Mosteller-Tukey, 1977), then γ does not help much. Consider for instance two series x and y of length n with $x_i = y_i$ for all i except i_0. If x is a typical sample of an AR-process, then it can be shown that for large n

$$\begin{aligned} &n(\hat{\theta}_n(x) - \hat{\theta}_n(y)) \\ &\approx \sum_{i=i_0-m+1}^{i_0} (IC_T(x_i, \cdots, x_{i+m-1}; \theta) - IC_T(y_i, \cdots, y_{i+m-1}; \theta)). \end{aligned} \tag{1.27}$$

Note that the right-hand side of (1.27) is independent of which version of IC we choose. Obviously $\gamma \leq \inf_g \sup_x | IC_T(x, \theta) + g(x) - g(Sx) |$ but the next example shows that there is in general strict inequality. So in particular it seems to be impossible to estimate the right-hand side of (1.27) with the help of γ.

EXAMPLE 1.4. Let $m = 2$ and choose ψ as follows: $\psi(x_1, x_2) = 0$ except in small neighborhoods of $(1, 0)$ and $(-1, 0)$, and $-2 = \psi(-1, 0) \leq \psi(x_1, x_2) \leq \psi(1, 0) = 2$. If $\nu^2 \in \mathscr{M}^2_{\text{stat}}$ has mass near $(1, 0)$, then it must have at least equal mass somewhere near the line $x_2 = 1$ where ψ is zero. Therefore

$$\sup_{\nu^2} \left| \int \psi(x_1, x_2) \nu^2(dx) \right| = 1.$$

But if $| \psi(x_1, x_2) + g(x_1) - g(x_2) | \leq 1$ for some g, then $g(1) - g(0) \leq -1$ and $g(-1) - g(0) \geq 1$ which would imply that $\psi(1, -1) + g(1) - g(-1) \leq -2$.

Because of this result we should like to bound instead of γ some quantity more closely related to the influence function itself. The right-hand side of (1.27) is not satisfactory either, since we get a different expression if we replace two or more observations. From the interpretation of IC_T^{cond} in Section 1.4 and the resistance point of view it is justified to bound IC_T^{cond}. Moreover we then can give at least upper bounds for γ as well as for the right-hand side of (1.27) and all other expressions obtained from similar considerations. So we will work in the rest of the paper with the following gross error sensitivity

$$\gamma^* = \sup_x | IC_T^{\text{cond}}(x, \theta) |.$$

Problems which arise from the fact that γ^* (and γ) depend on the chosen parametrization will be dealt with in Section 2.2.

2. Optimal robust estimators

2.1 *Unstandardized optimality with known σ.* We consider here Hampel's optimality problem: minimize the trace of the asymptotic covariance matrix $C(T, \theta)$ among all estimators of the type (1.9) which have an influence function and for which $\gamma^* = \sup_x |IC_T^{\text{cond}}(x, \theta)| \leq c(\theta)$. Using the results of the previous chapter, this is equivalent to finding an $m > p$ and a function $\psi\colon \mathbb{R}^m \times \Theta \to \mathbb{R}^q$ which minimizes

$$\text{Trace} \int \psi(x, \theta)\psi(x, \theta)^t \mu_\theta^m(dx) = \int \psi(x, \theta)^t \psi(x, \theta) \mu_\theta^m(dx) \tag{2.1}$$

under the side conditions

$$\int \psi(x_1, \cdots, x_m; \theta) \mu_\theta(dx_m \mid x_{m-1}, \cdots, x_1) = 0, \tag{2.2}$$

$$\int \psi(x_1, \cdots, x_m; \theta) \kappa(x_{m-p}, \cdots, x_m; \theta)^t \mu_\theta^m(dx) = Id, \tag{2.3}$$

$$\sup_x |\psi(x, \theta)| \leq c(\theta). \tag{2.4}$$

(compare (1.22)–(1.24)). Because the constant (2.4) and therefore also the optimal estimator depend on the chosen parametrization (see Krasker-Welsch, 1982, Stahel, 1981), we used the term "unstandardized" in the heading of this section.

Similarly to the i.i.d. case, the expression (2.1) can be transformed using (2.3). Namely, for any $q \times q$ matrix A depending only on θ we have

$$\begin{aligned}
&\text{Trace} \int \psi(x, \theta)\psi(x, \theta)^t \mu_\theta^m(dx) \\
&\quad = \text{Trace} \int (\psi(x, \theta) - A\kappa(S^{m-p-1}x, \theta)) \\
&\qquad\qquad \cdot (\psi(x, \theta) - A\kappa(S^{m-p-1}x, \theta))^t \mu_\theta^m(dx) \\
&\qquad + \text{Trace} \int (\psi(x, \theta)\kappa(S^{m-p-1}x, \theta)^t \mu_\theta^m(dx) \cdot A^t) \\
&\qquad + \text{Trace} \left(A \int \kappa(S^{m-p-1}x, \theta)\psi(x, \theta)^t \mu_\theta^m(dx)\right) \\
&\qquad - \text{Trace} \left(A \int \kappa(x, \theta)\kappa(x, \theta)^t \mu_\theta^m(dx) A^t\right) \\
&\quad = \int |\psi(x, \theta) - A\kappa(S^{m-p-1}x, \theta)|^2 \mu_\theta^m(dx) \\
&\qquad + 2 \text{ Trace } (A) - \text{Trace } (A \cdot I(\theta) \cdot A^t).
\end{aligned} \tag{2.5}$$

The last two terms are independent of ψ, so the optimal ψ has to be as close to

$A \cdot \kappa(S^{m-p-1}x, \theta)$ as the side conditions allow. Let H_c be the q-dimensional Huber-function $H_c(x) = x \min(1, c/|x|)$. Then we have:

THEOREM 2.1. *Assume the distribution of the innovations is symmetric about its mean, i.e.* $h(x) = h(-x)$, *and* σ *is known, while* $\eta, \beta_1, \cdots, \beta_p$ *depend on an unknown parameter* θ. *If the bound* $c(\theta)$ *is such that*

$$\int H_{c(\theta)}(A(\theta)\kappa(x, \theta))\kappa(x, \theta)^t \mu_\theta^{p+1}(dx) = Id$$

has a solution $A(\theta)$ *for all* θ, *then* (2.1) *is minimal under the side conditions* (2.2)–(2.4) *if* $m = p + 1$ *and* $\psi(x, \theta) = H_{c(\theta)}(A(\theta)\kappa(x, \theta))$.

PROOF. The optimal ψ under the condition (2.4) alone is by (2.5) equal to $H_{c(\theta)}(A(\theta)\kappa(S^{m-p-1}x, \theta))$. Since we have chosen $A(\theta)$ such that (2.3) is satisfied, we only have to show that the above ψ satisfies (2.2). But this follows easily by symmetry:

$$\kappa(x, \theta)_j = -\sigma^{-1} \frac{\partial}{\partial \theta_j} (\textstyle\sum_{k=1}^p \beta_k x^*_{p+1-k} + \eta) h'(u/\sigma)/h(u/\sigma)$$

where

$$x_i^* = x_i - \eta, \quad u = x^*_{p+1} - \textstyle\sum_{k=1}^p \beta_k x^*_{p+1-k}.$$

The term containing the derivative is independent of x_{p+1}, and the ratio $h'(u/\sigma)/h(u/\sigma)$ is an odd function of u; so ψ is for fixed $x_1, \cdots, x_p$ an odd function of u. □

In the following we show how to modify this optimal robust estimator if one takes constraints which are invariant to parameter transformations and if in addition σ is also unknown.

2.2 *Two possible standardizations.* The discussion here is completely analogous to the i.i.d. case (see Stahel, 1981), so we just state the results. One possibility to replace (2.4) is

$$\gamma^{**} = \sup_x (\psi(x, \theta)^t I(\theta) \psi(x, \theta))^{1/2} \leq c \tag{2.6}$$

where $I(\theta)$ is the Fisher information (1.26) and c does not depend on θ. In (2.6) we compare the bias with the scatter of the maximum likelihood estimator. For an optimal estimator we take $m = p + 1$ and

$$\psi(x, \theta) = J(\theta)^{-t} H_c(A(\theta)\kappa(x, \theta)) \tag{2.7}$$

where $J(\theta)J(\theta)^t = I(\theta)$ and $A(\theta)$ is the solution of

$$\int H_c(A(\theta)\kappa(x, \theta))\kappa(x, \theta)^t \mu_\theta^{p+1}(dx) = J(\theta)^t.$$

By construction, (2.7) satisfies (2.3) and (2.6), and under the hypotheses of Theorem 2.1 it satisfies also (2.2). By an easy modification of (2.5), it can be

shown that it minimizes the trace of $(I(\theta) \cdot \int \psi(x, \theta)\psi(x, \theta)^t \mu_\theta^m(dx))$ under the side conditions (2.2), (2.3) and (2.6), cf. Stahel (1981).

The second alternative to (2.4) is

$$(2.8) \quad \gamma^{***} = \sup_x \left[\psi(x, \theta)^t \left(\int \psi(x, \theta)\psi(x, \theta)^t \mu_\theta^m(dx) \right)^{-1} \psi(x, \theta) \right]^{1/2} \leq c.$$

Here we compare the bias with the scatter of the estimator itself. Again we take $m = p + 1$, but

$$(2.9) \quad \psi(x, \theta) = \left(\int H_c(A(\theta)\kappa(x, \theta))\kappa(x, \theta)^t \mu_\theta^{p+1}(dx) \right)^{-1} H_c(A(\theta)\kappa(x, \theta)),$$

where $A(\theta)$ is the solution of

$$\int H_c(A(\theta)\kappa(x, \theta))H_c(A(\theta)\kappa(x, \theta))^t \mu_\theta^{p+1}(dx) = Id.$$

Again, under the conditions of Theorem 2.1, this ψ satisfies (2.2), (2.3) and (2.8). It does not minimize a criterion, but by an easy modification of (2.5), no other estimator satisfying the same side conditions can have a smaller asymptotic covariance matrix, cf. Stahel (1981).

2.3 *Estimation of nuisance parameters σ and η.* In general σ is a nuisance parameter, and in most situations the same is true for η. We show here that we can estimate the nuisance parameter without affecting the estimation of the main parameter.

If only σ is the nuisance parameter, we write $\theta = (\theta_1, \theta_2)$ with $\theta_1 = \sigma$ while η and $\beta_1, \cdots, \beta_p$ depend on θ_2 alone. Similarly we partition $\kappa = (\kappa_1, \kappa_2)$, $\psi = (\psi_1, \psi_2)$.

THEOREM 2.2. *If $\psi_1(x_1, \cdots, x_{p+1}; \theta) = \chi(u/\sigma)$ where $u = x_{p+1}^* - \sum \beta_k x_{p+1-k}^*$, $x_j^* = x_j - \eta$, and χ is an even function with $\int \chi(u)h(u)\,du = 0$, and if ψ_2 is one of the optimal solutions of Sections* 2.1–2.2, *then $\hat{\sigma}$ is asymptotically independent of $\hat{\theta}_2$ and the asymptotic covariance for $\hat{\theta}_2$ is the same as for known σ.*

PROOF. $\int \chi(u)h(u)\,du = 0$ implies that (2.2) holds. Since κ_2 and ψ_2 are odd functions of u, we have by symmetry

$$\int \psi_1(x, \theta)\kappa_2(x, \theta)^t \mu_\theta^{p+1}(dx) = \int \psi_1(x, \theta)\psi_2(x, \theta)^t \mu_\theta^{p+1}(dx) = 0.$$

This means that $IC_T^{\text{cond}} = (\sigma\psi_1 \text{ const.}, \psi_2)$, so the theorem follows. □

If the nuisance parameter is (σ, η), then we write $\theta = (\theta_1, \theta_2, \theta_3)$ with $\theta_1 = \sigma$, $\theta_2 = \eta$ and $(\beta_1, \cdots, \beta_p)$ depends only on θ_3. Again we partition $\kappa = (\kappa_1, \kappa_2, \kappa_3)$ and $\psi = (\psi_1, \psi_2, \psi_3)$.

THEOREM 2.3. *If ψ_1 is as in Theorem* 2.2, *$\psi_2(x, \theta) = \xi(u/\sigma)$ with ξ odd and if ψ_3 is one of the optimal solutions of Sections* 2.1–2.2, *then $\hat{\sigma}$, $\hat{\eta}$ and $\hat{\theta}_3$ are*

asymptotically independent and the asymptotic covariance of $\hat{\theta}_3$ is the same as for known η and σ.

PROOF. U and $X_1, \cdots, X_p$ are independent, and $(X_1^*, \cdots, X_p^*)$ has the same distribution as $(-X_1^*, \cdots, -X_p^*)$. Therefore by symmetry

$$\int \psi_2(x, \theta)\kappa_3(x, \theta)^t \mu_\theta^{p+1}(dx) = \int \psi_2(x, \theta)\psi_3(x, \theta)^t \mu_\theta^{p+1}(dx) = 0.$$

The rest of the proof is the same as for Theorem 2.2. □

REMARK 2.4. Often one subtracts the median of $x_1, \cdots, x_n$ from all the x_i and then proceeds as if η is known and equal to zero. This corresponds to taking $\psi_2(x, \theta) = \operatorname{sign}(x_1 - \eta)$. But in this case the second component of IC_T^{cond} is not a multiple of ψ_2. Using the general formula (1.21) it follows that $\hat{\eta}$ and $\hat{\theta}_3$ are not asymptotically independent, although $\int \psi_2(x, \theta)\psi_3(x, \theta)^t \mu_\theta^{p+1}(dx) = 0$.

2.4 *Simplifications in the Gaussian case.* With Gaussian innovations the matrices $A(\theta)$ in (2.7) and (2.9) can be calculated up to a factor. Moreover the two standardizations give the same estimators. These results are based on the following

LEMMA 2.5. *Let the innovations U_i be Gaussian and assume that $\theta = (\beta_1, \cdots, \beta_p)$ is unknown while $\eta = 0$ and $\sigma = 1$ are known. If $J(\theta)$ is any matrix with $J(\theta)J(\theta)^t = I(\theta)$, then for all θ*

$$\int H_b(J(\theta)^{-1}\kappa(x, \theta))\kappa(x, \theta)^t \mu_\theta^{p+1}(dx) = f(b)J(\theta)^t$$

and

$$\int H_b(J(\theta)^{-1}\kappa(x, \theta))H_b(J(\theta)^{-1}\kappa(x, \theta))\mu_\theta^{p+1}(dx) = f^*(b)Id$$

where f and f^* are continuous functions $\mathbb{R}^+ \to \mathbb{R}^+$ defined in (2.10) *and* (2.11) *below.*

PROOF. In this case, the information matrix has elements $I(\theta)_{jk} = E_\theta(X_j X_k)$. Therefore, if we put $Y_i = \sum_k (J(\theta)^{-1})_{ik} X_{p+1-k}$ $(i = 1, \cdots, p)$, then the Y_i are under μ_θ^p distributed according to $\mathcal{N}(0, Id)$. Note that

$$H_b(J(\theta)^{-1}\kappa(x, \theta))_i = \begin{cases} y_i u & \text{if } |u| \le b/|y| \\ b \cdot \operatorname{sign}(u) y_i/|y| & \text{if } |u| \ge b/|y| \end{cases}$$

where $u = x_{p+1} - \sum_k \beta_k x_{p+1-k}$. The distribution of $(Y_i, \cdots, Y_p, U)$ is independent

of θ, and we conclude by symmetry that

$$\int H_b(J(\theta)^{-1}\kappa(x,\theta))_i(J(\theta)^{-1}\kappa(x,\theta))_j\mu_\theta^{p+1}(dx) = 0 \quad (i \neq j),$$

$$= 2\int \cdots \int y_i^2\left(\int_0^{b/|y|} u^2\phi(u)\,du + \frac{b}{|y|}\int_{b/|y|}^{\infty} u\phi(u)\,du\right)$$

$$\cdot \Pi\phi(y_i)\,dy_i \ (i = j)$$

where ϕ is the standard normal density. From this, the first assertion follows immediately with

$$f(b) = 2p^{-1}\int_0^{\infty} z\left(\int_0^{bz^{-1/2}} u^2\phi(u)\,du + bz^{-1/2}\int_{bz^{-1/2}}^{\infty} u\phi(u)\,du\right)\chi_p^2(z)\,dz \tag{2.10}$$

where χ^2 is the chi-squared density. The proof of the second assertion is similar. We obtain

$$f^*(b) = 2p^{-1}\int_0^{\infty} z\left(\int_0^{bz^{-1/2}} u^2\phi(u)\,du + \frac{b^2}{z}\int_{bz^{-1/2}}^{\infty} \phi(u)\,du\right)\chi_p^2(z)\,dz. \tag{2.11}$$

□

The desired $A(\theta)$ is therefore a constant times $J(\theta)^{-1}$. The choice of the constant will determine the value of γ^{**} and γ^{***} respectively. Since we can use any matrix multiple of the influence function for the definition of the optimal estimator, we get the following set of equations

$$\sum_{i=1}^{n-p} w_{b_1}(d(x_i, \cdots, x_{i+p-1};\hat\theta)\hat u_{i+p}/\hat\sigma)(x_{i+j} - \hat\eta)\hat u_{i+p} = 0$$

$$(j = 0, \cdots, p-1) \tag{2.12}$$

$$\sum_{i=p+1}^{n-p} w_{b_2}(\hat u_i/\hat\sigma)\hat u_i = 0 \tag{2.13}$$

$$\sum_{i=p+1}^{n} w_{b_2}(\hat u_i/\hat\sigma)^2\hat u_i^2 = (n-p-1)\alpha\cdot\hat\sigma^2. \tag{2.14}$$

Here

$$d(x_1, \cdots, x_p;\theta) = \left(\sum_{i,j}(x_i - \eta)(x_j - \eta)(R(\theta)^{-1})_{ij}\right)^{1/2},$$

$$R(\theta)_{ij} = \mathrm{Cov}_\theta(X_i, X_j) \quad (i, j = 1, \cdots, p), \qquad \hat u_i = x_i - \hat\eta - \sum_{k=1}^{p}\hat\beta_k(x_{i-k} - \hat\eta)$$

$$w_b(x) = \min\left(1, \frac{b}{|x|}\right) \quad (x \in \mathbb{R}), \qquad \alpha = \int x^2 w_b(x)^2\phi(x)\,dx.$$

We summarize the properties of this estimator as follows.

THEOREM 2.6. *$\hat\eta$, $\hat\sigma$, $\hat\beta$ are asymptotically independent under the model μ_θ. The*

β-component of IC^T_{cond} is

$$f(b_i)^{-1}w_{b_1}(d(x_1, \cdots, x_p; \theta)u_{p+1}/\sigma)u_{p+1}R(\theta)^{-1}(x_1 - \eta, \cdots, x_p - \eta)^t,$$

the asymptotic covariance for $\hat{\beta}$ is $f(b_1)^{-2}f^*(b_1)R(\theta)/\sigma^2$ and the standardized gross-error-sensitivities for β are $\gamma^{**} = b_1 f(b_1)^{-1}$, $\gamma^{***} = b_1 f^*(b_1)^{-1/2}$. The estimator for β is optimal among all estimators with the same or smaller values for γ^{**} or γ^{***} respectively.

The proof is a straightforward application of previous results.

First let us discuss shortly the choice of cut-off constant b_1. Instead of giving a value of γ^{**} or γ^{***}, it is common in robust statistics to choose b_1 such that the efficiency $f(b_1)^2 f^*(b)^{-1}$ is acceptable. In our implementation of the estimator (2.12)–(2.14) we determined b_1 such that $f(b_1)^{-2}f^*(b_1)$ is between 1.05 and 1.15. Values of b_1 and coresponding γ^{**} are given in Table 1 of Künsch (1983). For known $\beta_1, \cdots, \beta_p$, (2.13)–(2.14) is a location-scale problem for i.i.d. variables and we can use a standard value of b_2. (2.13) was chosen such that we also have an optimality property of $\hat{\eta}$. For $\hat{\sigma}$ there is no optimality, but (2.13)–(2.14) corresponds to Huber's "Proposal 2" and we use it for computational simplicity.

The equation (2.12) is the analogue of what one gets in Schweppe's form of bounded influence regression, see Krasker-Welsch (1982). The weights used for the $(p+1)$ − tuple $x_i, \cdots, x_{i+p}$ depends on how well $x_i, \cdots, x_{i+p-1}$ fit to the model and how big the estimated innovation $\hat{u}_{i+p}$ is. However, we don't have separate weights for $x_i, \cdots, x_{i+p-1}$ and for $\hat{u}_{i+p}$ as in Mallows form. A large $|\hat{u}_{i+p}|$ causes no downweighting if $x_i, \cdots, x_{i+p-1}$ fit well and vice versa. This property gives us a high efficiency at the true model. The only difference with robust regression is that here the model determines also the distribution of the "independent variables" $x_i, \cdots, x_{i+p-1}$ which simplifies the problem. Finally, we mention that we would get weights $w_{b_1}((\text{const.} + d(x_i, \cdots, x_{i+p-1}; \hat{\theta})^2)^{1/2}\hat{u}_{i+p}/\hat{\sigma})$ if η is not considered as a nuisance parameter and we bound the total influence on η and $\beta_1 \cdots \beta_p$. This means that we have to downweight if $\hat{u}_{i+p}$ is large although $x_i, \cdots, x_{i+p-1}$ fits well.

2.5 *An example.* The author has written an ALGOL program for the solution of (2.12)–(2.14). As an example, we compare here maximum likelihood with our robust estimator for a series consisting of 91 monthly interest rates of an Austrian bank (data kindly provided by W. Polásek, Vienna). They are plotted in Fig. 1, which is reprinted from Künsch (1983) with permission of the publisher, and have been analyzed already in Künsch (1983). From Table 1 we see that there are tremendous differences between the maximum likelihood and the robust estimates for β and σ. The series contains three large outliers for the months number 18, 28 and 29. So we wondered what might happen if we replaced these outliers by 9.85 which is close to the values nearby. From Table 2 we see that the maximum likelihood estimates are moving closer to the robust estimates which are almost the same as before. However the estimates for σ still differ by a factor 1.6. A closer inspection of the data shows that this difference is due to 5

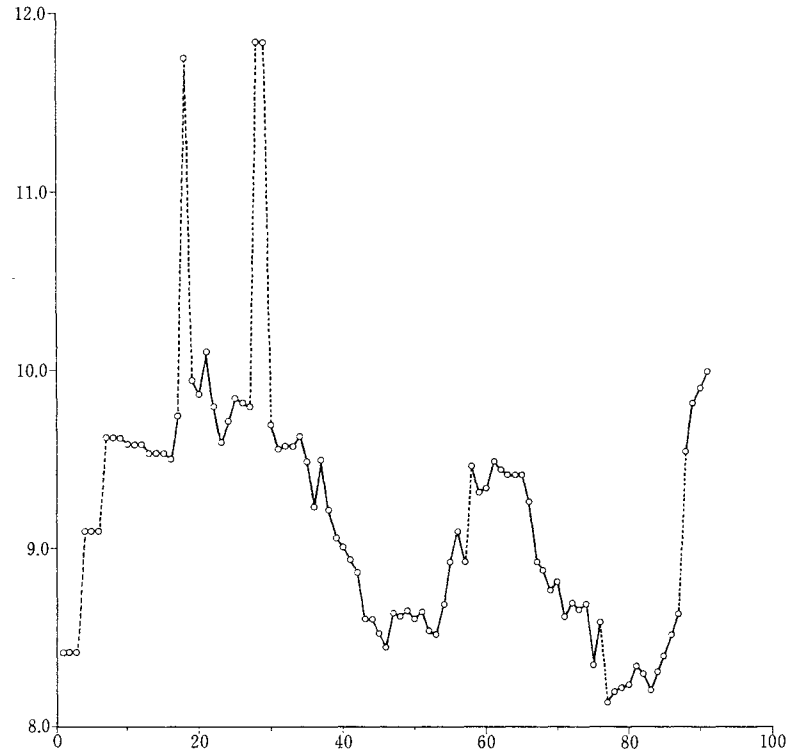

FIG. 1. *Monthly interest rate during* 91 *months. (Dotted lines indicate doubtful parts in the data).*

TABLE 1

Parameter estimates obtained by fitting an AR(p)*-process to the data of Fig.* 1

$p = 1$	b_1	b_2	$\hat{\beta}$	$\hat{\eta}$	$\hat{\sigma}$
	∞	∞	.789	9.19	.443
	2.5	1.5	.958	9.18	.154
	2.5	1.0	.959	9.11	.133
	1.5	1.5	.959	9.18	.154
	1.5	1.0	.959	9.11	.133

$p = 2$	b_1	b_2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\eta}$	$\hat{\sigma}$
	∞	∞	.766	.030	9.19	.442
	2.9	1.5	.977	−.023	9.19	.154
	2.9	1.0	.993	−.032	9.14	.127
	1.7	1.5	1.001	−.035	9.24	.154
	1.7	1.0	1.003	−.031	9.16	.125

large jumps in the series (indicated by dotted lines). This phenomenon is much less obvious than the three big outliers, but it is surely important to know that the innovations are usually small with some occasional large values. This example shows that our robust estimators are insensitive to gross errors and that they

TABLE 2
Parameter estimates for the data of Fig. 1 after correction of the three large outliers. $p = 1$

b_1	b_2	$\hat{\beta}$	$\hat{\eta}$	$\hat{\sigma}$
∞	∞	.923	9.12	.212
2.5	1.5	.958	9.21	.137
2.5	1.0	.959	9.16	.126
1.5	1.5	.964	9.23	.136
1.5	1.0	.965	9.17	.125

can draw our attention also to other irregularities in the data which are otherwise easily overlooked.

2.6 *Open problems.* We have found the optimal estimators under a bound on $IC_T^{\text{cond}}(x, \theta)$ in some metric. One might wonder what happens if we bounded instead $\int IC_T(x, \theta)\nu^m(dx)$ over all $\nu^m \in \mathcal{M}_{\text{stat}}^m$. For our optimal estimator

$$\sup_x |\, IC_T^{\text{cond}}(x, \theta)\,| = \sup_{\nu^{p+1}} \left| \int IC_T(x, \theta)\nu^{p+1}(dx) \right|,$$

we just take $\nu^{p+1} = \delta_{(c,\dots,c)}$ and let c go to infinity. However the following example shows that we can decrease the asymptotic covariance without increasing $\sup_{\nu^{p+1}} |\int IC_T(x, \theta)\nu^{p+1}(dx)|$.

EXAMPLE 2.7. We take a Gaussian AR(1)-process with $\eta = 0$ and $\sigma = 1$ known and fix a $\theta > 0$. If $\psi(x_1, x_2) = H_c(x_1(x_2 - \theta x_1))$, then ψ can be varied freely not only in $\{|x_1(x_2 - \theta x_1)| < c\}$, but also in $\{0 < x_1 < 2(c\theta)^{1/2}, x_1 < x_2\}$ without increasing $\sup_{\nu^2} |\int \psi(x_1, x_2)\nu^2(dx_1, x_2)|$. Namely, if (y_1, y_2) is a point in the above domain and $\nu^2 \in M_{\text{stat}}^2$ has mass near (y_1, y_2), then ν^2 must have equal mass somewhere near the line $x_2 = y_1$. But on this line $\psi(x_1, x_2) \le y_1^2/4\theta < c$. We therefore choose a function ω concentrated on this domain with

$$\int \omega(x_1, x_2)\phi(x_2 - \theta x_1)\, dx_2 = 0,$$

$$\int \omega(x_1, x_2)(x_2 - \theta x_1)\phi(x_2 - \theta x_1)\, dx_2 = 0$$

and

$$\int \omega(x_1, x_2)\psi(x_1, x_2)\phi(x_2 - \theta x_1)\, dx_2 \le 0, \quad < 0$$

for x_1 in a set of positive measure.

We do this for all $\theta > 0$ and put $\psi_\varepsilon = \psi + \varepsilon\omega$. If ε is small enough, the estimator given by ψ_ε has a smaller variance, but the same bound for the asymptotic bias as the estimator given by ψ.

It would be interesting to know how much smaller the asymptotic covariance can be, but we guess that the improvement will be small. In addition, the bound

on the integrated influence function is not quite satisfactory for the reasons given in Section 1.5.

Another open question is if a similar theory can be obtained without the restriction that the estimator depends on some finite dimensional marginal. We do not know if an analogue to Theorem 1.1 holds, but at least for M-estimators the generalization is straightforward. If $\hat{\theta}_n$ is given by

$$\sum_{i=1}^{n} \psi_i(x_i, x_{i-1}, \cdots, x_1; \hat{\theta}_n) = 0 \tag{2.15}$$

where ψ_i converges in some sense to a $\psi: \mathbb{R}^\infty \times \Theta \to \mathbb{R}^q$, then $\hat{\theta}_n$ converges μ – a.s. to $T(\mu)$ where T is defined by $\int \psi(x, T(\mu))\mu(dx) = 0$. Moreover under some regularity

$$\lim \varepsilon^{-1}(T((1-\varepsilon)\mu_\theta + \varepsilon\nu) - T(\mu_\theta)) = \int IC_T(x, \theta)\nu(dx)$$

with

$$IC_T(x, \theta) = E_\theta[(\partial/\partial\theta)\psi]^{-1}\psi(x, \theta).$$

However, the big problem comes from the fact that the segment $(1-\varepsilon)\mu_\theta + \varepsilon\nu$ is not a good approximation to the arc of a contamination model. (1.11) does not hold and $\varepsilon^{-1}(\mu_\varepsilon - \mu_0)$ does in general not converge to a finite signed measure.

It is clear that an infinite range will be needed for ARMA models, but it is of interest also in the autoregressive case. For instance, the asymptotic variance of an estimator depends in general on the whole distribution of the process. It is therefore an open problem how to generalize the change of variance function to autoregressive processes. Another problem concerns the estimation based on robust filtering, see Kleiner et al. (1979), and Martin and Thomson (1982). It is not an M-estimator in our sense here, and it is not clear if it is of the type (2.15) with ψ_i converging to some ψ. Moreover it seems to have a bias for the true model. For all these reasons we see no possibility of comparing this method with the optimal estimator here on a theoretical basis.

3. Proof of Theorems 1.1–1.3.

3.1 *Proof of Theorems* 1.1 *and* 1.2. We start with Theorem 1.2 since its proof is easier and we will use it for the proof of Theorem 1.1. So let us assume that $\int t(x)\nu^m(dx) = 0$ for all $\nu^m \in \mathscr{M}^m_{\text{stat}}$. We put $g(x_1, \cdots, x_{m-1}) = \sum_{k=1}^{m-1} t(x_k, \cdots, x_{m-1}, 0, \cdots, 0)$. For this g we have

$$\begin{aligned} t(x_1, \cdots, x_m) - g(x_1, \cdots, x_{m-1}) + g(x_2, \cdots, x_m) \\ = [t(x_1, \cdots, x_m) + g(x_2, \cdots, x_m) + \textstyle\sum_{k=1}^{m-1} t(0, \cdots, 0, x_1, \cdots, x_k)] \\ - [g(x_1, \cdots, x_{m-1}) + \textstyle\sum_{k=1}^{m-1} t(0, \cdots, 0, x_1, \cdots, x_k)]. \end{aligned}$$

In order to see that the first bracket is zero, we choose as $\nu^m\rho(y, 2m-1)^m$ where $y_i = x_i (i \le i \le m)$ and $y_i = 0$ $(m < i \le 2m-1)$, while for the second bracket we choose as $\nu^m\rho(z, 2m-2)^m$ where $z_i = x_i$ $(1 \le i < m)$ and $z_i = 0$ $(m \le i \le 2m-2)$. This completes the proof of Theorem 1.2.

The "only if" part of Theorem 1.1 is obvious. For the "if" part, we are going to use a result saying that every affine continuous function on a compact convex subset K of a topological vector space E can be approximated uniformly on K by continuous linear functionals on E, see Meyer (1966), Chapter XI, T6. Inour case we choose $E = \mathscr{M}' = \{\text{finite signed measures on } \mathbb{R}^m\}$ and as K the set $\{\nu^m \in \mathscr{M}^m_{\text{stat}}, \nu^m \text{ is concentrated on } [-N, N]^m\}$. By Prohorov's Theorem, K is weakly compact. But any continuous linear functional on $\mathscr{M}'$ is of the form $\nu^m \to \int t\, d\nu^m$ with t continuous and bounded, see Huber (1981), Lemma 2.2.1. So let t_j be a sequence of bounded continuous functions such that $\sup_{\nu^m \in k} | L(\nu^m) - \int t_j\, d\nu^m | \leq j^{-1}$. Obviously we may replace t_j by

$$\tilde{t}_j(x_1, \cdots, x_m) = t_j(x_1, \cdots, x_m) - \sum_{k=1}^{m-1} t_j(x_k, \cdots, x_{m-1}, 0, \cdots, 0) + \sum_{k=2}^{m} t_j(x_k, \cdots, x_m, 0, \cdots, 0).$$

Then we have for $k > j$, $| \tilde{t}_j(x_1, \cdots, x_m) - \tilde{t}_k(x_1, \cdots, x_m) | \leq 4/j$ for all $x \in [-N, N]^m$, as one can see easily if one chooses as ν^m the same measures as in the proof of Theorem 1.2. So $\tilde{t}_j$ converges on $[-N, N]^m$ uniformly to a function t which is therefore continuous.

Because of the special choice of $\tilde{t}_j$, $\tilde{t}_j(x)$ is constant on the set $\{x \in \mathbb{R}^m, x_m = 0\}$. Therefore $t(x) = t(0, \cdots, 0) = L(\delta_0)$ for all x with $x_m = 0$. Looking at the construction of the function g in Theorem 1.2, we see that we will get an extension of t when we take $N' > N$. So we have constructed a continuous function t on $\mathbb{R}^m$ such that $L(\nu^m) = \int t(x)\nu^m\,(dx)$ for all $\nu^m \in M^m_{\text{stat}}$ with compact support.

Because L is weakly continuous and affine, it is easy to show that $C = \sup_{\nu^m} | L(\nu^m) | < \infty$, cf. the proof of Lemma 2.2.1 of Huber (1981). Choosing as ν^m the measure $\rho(y, 2m-1)^m$ with $y_i = x_i$ $(1 \leq i \leq m)$, $y_i = 0$ $(m < i \leq 2m-1)$ we get

$$| t(x_1, \cdots, x_m) | \leq (2m-1)C + \sum_{k=1}^{m-1} | t(0, \cdots, 0, x_1, \cdots, x_k) | + (m-1)C$$

because t is equal to $L(\delta_0)$ for all x with $x_m = 0$. But if $x_m \neq 0$, then each argument of t in the sum on the right-hand side has more zero components than $x = (x_1, \cdots, x_m)$. Hence we may conclude by induction on the number of nonzero components of x that t is bounded. But then it follows immediately that $L(\nu^m) = \int t(x)\nu^m\,(dx)$ for all $\nu^m \in \mathscr{M}^m_{\text{stat}}$ and the theorem is proved.

3.2 *Proof of Theorem* 1.3. First we show that g is unique up to a constant. For this it is sufficient to prove that $f = 0$ implies $g = \text{const}$. Iterating the equation for g we find

$$g(X_1, \cdots, X_{m-1}) = E[n^{-1} \sum_{i=1}^{n} g(X_i, \cdots, X_{i+m-2}) \mid X_1, \cdots, X_{m-1}]. \tag{3.1}$$

By the ergodic theorem, the average on the right-hand side converges in L_1 to $\int g(x)\mu^{m-1}\,(dx)$ so $g(X_1, \cdots, X_{m-1}) = \int g(x)\mu^{m-1}\,(dx)$ a.s. By the continuity of g this means $g = \text{const}$.

For the existence we introduce the transition operator P. It takes bounded continuous functions on $\mathbb{R}^k (k \geq p)$ in bounded continuous functions on $\mathbb{R}^{\max(k-1,p)}$

and is defined by

$$(3.2)\quad Pf(x_1, \cdots, x_{k-1}) = \int f(x_1, \cdots, x_k)\mu\ (dx_k \mid x_{k-1}, \cdots, x_{k-p}) \quad (k > p)$$

$$(3.3)\quad Pf(x_1, \cdots, x_p) = \int f(x_2, \cdots, x_{p+1})\mu\ (dx_{p+1} \mid x_p, \cdots, x_1) \quad (k = p).$$

Furthermore we put $V_k(x_1, \cdots, x_k) = 1 + \sum_{j=1}^{k} |x_j|$ and

$$C_V^k = \{f\colon \mathbb{R}^k \to \mathbb{R} \text{ continuous}, \|f\|_V = \sup_x |f(x)|/V_k(x) < \infty\}.$$

Since

$$PV_k(x_1, \cdots, x_{k-1}) \leq 1 + \sum_{j=1}^{k-1} |x_j| + \sum_{j=1}^{p} |\beta_j| |x_{k-j}| + E|U_i|$$

$$\leq \text{const. } V_{k-1}(x_1, \cdots, x_{k-1}),$$

P can be extended to an operator from C_V^k to $C_V^{\max(k-1,p)}$.

It is then sufficient to show that for all $k \geq p$ and all $f \in C_V^k$ with $\int f(x)\mu^k\ (dx) = 0$ there is a $g \in C_V^k$ such that

$$(3.4)\quad g(x_1, \cdots, x_p) - Pg(x_1, \cdots, x_p) = f(x_1, \cdots, x_p) \quad (k = p)$$

$$(3.5)\quad g(x_1, \cdots, x_k) - Pg(x_2, \cdots, x_k) = f(x_1, \cdots, x_k) \quad (k > p).$$

(In order to find the g to the original f of Theorem 1.3, we solve (3.4) − (3.5) with Pf instead of f).

In the next step we show that the case $k > p$ can be reduced to the case $k = p$. Namely we put

$$(3.6)\quad g(x_1, \cdots, x_k) = \tilde{g}(x_{k-p+1}, \cdots, x_k) + \sum_{j=0}^{k-p-1} P^j f(x_{j+1}, \cdots, x_k)$$

where $\tilde{g}$ is the solution of

$$(3.7)\quad \tilde{g}(x_1, \cdots, x_p) - P\tilde{g}(x_1, \cdots, x_p) = P^{k-p}f(x_1, \cdots, x_p).$$

It is an easy calculation to show that (3.5) follows from (3.6) and (3.7).

For the proof that (3.4) has a solution, we use an idea due to S. Kotani (compare the similar argument in Kotani, 1983). We introduce the operator Q: $C_b(\mathbb{R}_p) \to C_b(\mathbb{R}^p)$ defined by

$$(3.8)\quad Qf(x) = V_p(x)^{-1}P(V_p f)(x)$$

(C_b = space of bounded continuous functions with supremum norm). (3.4) is then equivalent to

$$(3.9)\quad \tilde{g}(x) - Q\tilde{g}(x) = \tilde{f}(x)$$

where $\tilde{f} = f \cdot V_p^{-1}$, $\tilde{g} = g \cdot V_p^{-1}$. Here $\tilde{f} \in J = \{\tilde{f} \in C_b; \int \tilde{f}(x)V_p(x)\mu^p\ (dx) = 0\}$ and $Q\colon J \to J$. By Lemma 3.1 below, Q satisfies the conditions of Theorem VIII. 8.3, page 711, of Dunford-Schwartz (1959) and obviously the same is true for Q restricted to J. So because $\lambda = 1$ can be at most a pole of order one, we only have

to show that $\tilde{g} = Q\tilde{g}$, $\tilde{g} \in J$, implies $\tilde{g} = 0$. But this follows from (3.1) by the same argument as in the uniqueness proof for g.

LEMMA 3.1. *The operator Q of* (3.8) *has the following properties*:
(i) $\| Q^n \| \leq C$ *for all* n.
(ii) *there is an* n *and a compact operator* K *such that* $\| Q^n - K \| < 1$.

PROOF. By the special form of P we obtain

$$| Q^n f(x) | \leq \| f \| V_p(x)^{-1} P^n V_p(x)$$
$$\leq \| f \| V_p(x)^{-1}(1 + \textstyle\sum_{j=1}^{p} | \alpha_{nj}^* \| x_{p+1-j} | + E | X_i |)$$

where α_{nj}^* are the coefficients defined in Anderson (1971), 5.21. Since $| \alpha_{nj}^* | \to 0$ as $n \to \infty$, (i) is obvious. Let n be such that all $| \alpha_{nj}^* | < 1$. To any $\varepsilon > 0$ we then can find a $\psi \in C_b$ with compact support, $0 \leq \psi \leq 1$ and such that $(1 - \psi(x))Q^n 1(x) \leq \varepsilon$. Put $Rf(x) = f(x)\psi(x)$. Then $\| (1 - R)Q^n \| \leq \varepsilon$ and RQ^nR is compact. Moreover

$$Q^n = RQ^nR + (1 - R)Q^nR + Q^n(1 - R).$$

From this it follows easily that there is a compact operator K such that $\| Q^{2n} - K \| < 1$ if ε is small. □

Acknowledgment. I wish to thank R. Maronna for his introduction to robustness in time series given at ETH and to F. Hampel and W. Stahel for illuminating discussions on robust statistics. I am also grateful to the Japan Society for the Promotion of Science for the opportunity to work out and discuss these results in Japan. Finally I thank S. Kotani for the help with Theorem 1.3 and R. D. Martin and a referee for useful comments.

REFERENCES

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.
BOENTE, G., MARONNA, R. and YOHAI, V. (1982). Qualitative robustness in stochastic processes. Preprint, University of Buenos Aires.
BUSTOS, O. (1982). General M-estimates for contaminated p-th order autoregressive processes: Consistency and asymptotic normality. *Z. Wahrsch. verw. Gebiete* **59** 491–504.
DUNFORD, N. and SCHWARTZ, J. T. (1958). *Linear Operators, Part* I. Interscience, New York.
HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, University of California, Berkeley.
HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.
HAMPEL, F., MARAZZI, A., RONCHETTI, E., ROUSSEEUW, P., STAHEL, W. and WELSCH, R. E. (1982). Robust statistical methods. Handouts for an instructional meeting in Palermo, Italy.
HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.
KLEINER, B., MARTIN, R. D. and THOMSON, D. J. (1979). Robust estimation of power spectra. *J. Roy. Statist. Soc. B* **41** 313–351.
KOTANI, S. (1983). Limit theorems of hypoelliptic diffusion processes. In *Probability Theory and Mathematical Statistics*. Eds. K. Itô and J. V. Prokhorov. *Lecture Notes in Math.* **1021**. Springer, Berlin.

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

KÜNSCH, H. (1983). Robust estimation for autoregressive processes. *Proc. Institute for Statist. Math.* **31** 51–64. In Japanese, with summary and titles of tables in English, available upon request.

MARTIN, R. D. and THOMSON, D. J. (1982). Robust-resistant spectrum estimation. *Proc. IEEE* **10** 1097–1115.

MEYER, P. A. (1966). *Probabilites et Potentiel.* Hermann, Paris.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression.* Addison, Reading, Mass.

PAPANTONI-KAZAKOS, P. and GRAY, R. (1979). Robustness of estimators on stationary observations. *Ann. Probab.* **7** 989–1002.

STAHEL, W. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. thesis ETH Nr. 6881.

FACHGRUPPE FÜR STATISTIK
ETH-ZENTRUM
CH-8092 ZURICH, SWITZERLAND