ROBUST ESTIMATION OF LOCATION IN
A GAUSSIAN PARAMETRIC MODEL

Donald W. K. Andrews

ABSTRACT

This chapter considers the problem of optimal robust estimation of location in a model with stationary Gaussian parametric distributions. A Huber-type minimax criterion is used, in which minimaxing takes place over neighborhoods of the parametric Gaussian distributions. Four different neighborhood systems are considered. The neighborhoods include distributions of strong mixing processes whose univariate distributions are within (1) Hellinger, (2) $\varepsilon$-contamination, (3) variational, or (4) Kolmogorov neighborhoods of the parametric normal distributions. In addition, the neighborhood systems allow for deviations from stationarity and from the Gaussian structure of dependence.

For the Hellinger neighborhoods, an estimator is found that is within $\varepsilon$ of being asymptotically efficient at the Gaussian parametric distributions and is within $\varepsilon$ of being optimally robust. Qualitatively...
distinct results are obtained with the other three larger neighborhood systems. For these neighborhoods, optimal robust estimators are found to be bounded influence estimators with \( \psi \) functions that are very nearly of Huber shape. These estimators are quite robust against different “amounts” of dependence.

I. INTRODUCTION

The standard method of classical statistical estimation theory is to specify a parametric model, which is assumed to hold exactly, and then to derive the best estimator for the unknown parameters, according to some specified optimality criterion. Often a decision theoretic framework is used. The fruitfulness of this approach is reflected by its widespread usage. Parametric methods do have some drawbacks, however. In consequence, the field of robust statistics has emerged. It attempts to extend the usefulness and relevance of parametrically based models.

To illustrate some of the issues dealt with in robust statistics, consider the standard linear model. This model is based on the following assumptions:

1. Linearity in the parameters: \( Y = X\theta + U \), where \( X \) is a fixed known \( n \times k \) matrix of regressor variables, \( \theta \) is an unknown \( k \)-vector of parameters, \( U \) is an \( n \)-vector of random variables, and \( Y \) is an \( n \)-vector of observations.

2. Identical distributions: \( U_j, j = 1, \ldots, n \), are identically distributed.

3. Independence: \( U_j, j = 1, \ldots, n \), are independent.

4. Distributional shape: \( U_j \sim N(0, \sigma^2) \).

The special case where \( X \) is an \( n \times 1 \) vector of ones and \( \theta \) is a scalar parameter yields the location model: \( Y_j = \theta + U_j, j = 1, \ldots, n \).

Under assumptions 1–4, estimators of \( \theta \) and \( \sigma^2 \) can be derived, viz. the least-squares (LS) estimators, that are optimal according to reasonable criteria (e.g., see Andrews and Phillips, 1987). In practice, none of the above assumptions can be expected to hold exactly. It is possible, however, that they hold more or less “approximately.” The question arises: Does “approximate” validity of the assumptions still imply optimality or even “approximate” optimality of the best estimator for the exact model? This question has been investigated in some detail with respect to assumption 4, shape of the error distribution. The answer given by results in the field of robust statistics is
Robust Estimation of Location

no [see Huber (1981) and Hampel et al. (1986) for discussions and references].

Statisticians have studied most intensely the location model (for obvious reasons of simplicity). In this case, the sample mean \( \bar{Y} \) is the LS estimator and is optimal under assumptions 1-4. When the error distribution is not normal but has fatter tails, the use of \( \bar{Y} \) is found to be ill advised due to its sensitivity to outliers. The drawbacks of the mean in the location model also arise with the LS estimator in the more general linear model. As a result, numerous alternative estimators have been suggested that are designed to be robust under deviations from 4, the shape assumption [again see references in Huber (1981) and Hampel et al. (1986)].

In contrast to the voluminous literature dealing with shape robustness, the literature concerning robustness against distributional shape and against dependence (assumption 3) is somewhat limited. This is true in spite of the fact that several prominent statisticians have expressed the belief that the problem of robustness against distributional shape and dependence is worthy of much more intensive scrutiny (see Box, 1977; Cox, 1977; Bickel, 1978). For example, Box (1977) states, “emphasis on non-normality of the marginal distribution has been rather overdone and... considerably more attention should be paid to the assumption of independence.” This is particularly important for applications made to the analysis of economic time series data. In most situations, assumption 3 has no a priori justification from experience or theory and, in some situations, is patently false. Often this assumption may not be even “approximately” valid, as has been widely recognized by econometricians.

Although the literature on robustness with dependent data is limited, it is growing and is attracting considerable attention in the statistics community. The recent invited paper to the *Annals of Statistics* by Martin and Yohai (1986) on “Influence Functionals for Time Series” is a prime example. Other related papers include Gastwirth and Rubin (1975), Koul (1977), Portnoy (1977, 1979), Denby and Martin (1979), Kleiner et al. (1979), Papantoni-Kazakos and Gray (1979), Martin (1980), Andrews (1982, 1986), Küensch (1984), and Bustos and Yohai (1986).

Most of the above papers consider robust estimation of parameters of dependence, such as autoregressive-moving average parameters. A different problem is that of robustly estimating parameters of the marginal distributions of random variables (rv’s) when the rv’s are not necessarily independent. In this context, the existence of temporal
dependence of the rv's constitutes a nuisance that serves to complicate estimation. Given the prevalence of temporal dependence in economic data, problems of this nature arise quite frequently. Such problems are addressed in the present chapter. For simplicity, the location model is considered. More complicated multiparameter models are left unexplored. They present a challenge for future work.

The problem we consider is to derive an optimal robust estimator of the location parameter \( \theta \) in the model

\[
Y_j = \theta + U_j, \quad j = 1, \ldots, n,
\]

where the errors \( U_j \) are dependent rv's. Our approach extends that of Bickel (1981), who considers iid models. Parametric distributions of the infinite process \( \{ Y_j \} \equiv (Y_1, Y_2, \ldots) \) are specified and neighborhoods of these parametric distributions are constructed. In fact, four different neighborhood systems are considered—each corresponding to a distinct model.

The parametric distributions are mean \( \theta \) stationary strong mixing Gaussian distributions. The neighborhoods contain strong mixing distributions that are "close" to the Gaussian distributions in the sense that their univariate marginal distributions are close in terms of (1) Hellinger (H) distance, (2) \( \epsilon \)-contamination (C), (3) variational (V) distance, or (4) Kolmogorov (K) distance and their bivariate distributions are close in terms of some weak-convergence-inducing metric. With any of the four models, the true distribution of \( \{ Y_j \} \) is assumed to lie in the neighborhood of the parametric distribution. The true distribution of \( Y_1, \ldots, Y_n \) then is just the distribution of the first \( n \) rv's of the infinite process \( \{ Y_j \} \). These neighborhood systems allow for the relaxation of the assumptions of normality, independence, and stationarity of the rv's.

The univariate contamination and variational neighborhoods used in this chapter can be given gross error interpretations [for the latter a theorem of Strassen (1965) is used]. In the present model, however, the neighborhoods are not intended to be interpreted solely as the consequence of gross errors. The neighborhoods are designed to reflect the belief that the parametric model is merely a close approximation of the true distribution of the process \( \{ Y_j \} \), due to the complexity of nature. Thus, the parametric model is an oversimplification of reality, of which the existence of gross errors is only one possible cause.

The estimation problem under consideration is cast in an asymptotic framework. An optimal robust estimator is determined using a criterion of minimax asymptotic risk (following Huber, 1964). That is,
the optimal estimator minimizes, over a class of estimators, the asymptotic maximum risk over all distributions in the neighborhood system. In the present models, this includes minimaxing over a pre-specified set of correlation matrices of the parametric distributions. The use of the minimax principle requires careful design of the size and character of the neighborhood system. The size of the neighborhoods considered here is made to shrink to zero as the sample size \( n \) tends to infinity. Thus, for each \( n \) we have a neighborhood system of the whole process \( \{ Y_n \} \) and as \( n \) increases the size of the neighborhoods shrinks. This modeling construct is somewhat analogous to that of local power used in testing theory. It has been used extensively in iid robust models by Huber-Carol (1970), Jaeckel (1971), Beran (1977a,b, 1980), Rieder (1979, 1980, 1981a,b), Bickel (1981, 1982, 1983), Wang (1979, 1981), and Holmes (1981). The contiguity concept of LeCam (1960) underlies these neighborhood models.

The class of estimators considered here is that of weighted M-estimators. Within this class, an estimator is found that is within \( \varepsilon \) of being optimally robust for Hellinger neighborhoods and is within \( \varepsilon \) of being asymptotically efficient for the parametric model. Thus, robustness and efficiency are reconciled in the dependent Hellinger model, as they are in the iid Hellinger models of Beran (1977a) and Bickel (1981). The \( \varepsilon \)-optimal estimator has a defining equation with equal weights and \( \psi \) function given by the normalized score function, viz., \( y - \theta \), truncated at a very high point. This estimator gives the same estimate as the sample mean, except in samples with extremely large outliers, and, hence, is essentially the same as the sample mean.

It may be argued that the univariate Hellinger neighborhoods are too small. With the three models that have larger univariate neighborhoods, viz. contamination, variational, and Kolmogorov neighborhoods, qualitatively distinct results are obtained. In fact, the optimal estimator for Hellinger neighborhoods has arbitrarily large risk for certain distributions in these larger neighborhoods.

The optimal estimators for the C, V, and K models are very closely related, and, hence, it suffices to describe the optimal weighted M-estimator for the C model. It has equal weights and \( \psi \) function of the form

\[
\psi(y) = \begin{cases} 
\begin{align*}
\varepsilon & \quad \text{for } y > b \\
-\gamma & \quad \text{for } |y| < b, \text{ for a constant } b < \infty \\
-\gamma & \quad \text{for } y < -b.
\end{align*}
\end{cases}
\]
Thus, the optimal estimator has a bounded influence function. In fact, the dependent robust model considered here is closely related to the bounded influence method of Hampel [see Hampel et al. (1986)].

The shape of the optimal $\psi$ function for $y \in [-b, b]$ is given by the solution to a calculus of variations problem. This problem yields a constrained integral equation. The solution of this integral equation is given as an infinite expansion in terms of orthonormal polynomials. The coefficients of this expansion, the truncation height $\gamma$, and the constant $b$ are calculated numerically below. The appearance of an integral equation solution to the calculus of variations problem, as opposed to a closed-form solution, is directly attributable to the dependence of the errors.

The optimal estimator is found numerically to be nearly linear for $y \in [-b, b]$ and to be very well approximated by a Huber estimator (see Huber, 1964) whose truncation height depends upon the “size” of the univariate neighborhoods and upon the least-favorable dependence structure. It is shown that the truncation height of the optimal estimator (or of the optimal Huber estimator) is relatively insensitive to the least favorable dependence structure, unless considerable positive dependence is allowed. Thus, the optimal $\psi$ function for dependent rv’s is remarkably similar to that for independent rv’s.

A clear dichotomy exists between the optimal $\psi$ function for the Hellinger model and those for the other three models. This stems from the relatively small size of the Hellinger neighborhoods. Whether or not the Hellinger neighborhoods are too small depends, of course, on the unknown true distribution. If there is a possibility that the true distribution lies outside of a Hellinger neighborhood of the Gaussian distribution, then the C, V, or K neighborhoods clearly are more appropriate.

For brevity, the results presented in this chapter concern the case in which the parametric distributions are mean $\theta$ stationary Gaussian distributions. The neighborhood systems and theoretical results can be extended, however, to more general single parameter stationary models (see Andrews, 1982).

The remainder of this chapter is organized as follows: Section II sets up the problem to be solved. It contains definitions, regularity conditions, and a formal statement of the problem. Section III gives the theoretical results, i.e., the determination of the optimal robust estimators for the different models. Section IV describes numerical results, including comparisons of the performance of the optimal estimators with the simpler Huber estimators and with the optimal
nonrobust estimator, under a variety of scenarios. Section V draws conclusions. An Appendix (Section VI) contains proofs of results given in Section III.

II. STATEMENT OF THE PROBLEM

2.1. The model under consideration is given in eq. (1). The rv's $Y_1, \ldots, Y_n$ are observed and $\theta$ is an unknown parameter to be estimated. The parameter space, $\Theta$, is an open subset of $R$.

2.2. Parametric distributions, $\Phi_{\Omega}$, of the infinite process $\{Y_j\}$ are specified to be mean $\theta$, (known) variance $\sigma^2$, stationary, strong mixing, Gaussian distributions with (infinite dimensional) correlation matrix $\Omega$. $\Omega \in S$, a specified set of correlation matrices defined in Section 2.3 below. $\Phi_{\Omega}$ satisfies the mixing condition M1X defined in Section 2.5. Under $\Phi_{\Omega}$, the univariate marginal distribution of $Y_j$ (for any $j$) is denoted by $\Phi_\theta$ and the bivariate marginal distribution of rv's $Y_j$, $Y_k$ of distance $r$ apart (i.e., $|j - k| = r$) is denoted $\Phi_{\Omega\rho}$. The densities of $\Phi_\theta$ and $\Phi_{\Omega\rho}$, with respect to Lebesgue measure, are denoted by $\phi_\theta$ and $\phi_{\Omega\rho}(\cdot, \cdot, \rho_j)$, respectively, where $\rho = (\rho_1, \rho_2, \ldots)'$ is the correlation vector corresponding to $\Omega$.

2.3. Each correlation matrix $\Omega$ is completely characterized by its correlation vector $\rho$. We specify restrictions, in terms of the vector $\rho$, on the matrices $\Omega$ over which minimaxing is to take place. Correlation matrices $\Omega$ of parametric distributions $\Phi_{\Omega\rho}$ must lie in the set $S$, where $S$ is defined by

$$S = \{\Omega: \Omega \text{ is a positive definite, semiinfinite, Toeplitz matrix with unit main diagonal elements and } j \text{th diagonal elements } \rho_j, \text{ where } |\rho_j| \leq \rho_j^b, \forall j \geq 1\},$$

where $\rho^b = (\rho_1^b, \rho_2^b, \ldots)'$ is some specified positive definite correlation vector (i.e., $\rho^b$ corresponds to a positive definite Toeplitz matrix $\Omega^b$) with $\rho_j^b > 0, \forall j$. (By semiinfinite we mean $\Omega$ corresponds to a process on the positive integers.) $\rho^b$ and $\Omega^b$ are referred to as the boundary correlation vector and matrix, respectively.

Comments: (i) For example, the boundary correlation vector can be taken as that corresponding to any stationary autoregressive-moving average (ARMA) process.

(ii) The boundary correlation vector is specified on the basis of experience in the same manner as is the choice of parametric family.
The results of Section IV, however, show that the optimal robust estimator is not sensitive to the choice of $\rho^b$, unless quite large "amounts" of correlation are considered.

2.4. Four different models are considered, each corresponding to a different sequence of neighborhoods of the parametric distributions $\Phi_{\theta}$. The four sequences of neighborhoods are denoted by $\{C_{\theta}(\theta), \Omega_n\}_n$, for $i = H$ (Hellinger), $C$ (contamination), $V$ (variational), and $K$ (Kolmogorov). Each neighborhood consists of distributions $F_{\theta_n}$ of a real-valued process $\{Y_j\}$ with the properties specified below. Let $F_{\theta_n}$ be the univariate marginal distribution of $Y_j$ under $F_{\theta_n}$ for $j = 1, 2, \ldots$. Let $F_{\theta_n}^{j,k}$ denote the bivariate distribution of $Y_j, Y_k$ under $F_{\theta_n}$. The density of $F_{\theta_n}$, with respect to a specified measure, is denoted by $f_{\theta_n}$, $\forall j = 1, 2, \ldots$.

A distribution $F_{\theta_n}$ is in the $n$th type-$i$ neighborhood of $\Phi_{\theta}$, i.e., $F_{\theta_n} \in \{C_{\theta}(\theta), \Omega_n\}_n$ for $i = H, C, V, or K$, if:

ASSUMPTION F1: $F_{\theta_n}$ is the distribution of a strong mixing, real-valued process on the positive integers with mixing numbers $\tilde{\alpha}(j)$, $j = 1, 2, \ldots$, where $\{\tilde{\alpha}(j)\}$ satisfy condition MIX defined below.

ASSUMPTION F2: $F_{\theta_n} \in C_{\theta}(\theta), \forall j = 1, 2, \ldots$, where $C_{\theta}(\theta)$ is a neighborhood of the univariate parametric distribution $\Phi_{\theta}$ defined by

(a) Contamination ($i = C$):

$C_{\theta}(\theta) = \{F: F(y) = (1 - t/\sqrt{n})\Phi_{\theta}(y) + H(y)t/\sqrt{n}, for H(\cdot) \text{ arbitrary}\}$,

where $F(\cdot)$ and $H(\cdot)$ are distribution functions (dfs) on $R$.

(b) Hellinger, variational, and Kolmogorov ($i = H, V, and K$):

$V_{\theta}(\theta) = \{F: \Delta_v(F, \Phi_{\theta}) < t/\sqrt{n}\}$,

where

$\Delta_v(F, \Phi_{\theta}) = \left( \int (\sqrt{f} - \sqrt{\Phi_{\theta}})^2 \, d\mu \right)^{1/2}$,

$\Delta_v(F, \Phi_{\theta}) = \sup \{|F(A) - \Phi_{\theta}(A)|: A \text{ is a Borel set in } R\}$

$= \frac{1}{2} \int |f - \phi_{\theta}| \, d\mu$,

$\Delta_k(F, \Phi_{\theta}) = \sup_{y \in R} |F(y) - \Phi_{\theta}(y)|$.

and $f$ is the density of $F$ with respect to some measure $\mu$ that dominates $F$ and $\Phi_{\theta}$. 
ASSUMPTION F3: The bivariate distributions of $F_{\bar{X}_n}$ lie in $\Delta_w$-metric neighborhoods of $\Phi_{x\Omega}$, where $\Delta_w$ is any fixed but arbitrary weak-convergence-inducing metric on $R^2$, and the sizes of such neighborhoods are $o(1)$ as $n \to \infty$. That is, for some $d(n)$ that satisfies $\lim_{n \to \infty} d(n) = 0$, $\Delta_w(F_{\bar{X}_n}, \Phi_x) \leq d(n)$ for all $r$ and for all $j$, $k$ such that $|j - k| = r$.

Comments: (i) The univariate neighborhoods are related as follows:

\[ \mathcal{F}^c_{\bar{X}_n}(\theta) \subset \mathcal{F}^v_{\bar{X}_n}(\theta) \subset \mathcal{F}^k_{\bar{X}_n}(\theta). \]

(ii) The neighborhoods as defined are quite rich. In particular, all univariate distributions in $\mathcal{F}^c_{\bar{X}_n}(\theta)$ are attained as the univariate marginal distributions of some distribution in the neighborhoods (see Andrews, 1982).

(iii) The use of shrinking neighborhoods is justified as follows. The objective is to embed a fixed sample size problem in a sequence of problems so that the asymptotic properties of estimators approximate those of the finite sample as closely as possible. If the "size" of the neighborhood is not a function of the sample size, then the asymptotic bias dominates the asymptotic variance, provided asymmetric distributions are allowed in the neighborhoods. If only symmetric neighborhood distributions are considered, then the asymptotic bias is zero (for reasonable estimators) and the asymptotic variance is the sole criterion of performance. Since the symmetry assumption is artificial, we are left with the unsatisfactory result that bias and variance are of different orders of magnitude. On the other hand, if the "size" of the neighborhood is made to shrink at the rate $O(1/\sqrt{n})$ as $n \to \infty$, then the bias and variance are of the same order of magnitude. In this case both the bias and the variance appear as criteria of performance asymptotically, which reflects the finite sample situation.

2.5. The following mixing condition is assumed to hold:

ASSUMPTION MIX: $\Phi_{x\Omega}$ and $F_{\bar{X}_n}$, where $F_{\bar{X}_n} \in \{\mathcal{F}^c_{\bar{X}_n}(\theta), \Omega \}$, $\forall n$, $\forall \theta \in \Theta$, $\forall \Omega \in S$, are distributions of sequences of strong mixing rv's with mixing numbers generically denoted by $\bar{\alpha}(m)$, $m = 1, 2, \ldots$. There exists a sequence of dominating numbers $\{\alpha(m)\}$ such that $\bar{\alpha}(m) \leq \alpha(m) \forall m, \forall \Phi_{x\Omega}$, and $\forall F_{\bar{X}_n}$ and $\sum_{m=1}^{\infty} \alpha(m) < \infty$ for some $\tau < 1$.

2.6. Let $\{F_{\bar{X}_n}\}$ denote an arbitrary sequence of neighborhood distributions $F_{\bar{X}_n}$, $n = 1, 2, \ldots$. That is, $F_{\bar{X}_n} \in \{\mathcal{F}^c_{\bar{X}_n}(\theta), \Omega \}$, $\forall n$,
and $\Omega \in \mathcal{S}$ where $i = H, C, V,$ or $K$. Let $\text{unif}(\theta, F, \Omega)$ abbreviate “uniformly over $\theta \in \Theta_C$, where $\Theta_C$ is any compact set in $\Theta$, uniformly over distributions $\{F_{\theta_n}\}$ in the type-$i$ neighborhood, and uniformly over $\Omega \in \mathcal{S}$,” for $i = H, C, V,$ or $K$.

2.7. Define $T = \{T_n\}$ to be a weighted $M$-estimator of $\theta$ for some Lebesgue measurable function $\psi: \Theta \times R \to R$ and for some triangular array of weights $\{d_{nj}\}$, if

$$T_n = \begin{cases} 
\text{Closest root of} \sum_{j=1}^n d_{n}\psi(\theta, y) = 0 & \text{to} \ T_n \\
T_n & \text{if no root exists},
\end{cases}$$

where $T = \{T_n\}$ is some estimator of $\theta$ that is consistent $\text{unif}(\theta, F, \Omega)$. Such an estimator exists (see Andrews, 1982).

**Comment:** The class of equally weighted M-estimators is well known (see Huber, 1964). When the observations are iid the order statistics are sufficient and the Rao–Blackwell theorem implies that it suffices to consider only estimators that are invariant under permutations of the data. Equally weighted M-estimators have this property. In the dependent case, however, the order statistics are not sufficient and it is possible that estimators that are not invariant have advantageous properties. For finite sample sizes the UMVU estimator is the generalized least-squares estimator, which is not invariant under permutations of the data and is not an M-estimator. It is a weighted $M$-estimator, however, with weights that are a function of the correlation matrix $\Omega$. This estimator motivates the use of the class of weighted M-estimators.

The following assumptions on $\psi$ hold when specified:

**Assumption A1:** $\psi \in \Psi = \{\psi: \Theta \times R \to R | E_{\theta_0}\psi(\theta, Y) = 0, E_{\theta_0}\psi(\theta, Y) \cdot (Y - \theta) / \sigma^2 = 1, \text{and} E_{\theta_0}\psi^2(\theta, Y) < \infty, \forall \theta \in \Theta\}$.

**Assumption A2:** $\sup_{\theta \in \mathcal{R}} |\psi(\theta, y)| \leq C_1(\theta)\), where $C_1(\theta)$ is bounded for $\theta \in \Theta_C$, any compact set in $\Theta$.

**Assumption A3:** $\sup_{\theta \in \mathcal{R}} |\psi(\theta + h_1, y) - \psi(\theta + h_2, y)| \leq C_1(\theta) \cdot \|h_2 - h_1\| \text{ for all } |h_1|, |h_2| < \varepsilon, \text{ where } \varepsilon > 0 \text{ is fixed and where } C_2(\theta) \text{ is bounded for } \theta \in \Theta_C, \text{ any compact set in } \Theta$.

**Assumption A4:** $\psi(\theta, y) = \psi(0, y - \theta) [\equiv \psi(y - \theta)]$. 

Comments: (i) The first two parts of assumption A1 are merely normalization conditions. The third part of A1 must hold if T is to have finite asymptotic variance under the parametric model, and, hence, is not restrictive.

(ii) Assumption A2 is restrictive, although any unbounded $\psi$ function may be approximated by a sequence of bounded $\psi$ functions.

(iii) Assumption A3 is only mildly restrictive since it corresponds to a desirable property for an estimator to have, viz. bounded local shift sensitivity as defined by Hampel (1974).

(iv) Assumption A4 implies that the estimator is location equivariant with probability that converges to 1 as $n \to \infty$. This condition is not needed for most of the theoretical results of the chapter, but it simplifies the numerical calculations of Section IV considerably.

For the $K$ neighborhood model the following two stronger assumptions replace assumptions A2 and A3:

**Assumption A2’:** There exists a finite, signed measure $v_\theta$ and some constant $C_2$ such that $\psi(\theta, y) = v_\theta((-\infty, y]) + C_2$ a.e. and $\|v_\theta\|_V \leq C_1(\theta)$, where $C_1(\theta)$ is bounded on compact sets and $\|\cdot\|_V$ is the variational norm.

**Assumption A3’:** $\|v_{\theta+h_2} - v_{\theta+h_1}\|_V \leq C_2(\theta)|h_2 - h_1|$ for $v_\theta$ as in A2’, where $C_2(\theta)$ is bounded on compact sets.

Comment: The existence of a measure $v_\theta$ in assumptions A2’ and A3’ is not restrictive since the minimand is finite if and only if there exists such a measure (see Bickel, 1981).

The triangular array of weights $\{d_{ij}\}$ is assumed to satisfy the following:

**Assumption D1:** $|d_{ij}| \leq B_j < \infty$, $\forall j = 1, \ldots, n; n = 1, 2, \ldots$

**Assumption D2:** The limit $\bar{d}_n(\equiv \frac{1}{n} \sum_{j=1}^{n} d_{ij})$, $n = 1, 2, \ldots$, exists and equals some constant $\tilde{\eta} \neq 0$.

Without loss of generality (see Andrews, 1982), we can take $\tilde{\eta} = 1$ and assume the following limits exist:

$$\eta = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} |d_{ij}|$$

and

$$\omega_r = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} (d_{ij} - \bar{d}_n)(d_{i(j+r)} - \bar{d}_n), \quad \forall r = 0, 1, 2, \ldots.$$
**Comment:** Triangular arrays that satisfy assumptions D1 and D2 exist in abundance. In fact, with probability one, the observed sequence from an ergodic process of bounded random variables satisfies the conditions.

2.8. Let \( \mathcal{M}_i \), for \( i = H, C, V, \) or \( K \), be classes of weighted M-estimators with weights \( \{d_n\} \) that satisfy assumptions D1 and D2 and \( \psi \) functions that satisfy assumptions A1, A2, and A3 for \( i = H, C, \) or \( V \) and assumptions A1, A2', and A3' for \( i = K \). Let \( \mathcal{M}_i^l \) be the location equivariant subset of \( \mathcal{M}_i \). That is, estimators in \( \mathcal{M}_i^l \) also satisfy assumption A4.

2.9. Let \( L: R \rightarrow [0, \infty) \) be a given loss function that satisfies assumptions L1 and L2 when specified:

**Assumption L1:** \( L \) is symmetric about 0 and increasing on \( R^+ \).

**Assumption L2:** \( L \) is convex.

2.10. The problem for \( i = H, C, V, \) or \( K \) can now be stated as follows. Find an estimator \( T = \{T_n\} \in \mathcal{M}_i^l \) of \( \theta \) that attains the infimum of the expression

\[
R_i(\psi, d, \theta, t, S) = \sup_{\theta > 0} \lim_{n \to \infty} \sup_{x \in \mathcal{S}} \left( L_n(T_n - \theta) \right)
\]

where \( L_n(x) = \min\{L(x), a\} \). That is, we want to find an estimator in \( \mathcal{M}_i^l \) that yields asymptotic minimax risk over the distributions in the neighborhoods.

### III. THEORETICAL RESULTS

#### A. Simplification of the Minimand

We now proceed to simplify the minimand \( R_i(\psi, d, \theta, t, S) \).

**Theorem 1:** For \( T \in \mathcal{M}_i^l \),

\[
\mathbb{E}_{\mathbb{P}_{\mathbb{F}_{\text{finn}}}} \left\{ \sqrt{n} \left[ T_n - \theta - \frac{1}{n} \sum_{j=1}^{n} d_n \lambda(\psi, \theta, F_{\text{finn}}) \right] \right\}
\]

\[
= N \left[ 0, \sum_{j=1}^{n} (1 + \omega_{ij}) E_{\mathbb{P}_{\mathbb{F}_{\text{finn}}}} \psi(\theta, Y) \psi(\theta, Y_{1+i}) \right]
\]

as \( n \to \infty \) unif(\( \theta, F, \Omega \)), for \( i = H, C, V, \) or \( K \), where \( \lambda(\psi, \theta, F) = E_{\mathbb{P}} \psi(\theta, Y) \), \( \mathbb{P}_{\mathbb{F}_{\text{finn}}} \) denotes the distribution (or law) of \( \cdot \) under \( F_{\text{finn}} \), and \( \Rightarrow \) denotes convergence in distribution.
Comments: (i) The proof of Theorem 1 is rather lengthy. For brevity, it is not given here, but is available from the author upon request. The proofs of all other results stated below are given in the Appendix.

(ii) The theorem is proved uniformly for \( F_{\theta, n} \in \{ \mathcal{F}_{\psi, \theta, n}^H(\theta), \Omega \}_n \), \( \forall n \), and uniformly for \( \Omega \in S \) because the minimand \( R_{\psi}(\psi, d, \theta, t, S) \) takes the supremum over \( \{ \mathcal{F}_{\psi, \theta, n}^H(\theta), \Omega \}_n \) and over \( S \), before taking the limit supremum as \( n \to \infty \). This ordering of suprema is necessary to give a meaningful finite sample interpretation to the results.

(iii) The theorem is proved uniformly for \( \theta \) in compact sets, again for the reason of finite sample interpretation. Uniformity is needed so that the sample size required for the asymptotic results to be applicable to a finite sample situation is not a function of the unknown parameter.

Theorem 1 is used in obtaining the following result:

**Theorem 2:** If \( T \in \mathcal{M} \) and \( L \) satisfies assumptions L1 and L2, then

\[
R(\psi, d, \theta, t, S) = E_{\psi, d}[b_i(\psi, d, \theta) + \sup_{\Omega \in S} \sigma(\psi, d, \theta, \Omega) \cdot Z],
\]

for \( i = H, C, V, \) or \( K \), where \( Z \sim N(0, 1) \) and

\[
b_H(\psi, d, \theta) = \eta \cdot 2t \cdot \left[ \int \psi^2(\theta, y) \, d\Phi(y) \right]^{1/2},
\]

\[
b_C(\psi, d, \theta) = \eta \cdot t \cdot \left[ \inf_{y \in \mathcal{R}} \psi(\theta, y) \right]^{1/2}
\]

\[
+ \sup_{y \in \mathcal{R}} \psi(\theta, y) \]

\[
+ \left[ \inf_{y \in \mathcal{R}} \psi(\theta, y) \right]^{1/2}
\]

\[
b_V(\psi, d, \theta) = \eta \cdot t \cdot \left[ \sup_{y \in \mathcal{R}} \psi(\theta, y) \right]^{1/2}
\]

\[
- \left[ \inf_{y \in \mathcal{R}} \psi(\theta, y) \right]^{1/2}
\]

\[
b_K(\psi, d, \theta) = \eta \cdot t \cdot \| \nu_0 \|_V,
\]

and

\[
\sigma^2(\psi, d, \theta, \Omega) = \sum_{j=-\infty}^{\infty} (1 + \omega_j) E_{\theta, \psi(\theta, Y_j)} \psi(\theta, Y_{j+1})
\]

Comment: The asymptotic maximum bias \( b_i(\psi, d, \theta) \) is independent of the correlation matrix \( \Omega \), whereas the asymptotic maximum variance is a function of \( \Omega \).

Next we calculate the infimum over the triangular arrays of weights \( d_{ij} \).
THEOREM 3: If $T \in \mathcal{M}$, and $L$ satisfies assumptions L1 and L2, then
\[
\inf_{(\omega_0, \omega_1) \in \mathcal{D}_1 \text{ and } d \in \mathcal{D}_2} R_*(\psi, \theta, t, S) = R_*(\psi, d^*, \theta, t, S),
\]
(11)
where $d^* = \{d_{ij}^*\}$ is any triangular array of identical nonzero constants, for $t = H, C, V, \text{ and } K$.

Without loss of generality we can take $d_{ij}^* = 1, \forall i, \forall n$. Then, define $R_*(\psi, \theta, t, S) = R_*(\psi, d^*, \theta, t, S)$, $b_*(\psi, \theta) = b_*(\psi, d^*, \theta)$, and $\sigma^2(\psi, \theta, \Omega) = \sigma^2(\psi, d^*, \theta, \Omega)$, for $i = H, C, V, \text{ and } K$. The asymptotic maximum bias and the asymptotic variance simplify to
\[
b_*(\psi, \theta) = 2t \cdot \left[ \int \psi^2(\theta, y) d\Omega_0(y) \right]^{1/2},
\]
\[
b_*(\theta, \psi) = t \cdot \text{ess sup}_{y \in R} |\psi(\theta, y)|,
\]
\[
b_*(\psi, \theta) = t \cdot [\text{ess sup}_{y \in R} \psi(\theta, y) - \text{ess inf}_{y \in R} \psi(\theta, y)],
\]
\[
b_*(\theta, \psi) = t \cdot \|v_0\|_v,
\]
and
\[
\sigma^2(\psi, \theta, \Omega) = \sum_{j = -\infty}^{\infty} E_{\theta_0} \psi(\theta, Y_j) \psi(\theta, Y_{j+1}).
\]

The boundary correlation matrix $\Omega^b$ is found to be least favorable.

THEOREM 4: For $\psi$ satisfying assumption A2,
\[
\sup_{t \in \mathcal{N}} \sigma^2(\psi, \theta, \Omega) = \sigma^2(\psi, \theta, \Omega^b).
\]
(12)

Comments: (i) This shows that the (perhaps unrealistic) inclusion of $\Omega$ matrices in $S$ with large amounts of negative correlation is irrelevant.

(ii) Suppose the true $\Omega$ is known, call it $\Omega^T$. Then, the optimal estimator for the case of known covariance matrix can be calculated by taking $\Omega^b = \Omega^T$.

At this point, we need to treat the cases where $i = H$ and $i = C, V$, and $K$ separately.

B. Hellinger Neighborhood Results

Define $H*(\psi, \theta, t, S) = E_\psi L[b_*(\psi, \theta) + \sigma(\psi, \theta, \Omega^b) \cdot Z]$. Under the assumptions of Theorem 3 [which, in particular, require $\psi(\theta, y)$ to be
bounded and satisfy a Lipschitz condition, we have

\[ H(\psi, \theta, t, S) = R_{H}(\psi, \theta, t, S). \]  \hspace{1cm} (13)

We proceed to solve problem (3) for \( i = H \) by minimizing \( H(\psi, \theta, t, S) \) without imposing the boundedness and Lipschitz conditions on \( \psi \) and then adjusting the solution to get an \( \epsilon \)-optimal solution that satisfies these conditions. Equation (13) implies that such a solution also is within \( \epsilon \) of minimizing \( R_{H}(\psi, \theta, t, S) \) and is an \( \epsilon \)-optimal solution to problem (3).

**Theorem 5:** For \( L \) satisfying assumptions L1 and L2, \( \xi(\theta, y) \equiv y - \theta \) minimizes \( H(\psi, \theta, t, S) \) with respect to \( \psi \) subject to

\[ A1(a): \quad E_{h}\psi(\theta, Y) = 0 \]

and

\[ A1(b): \quad E_{h}\psi(\theta, Y) \cdot (Y - \theta)/\sigma^2 = 1. \]

**Comments:** (i) \( A1(a) \) and \( A1(b) \) are the normalization conditions of assumption A1.

(ii) The estimator based on \( \xi(\theta, y) \) is the sample mean and is the maximum likelihood estimator for the iid \( \Phi_{h} \) model. Thus, this estimator is asymptotically efficient for the iid \( \Phi_{h} \) model. In fact, this estimator is asymptotically efficient for all parametric distributions \( \Phi_{\theta} \) (provided \( \Omega \) is such that \( \Phi_{\theta} \) has a positive, continuous spectral density) (see Grenander and Rosenblatt, 1957, Chapter 7).

The function \( \xi(\theta, y) \) that minimizes \( H(\cdot, \theta, t, S) \) is unbounded, and, hence, does not satisfy assumption A2. The following lemma is useful for deriving an estimator whose \( \psi \) function is *bounded*, satisfies conditions \( A1(a) \) and \( A1(b) \), and has \( H(\cdot, \theta, t, S) \) value arbitrarily close to \( H(\xi, \theta, t, S) \). Let \( r \) be any function from \( R \) to \( R \). We introduce a truncated, shifted, and scaled transform of \( r \), denoted \( r_{\gamma} \), such that \( r_{\gamma} \) is bounded and satisfies \( A1(a) \) and \( A1(b) \). Denote

\[ [r(y)]_{\gamma}^{-} = \begin{cases} \gamma & \text{if } r(y) > \gamma \\ r(y) & \text{if } |r(y)| \leq \gamma \\ -\gamma & \text{if } r(y) < -\gamma. \end{cases} \]

\[ (14) \]

Now define \( r_{\gamma}(y) = \{[r(y)]_{\gamma}^{-} - \beta_{\gamma}\}/\alpha_{\gamma} \), where \( \beta_{\gamma} = E_{h}[r(Y)]_{\gamma}^{-} (< \infty) \) and \( \alpha_{\gamma} = E_{h}\{[r(Y)]_{\gamma}^{-} - \beta_{\gamma}\} \cdot (Y - \theta)/\sigma^2. \)
Lemma 1: Suppose \( r: R \to R, r \in L^{2(1-\tau)}[\Phi_b] \) where \( \tau < 1 \) is as in MIX, \( r \) satisfies \( E_{\cdot b}(r(Y) \cdot (Y - \theta)/\sigma^2) = 1 \) and \( E_{\cdot b}(r(Y)) = 0 \), and \( L \) satisfies assumptions L1 and L2, then for \( r \) as defined above

\[
\lim_{\gamma \to \infty} H(r, \theta, \gamma, t, S) = H(r, \theta, t, S). \tag{15}
\]

Applying Lemma 1 to the function \( \xi(\theta, y) \), we get a bounded function, viz. \( \xi_{\gamma}(\theta, y) \), that is within \( \varepsilon \) of minimizing \( H(\cdot, \theta, t, S) \) subject to A1(a) and A1(b) for \( \gamma \) sufficiently large. In particular, \( \xi_{\gamma}(\theta, y) = [y - \theta]_{-\gamma}^\gamma / \alpha_{\gamma} \), where \( \alpha_{\gamma} = E_{\Phi_b}[Y]_{-\gamma}^\gamma \cdot \gamma / \sigma^2 \), and \( \beta_{\gamma} = 0 \) by symmetry of \( \Phi_b \) and antisymmetry of \( \xi(\theta, y) \) about \( \theta \). \( \xi_{\gamma}(\theta, y) \) is just a truncated version of the (normalized) score function of \( \Phi_b \). Note that the calculation of specific estimates using the estimator defined by \( \xi_{\gamma} \) does not require calculation of \( \alpha_{\gamma} \), since estimates based on \( \alpha_{\gamma} \cdot \xi_{\gamma} \) are identical to those based on \( \xi_{\gamma} \).

Corollary 1: For \( L \) satisfying assumptions L1 and L2 and given \( \varepsilon > 0 \), the estimator \( T \) corresponding to \( (\psi, d) = (\xi_{\gamma}, d^*) \) satisfies

\[
R_H(\xi_{\gamma}, d^*, \theta, t, S) \leq \inf_{(\psi, d) \text{ corresponds to some } \Phi_b} R_H(\psi, d, \theta, t, S) + \varepsilon
\]

for any given \( t \geq 0 \), for \( \gamma \) sufficiently large.

Comments: (i) The conclusion of Corollary 1 is striking. The estimator corresponding to \( (\xi_{\gamma}, d^*) \) is not only \( \varepsilon \)-optimal for the robust model (the case where \( t > 0 \)), but is asymptotically \( \varepsilon \)-efficient at each parametric distribution \( \Phi_{b\Omega} \) (the case where \( t = 0 \)) for \( \Omega \in S \), provided \( \Omega \) has positive continuous spectral density. Thus, asymptotic efficiency and optimal robustness against distributional shape and dependence over time are attained simultaneously in this model. Bickel (1981) and Beran (1977a) obtain analogous results (with respect to distributional shape only) for the iid Hellinger neighborhood model.

(ii) Note that \( \xi(\theta, y) \) is independent of \( t, \sigma^2 \), and \( \Omega \). \( \xi_{\gamma}(\theta, y) \) also is independent of \( \Omega \), and depends on \( t \) and \( \sigma^2 \) only for the determination of the magnitude of \( \gamma \) necessary to make \( R(\xi_{\gamma}, d^*, \theta, t, S) \) within \( \varepsilon \) of the optimal value. If it is known that \( t \) and \( \sigma^2 \) are bounded above by some values \( t_0 \) and \( M \), respectively, then \( \xi_{\gamma}(\theta, y) \) also is independent of \( t \) and \( \sigma^2 \).
C. Contamination, Variational, and Kolmogorov Neighborhood Results

Now we continue the solution of problem (3) for the cases where \( i = C, V, \) or \( K \). The results obtained thus far are summarized by the following:

**Corollary 2:** For \( T \in \mathcal{M} \), and for \( L \) satisfying assumptions L1 and L2,

\[
R_c(\psi, \theta, t, S) = E_N L[b_c(\theta, \psi) + \sigma(\psi, \theta, \Omega^B) \cdot Z] \quad \text{for } t = C, V, \text{ and } K.
\]

(16)

Under assumption A4, \( \psi(\theta, y) = \psi(0, y - \theta) \). Define \( \psi(\theta, y) = \psi(0, y - \theta) \). The next result shows that only odd \( \psi(\cdot) \) functions need to be considered:

**Theorem 6:** If there exists a \( \psi(\cdot) \) function that satisfies assumptions A1–A4 and minimizes (16), then there exists an odd \( \psi(\cdot) \) function that does likewise, for \( i = C, V, \) or \( K \). If there only exists a sequence of functions \( \{\psi_q\} \) that approaches the infimum of (16) as \( q \to \infty \), then there also exists such a sequence of odd \( \psi(\cdot) \) functions.

For odd \( \psi \) functions,

\[
b_c(\psi, \theta) = 2t \cdot \text{ess sup} |\psi(0, y - \theta)| = 2b_c(\psi, \theta).
\]

(17)

Thus, the variational model with \( t = t'/2 \) has the same asymptotic maximum bias as the contamination model with \( t = t' \). Since the asymptotic variance is the same in both models,

\[
R_c(\psi, \theta, t', S) = R_v(\psi, \theta, t'/2, S).
\]

(18)

Further, if the optimal \( \psi \) function for the variational (or contamination) model, call it \( \psi^* \), is monotone (as it is for the numerical calculations carried out), then

\[
b_k(\psi^*, \theta) = b_v(\psi^*, \theta) \quad [= 2 \cdot b_c(\psi^*, \theta)],
\]

(19)

and

\[
R_k(\psi^*, \theta, t, S) = R_v(\psi^*, \theta, t, S).
\]

(20)

Since \( \{F^{V}_{n}(\theta), \Omega \}_n \subset \{F^{K}_{n}(\theta), \Omega \}_n \), \( \forall n, \forall \Omega \in S \), Eq. (20) implies that \( \psi^* \) also is optimal for the Kolmogorov model. Hence, it suffices to consider just the contamination model.

Minimizing (16) is easier if the doubly infinite sum of \( \sigma^2(\psi, \theta, \Omega^B) \) is truncated at a finite number of terms. By truncating at a sufficiently high point the optimal \( \psi \) function for the "truncated" minimand has
risk \( R_C(\psi, \theta, t, S) \) that is arbitrarily close to the minimum of (16) over \( T \in \mathcal{M}_C^\epsilon \).

**Lemma 2:** Suppose \( T \in \mathcal{M}_C^\epsilon \) and \( L \) satisfies assumptions L1 and L2. Given any \( \epsilon > 0 \), there exists an integer \( M \in (0, \infty) \) such that
\[
|\sigma_m^2(\psi, \theta, \Omega^\theta) - \sigma^2(\psi, \theta, \Omega^\theta)| < \epsilon \quad \text{and} \quad |H_{CM}(\psi) - R_C(\psi, \theta, t, S)| < \epsilon,
\]
(21)
where
\[
\sigma_m^2(\psi, \theta, \Omega^\theta) \equiv \sum_{j=-M}^M E_{\delta_{\theta j}} \psi(\theta, Y) \psi(\theta, Y_{i+j}) \quad (> 0)
\]
(22)
and
\[
H_{CM}(\psi) \equiv E_N L[b_C(\psi, \theta) + \sigma_m(\psi, \theta, \Omega^\theta) \cdot Z].
\]
(23)

By Lemma 2, we can solve
\[
\inf_{\psi, A_1, A_2, A_3 \text{ and } A_4 \text{ hold}} H_{CM}(\psi)
\]
(24)
to get an \( \epsilon \)-optimal solution to (3). Of course, if the parametric distributions are \( M \) dependent this solution is strictly optimal. For convenience, and without loss of generality by assumption A4, take \( \theta = 0 \). As above, \( \psi(y) \equiv \psi(0, y) \). The “truncated” problem (24) can be split into two parts. Let
\[
\Psi_\gamma = \{ \psi(y) : R \rightarrow R | \psi \text{ satisfies A1, A2, and A3 with } \theta = 0 \text{ and } \\
\text{ess sup}_{y \in R} |\psi(y)| \leq \gamma \}.
\]
(25)
Then, assuming L1 and L2 hold, Eq. (24) can be written as
\[
\inf_{\gamma \in R^+} E_N L[t \cdot \gamma + \inf_{\Psi_{\gamma}} \sigma_m(\psi, 0, \Omega^\theta) \cdot Z]
\]
(26)
by the argument of the proof of Theorem 2. Thus, we need to solve
\[
\inf_{\gamma \in \Psi_\gamma} \sigma_m^2(\psi, 0, \Omega^\theta).
\]
(27)
Once this is solved for different values of \( \gamma \), yielding solutions \( \psi_\gamma^* \), the optimal value of \( \gamma \), call it \( \gamma^* \), is simply the smallest value of \( \gamma \) that solves
\[
\inf_{\gamma \in R^+} E_N L[t \cdot \gamma + \sigma_m(\psi_\gamma^*, 0, \Omega^\theta) \cdot Z].
\]
(28)
Equation (28) is relatively easy to solve, at least for simple loss functions. For example, if loss is quadratic, Eq. (28) is just

$$\inf_{\gamma \in \mathbb{R}^+} \left[ r^2 \cdot \gamma^2 + \sigma^2_M(\psi_\gamma, 0, \Omega^b) \right].$$  \hfill (29)

Note that $\gamma^*$ must be finite (whether or not quadratic loss is assumed), since $R_C(\psi_\gamma, 0, t, S) \to \infty$ as $\gamma \uparrow \infty$, for all $\psi_\gamma$ under consideration.

Before attempting to solve problem (27) via standard calculus of variations methods, the question of existence must be addressed, since there may not exist a $\psi_\gamma$ function that satisfies all the constraints and attains the infimum.

**Lemma 3:** From the class of $\psi_\gamma$ functions that satisfy assumptions A1, A2, A4, and $\text{ess sup}_{y \in R} |\psi_\gamma(y)| \leq \gamma$ (but do not necessarily satisfy assumption A3), there exists an (a.s.) unique extremal function $\psi_\gamma^*$ for the problem

$$\inf_{\psi_\gamma} \sigma^2_M(\psi_\gamma, 0, \Omega^b).$$  \hfill (30)

Further, $\forall \varepsilon > 0, \quad \exists \psi_\gamma^{**} \in \Psi_\gamma$ (i.e., $\psi_\gamma^{**}$ also satisfies assumption A3) such that

$$\sigma^2_M(\psi_\gamma^{**}, 0, \Omega^b) < \inf_{\psi_\gamma} \sigma_M^2(\psi_\gamma, 0, \Omega^b) + \varepsilon.$$  \hfill (31)

Lemma 3 implies the existence of an (a.s.) unique extremum for the problem (27) provided the Lipschitz condition is ignored. Hence, the Euler–Lagrange necessary conditions for this calculus of variations problem can be applied. [See Gelfand and Fomin (1963) or Akhiezer (1962) for the classical calculus of variations theory.] For points $y \in R$ not on the boundary of the constraint

$$\text{ess sup}_{y \in R} |\psi_\gamma(y)| \leq \gamma,$$  \hfill (32)

a necessary condition for an extremum is

$$\psi_\gamma(y) = - (\lambda y / 2\sigma^2) - 2 \int_R \psi_\gamma(x) \frac{\varphi_\nu(y, x, \rho^b)}{\varphi_\nu(y)} \, dx.$$  \hfill (33)

where the multiplier $\lambda$ is such that $E_{\Phi_0} \psi_\gamma(Y)Y / \sigma^2 = 1$. This is a constrained integral equation of Fredholm second type. [Note that the truncation of the sum in Eq. (33) at $M$ has allowed the change in order of integration and summation that otherwise would not be justified.] Points $y \in R$ on the boundary of Eq. (32) are on the upper (lower) boundary if Eq. (33) holds with "=" replaced by "$\leq$" ("$\geq$").
Whether a point is on the boundary of Eq. (32) or not is determined by the coincident satisfaction of the integral Eq. (33) and the constraints. Since \( \psi \) is odd, Eq. (33) can be rewritten for \( y \) in \( R^+ \) (the set of positive real numbers) as

\[
\psi(y) = (-\lambda y/2\sigma^2) + g(y) + \int_{D_+} \psi(x)K_m(y, x)\,dx,
\]

(34)

where

\[
K_m(y, x) = -2 \sum_{j=1}^{M} [\varphi_0(y, x, \rho^b_j) - \varphi_0(y, -x, \rho^b_j)]/\varphi_0(y),
\]

(35)

\[
g(y) = \gamma \int_{D_+} K_m(y, x)\,dx - \gamma \int_{D_-} K_m(y, x)\,dx,
\]

(36)

and \( D_+ \), \( D_- \), and \( D^*_+ \) are the regions in \( R^+ \) where \( -\gamma < \psi(y) < \gamma \), \( \psi(y) = -\gamma \), and \( \psi(y) = +\gamma \), respectively. For \( y \in R^- \), \( \psi(y) = -\psi(-y) \). Manipulation of the integral Eq. (34) yields the following:

**Lemma 4:** The optimal \( \psi \) function, \( \psi^* \), which solves problem (30) of Lemma 3, satisfies

\[
\psi^*(y) = \begin{cases} 
\gamma & \text{for all } y \text{ sufficiently large} \\
-\gamma & \text{for all } y \text{ sufficiently small}.
\end{cases}
\]

(37)

**Comment:** The optimal \( \psi \) function, \( \psi^* \), is similar to the \( \psi \) function generated via the bounded influence approach (see Hampel, 1968, 1974)—both are constant for large and small values of \( y \).

The solution \( \psi^* \) to the constrained integral Eq. (33) can be expressed in terms of orthonormal functions. See Courant and Hilbert (1937, Chapter III, §5) for a discussion of this method. For notational convenience, let \( \langle s_1, s_2 \rangle \) denote \( \int_{D} s_1(y)s_2(y)\varphi_0(y)\,dy \) for any two functions \( s_1(\cdot) \) and \( s_2(\cdot) \) defined on \( D \). Let \( \xi_l(y) \) for \( y \in D, l = 1, 2, \ldots \), be the orthonormal polynomials (with respect to \( \varphi_0 \) over the region \( D \)) obtained by applying the Gram–Schmidt process to the polynomials \( 1, y, y^2, \ldots \). Then, by defining, \( \langle \xi_k, \xi_l \rangle = \delta_{kl} \) for \( k, l = 1, 2, \ldots \), where \( \delta_{kl} \) is the Kronecker delta function. By standard Fourier series results, the bivariate normal density \( \varphi_0(y, x, \pm \rho^b_j) \) can be written as

\[
\varphi_0(y, x, \pm \rho^b_j) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \xi_k(y)\xi_l(x)\varphi_0(y)\varphi_0(x)\rho_{kl}(\pm \rho^b_j), \quad \forall y, x \in D,
\]

(38)
where equality is in the sense of mean square convergence [against the
density \( \varphi_0(y) \varphi_0(x) \) over the region \( (D_j)^2 \)] and

\[
r_{kl}(\rho) = \int \int_{(D_j)^2} \xi_k(y)\xi_l(x) \varphi_0(y, x, \rho) \, dy \, dx.
\]  

(39)

Thus,

\[
K_m(y, x) = -2 \sum_{i=1}^{\infty} \sum_{l=1}^{\infty} \xi_k(y)\xi_i(x) \varphi_0(x) \tau_{il}
\]  

(40)

where \( \tau_{kl} = \sum_{i=1}^{\infty} [r_{il}(\rho_j^B) - r_{il}(-\rho_j^B)] \) for \( k, l = 1, 2, \ldots \). Note that \( \tau_{kl} \geq 0 \) because \( \rho_j^B \geq 0 \) implies \( r_{il}(\rho_j^B) \geq r_{il}(-\rho_j^B), \) \( \forall i, j, k, l. \) Define the following infinite dimensional vectors: \( \psi \equiv (\langle \psi, \xi_1 \rangle, \langle \psi, \xi_2 \rangle, \ldots, \rangle, \) \( g_j \equiv (\langle g_j, \xi_1 \rangle, \langle g_j, \xi_2 \rangle, \ldots, \rangle, \) \( Y \equiv (\langle Y/\sigma^2, \xi_1 \rangle, \langle Y/\sigma^2, \xi_2 \rangle, \ldots, \rangle, \) and \( \xi(y) \equiv [\xi_1(y), \xi_2(y), \ldots] \). Let \( Y \) be the semiinfinite matrix with \( (k, l) \) element \( \tau_{kl} \), for \( k, l = 1, 2, \ldots \).

**Theorem 7:** Let \( D_+, D_-, D^+_i \) be the regions in \( R^+ \) where \( \psi^*_\gamma \), the solution to problem (30) of Lemma 3, is unconstrained, contrained below, and constrained above, respectively. Then \( \psi^*_\gamma \) is given by

\[
\psi^*_\gamma(y) = \begin{cases}
\gamma & \text{for } y \in D^+_i \\
(-\lambda Y/2 - 2g_\gamma)(I + 2Y)^{-1} \xi(y) & \text{for } y \in D_i \\
\gamma & \text{for } y \in D^-_i,
\end{cases}
\]  

(41)

where \( I \) is the identity matrix, \( (I + 2Y)^{-1} \) is some generalized inverse of \( I + 2Y, \lambda = [2g_\gamma(I + 2Y)^{-1} Y + \gamma^T \Phi_0(D^+) - \Phi_0(D^-)] - 1] / Y(I + 2Y)^{-1} \) \( Y \) [where \( \Phi_0(D^+) \) and \( \Phi_0(D^-) \) are the \( \Phi_0 \) probabilities of \( D^+_i \) and \( D^-_i \), respectively], and the equality in (41) is in the sense of mean square convergence. For \( y \in R^-, \psi^*_\gamma(y) = -\psi^*_\gamma(-y) \).

**Comments:**

(i) The proof of this result follows the classical treatment of linear integral equations (see Courant and Hilbert, 1937, Chapter III).

(ii) By Lemma 3 the solution to problem (30) is (a.s.) unique. Thus, the sets \( D_+, D_-, D^+_i \) of Theorem 7 are (a.s.) unique. Also, given \( D_+, D^-_i, \) and \( D^+_i \), the integral Eq. (34) has an (a.s.) unique solution, provided \( I + 2Y \) is nonsingular. For the numerical calculations described below, singularity or near-singularity of \( I + 2Y \) never occurred.

(iii) If the \( \psi^*_\gamma \) function of Theorem 7 does not satisfy assumption A3, then a smoothed version of it, call it \( \psi^*_\gamma^* \), which does satisfy assumption A3, can be used (see Lemma 3 and its proof). Minimization of (28) over \( \gamma \) then yields an \( \varepsilon \)-optimal solution to problem (3) for
contamination neighborhoods. As discussed above, this immediately
gives the solution for variational neighborhoods and if the optimal
contamination neighborhood $\psi$ function is monotone, for Kolmogorov
neighborhoods as well. The next section presents the results of carrying
out these procedures numerically.

The analysis above assumes a known error variance $\sigma^2$ for the
specified parametric model. Using contiguity arguments, it should be
possible to extend these results (regarding optimal robust estimation
of $\theta$) to the case where $\sigma^2$ is estimated simultaneously with $\theta$. Since the
asymptotic covariance matrix for estimates of $\theta$ and $\sigma^2$ is block
diagonal for a wide class of estimation methods [e.g., see Huber (1964,
1981)], the asymptotic risk properties of $\hat{\theta}$ will be the same whether or
not $\sigma^2$ is estimated, provided $\hat{\sigma}^2$ is consistent.

IV. NUMERICAL RESULTS

The solution to problem (3) given by Eqs. (28) and (41) has been cal-
culated numerically for contamination neighborhoods and quadratic
loss. The results are given in Table 1 for a number of different
boundary correlation vectors $\rho^B$. More specifically, first-order auto-
regressive [AR(1)] and moving average [MA(1)] boundary correlation
vectors are considered with several different correlation parameters. $\sigma^2$
is taken to be 1. (For $\sigma^2 \neq 1$, the $\psi$ function just needs to be rescaled.)

The degenerate kernel method utilizing orthonormal expansions
was used to solve the integral Eq. (34) numerically. This method
consists of truncating the expansion of the solution (41) given in
Theorem 7 at a finite number of terms. For the degree of accuracy
given, it was necessary to calculate the first three terms of the
expansion. Hence, the optimal $\psi$ function for $y \in D_y$ is given by a
polynomial of degree three [without a constant term, since the solution
to (34) at $y = 0$ is necessarily 0]. The sets $D_y$, $D^{-}_y$, and $D^{+}_y$ and the
optimal truncation height $\gamma$ were found by search procedures. $D_y$, $D^{-}_y$, and $D^{+}_y$
are of the form $[0, b]$, the null set, and $[b, \infty)$, respectively.

Table 1 shows the optimal $\psi$ function corresponding to $\theta = 0$; for
$\theta \neq 0$, $\psi(\theta, y) = \psi(y - \theta)$ by assumption A4. Four values of $t$ (the
parameter that indexes the size of the neighborhoods) are considered,
viz. $t = 0.5, 1.0, 1.5$, and $2.0$. For samples of size 100, these correspond
to the "reasonable" contamination levels of 5, 10, 15, and 20%,
respectively. Bickel (1981) provides additional arguments that suggest
that this is a reasonable range of values for $t$. As discussed above, the
Table 1. $\psi$ Functions for the Optimal Estimators (E1) in the Contamination Model with Squared Error Loss

<table>
<thead>
<tr>
<th>Boundary correlation vector</th>
<th>$t$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$\gamma$</th>
<th>$b$</th>
<th>R1 risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1) $p_1^R = 0.0$</td>
<td>0.5</td>
<td>1.636</td>
<td>0.000</td>
<td>0.000</td>
<td>1.410</td>
<td>0.862</td>
<td>1.636</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.271</td>
<td>0.000</td>
<td>0.000</td>
<td>1.293</td>
<td>0.435</td>
<td>2.964</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
<td>0.000</td>
<td>0.000</td>
<td>1.207</td>
<td>0.256</td>
<td>5.000</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.634</td>
<td>0.000</td>
<td>0.000</td>
<td>1.259</td>
<td>0.165</td>
<td>7.787</td>
</tr>
<tr>
<td>AR(1) $p_1^R = 0.1$</td>
<td>0.5</td>
<td>1.632</td>
<td>0.000</td>
<td>0.000</td>
<td>1.411</td>
<td>0.864</td>
<td>1.859</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.271</td>
<td>0.000</td>
<td>0.001</td>
<td>1.293</td>
<td>0.435</td>
<td>3.186</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
<td>-0.002</td>
<td>0.006</td>
<td>1.267</td>
<td>0.256</td>
<td>5.223</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.634</td>
<td>-0.002</td>
<td>0.007</td>
<td>1.259</td>
<td>0.158</td>
<td>8.010</td>
</tr>
<tr>
<td>AR(1) $p_1^R = 0.3$</td>
<td>0.5</td>
<td>1.617</td>
<td>0.000</td>
<td>0.008</td>
<td>1.414</td>
<td>0.871</td>
<td>2.499</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.236</td>
<td>-0.001</td>
<td>0.011</td>
<td>1.294</td>
<td>0.441</td>
<td>3.829</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
<td>-0.006</td>
<td>0.025</td>
<td>1.267</td>
<td>0.256</td>
<td>5.867</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.636</td>
<td>-0.004</td>
<td>0.023</td>
<td>1.259</td>
<td>0.165</td>
<td>8.653</td>
</tr>
<tr>
<td>AR(1) $p_1^R = 0.5$</td>
<td>0.5</td>
<td>1.538</td>
<td>0.002</td>
<td>0.043</td>
<td>1.431</td>
<td>0.908</td>
<td>3.666</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.373</td>
<td>-0.023</td>
<td>0.098</td>
<td>1.297</td>
<td>0.486</td>
<td>5.010</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.631</td>
<td>-0.009</td>
<td>0.089</td>
<td>1.269</td>
<td>0.274</td>
<td>7.050</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.364</td>
<td>-0.005</td>
<td>0.088</td>
<td>1.259</td>
<td>0.165</td>
<td>9.839</td>
</tr>
<tr>
<td>AR(1) $p_1^R = 0.7$</td>
<td>0.5</td>
<td>1.233</td>
<td>0.016</td>
<td>-0.198</td>
<td>1.490</td>
<td>1.030</td>
<td>6.404</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.469</td>
<td>0.008</td>
<td>0.305</td>
<td>1.310</td>
<td>0.513</td>
<td>7.809</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.345</td>
<td>0.116</td>
<td>0.022</td>
<td>1.271</td>
<td>0.290</td>
<td>9.870</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.038</td>
<td>-0.003</td>
<td>0.396</td>
<td>1.260</td>
<td>0.179</td>
<td>12.664</td>
</tr>
<tr>
<td>MA(1) $p_1^R = 0.1$</td>
<td>0.5</td>
<td>1.636</td>
<td>0.000</td>
<td>0.000</td>
<td>1.410</td>
<td>0.862</td>
<td>1.837</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.271</td>
<td>0.000</td>
<td>0.001</td>
<td>1.293</td>
<td>0.435</td>
<td>3.164</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
<td>-0.002</td>
<td>0.006</td>
<td>1.267</td>
<td>0.256</td>
<td>5.201</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.636</td>
<td>-0.002</td>
<td>0.008</td>
<td>1.258</td>
<td>0.150</td>
<td>7.987</td>
</tr>
<tr>
<td>MA(1) $p_1^R = 0.3$</td>
<td>0.5</td>
<td>1.617</td>
<td>0.000</td>
<td>0.007</td>
<td>1.414</td>
<td>0.871</td>
<td>2.242</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.271</td>
<td>-0.001</td>
<td>0.011</td>
<td>1.294</td>
<td>0.446</td>
<td>3.572</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
<td>-0.005</td>
<td>0.022</td>
<td>1.267</td>
<td>0.258</td>
<td>5.609</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.636</td>
<td>-0.003</td>
<td>0.021</td>
<td>1.259</td>
<td>0.165</td>
<td>8.394</td>
</tr>
<tr>
<td>MA(1) $p_1^R = 0.499,99$</td>
<td>0.5</td>
<td>1.422</td>
<td>0.307</td>
<td>-0.171</td>
<td>1.427</td>
<td>0.903</td>
<td>2.662</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.285</td>
<td>-0.001</td>
<td>0.051</td>
<td>1.296</td>
<td>0.451</td>
<td>4.005</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.951</td>
<td>-0.006</td>
<td>0.077</td>
<td>1.267</td>
<td>0.256</td>
<td>6.045</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.635</td>
<td>-0.004</td>
<td>0.077</td>
<td>1.259</td>
<td>0.165</td>
<td>8.833</td>
</tr>
</tbody>
</table>

*The $\psi$ functions are odd and on $R^+$ are of the form

$$\psi(y) = \begin{cases} x_1y + x_2y^2 + x_3y^3 & \text{for } y \in [0, b) \\ \gamma & \text{for } y \in (b, \infty) \end{cases}$$

For $\theta \neq 0$, $\psi(\theta y) = \psi(y - \theta)$
contamination results for $t = t'$ correspond to the variational results for $t = t'/2$, thus Table 1 actually gives solutions for both the contamination and variational neighborhood models. Further, since the optimal $\psi$ functions given in Table 1 are monotone, the variational and Kolmogorov models have the same solution and the results of the table apply to the Kolmogorov model as well (with $t = t'/2$).

Table 1 gives the optimal $\psi$ functions and the corresponding minimum risk values (labeled R1 risk) for the different situations described above. As expected, $\gamma$ [i.e., the value of $\psi(y)$ for $y \in [b, \infty)$] and $b$ decrease in $t$ (for any given $\rho^b$). That is, as the size of the neighborhoods increases the truncation height decreases. For any given $t$, $\gamma$ and $b$ increase as the "amount" of dependence (as measured by the value of $\rho^b$) increases. Increased dependence inflates the asymptotic variance but not the asymptotic maximum bias. Hence, the trade-off between the two shifts in favor of increased weight given to the asymptotic variance as $\rho^b$ increases. It is worth noting, however, that the sensitivity of the optimal truncation height to the value of $\rho^b$ is slight for $\rho^b$ less than 0.7. A very large value of $\rho^b$ is needed to significantly increase the truncation height relative to the $\rho^b = 0$ optimal height.

Most of the optimal $\psi$ functions have slopes that are slightly increasing on $[0, b]$ or are decreasing very near zero and increasing from there to $b$. This contrasts with the constant slope on $[0, b]$ of the optimal, independent, robust $\psi$ function. The intuition to explain this pattern of change in slope is not clear. The change in slope is very small, however, for all cases except those with quite large $\rho^b$ value and even in those cases it is small. Generally speaking, the optimal $\psi$ functions are remarkably close to the Huber $\psi$ functions (see Huber, 1964) which are continuous, symmetric, linearly increasing in the middle of $R$, and constant in the tails. This is not surprising, especially for small $\rho^b$ values, since Huber $\psi$ functions are optimal in the above robust model with independent data (i.e., the model with $\Omega^b = I$; see Bickel, 1981).

A number of interesting questions can be asked concerning the relative performances of

(E1) the estimator that is optimally robust for dependent, Gaussian, parametric distributions for given $t$ and $\rho^b$ (see Table 1);

(E2) the estimator that is optimally robust for independent Gaussian parametric distributions (see Table 1, $\rho^b = 0.0$) which has $\psi(y) = (x_1 y)^{\gamma-1}$, where $\gamma$ and $x_1$ are constants that depend on $t$;
(E3) the Huber estimator with $\psi(y) = [x^T y]^2 \gamma$, where $\gamma$ is taken to minimize risk in the dependent robust model over all $\psi$ functions of the form $\psi(y) = [x^T y]^p \gamma$ (so $\gamma$ depends on $t$ and $\rho^B$, see Table 2); and

(E4) the sample mean, $\bar{y}_n = (1/n) \sum_{j=1}^{\infty} y_j$, which is (asymptotically) optimal for the (nonrobust) parametric Gaussian location model (regardless of the boundary correlation vector $\rho^B$, provided $\rho^B$ has a continuous positive spectral density).

The optimal Huber estimators E3 for the dependent robust model are given in Table 2 for different boundary correlation vectors. These estimators are normalized such that their defining $\psi$ function also is their influence curve (which yields $x$ not equal to one). These estimators are considered because of their simplicity. It is far easier to determine $\gamma$ numerically than to determine the strictly optimal estimator numerically, as has been done in Table 1.⁴ One of the questions of interest is how close an approximation of the optimal estimator E1 is the optimal Huber estimator E3.

The estimators can be compared for different underlying models. For each model, a comparison of two estimators can be made by looking at their ratio of risks. As in Table 1, loss is taken to be squared error and $\sigma^2 = 1$. The risk functions of the underlying models considered are

(R1) the asymptotic maximum risk over neighborhoods of dependent processes [i.e., the value of (3)] for given $\rho^B$ and $t > 0$;

(R2) the asymptotic maximum risk over neighborhoods of independent processes [i.e., the value of (3) for $\Omega^B = I$ and given $t > 0$];

(R3) the asymptotic risk (equivalently variance) under parametric Gaussian processes with a given boundary correlation vector $\rho^B$ and no neighborhood distributions [i.e., the value of (3) for given $\rho^B$ and $t = 0$]; and

(R4) the asymptotic risk (equivalently variance) under independent parametric Gaussian process with no neighborhood distributions [i.e., the value of (3) for $\Omega^B = I$ and $t = 0$].

Ratios of R1 risk for the estimators E1 and E3 have been calculated for the cases considered in Table 1. In each case the optimal estimator E1 has smaller or equal R1 risk compared to that of E3, but the
Table 2. $\psi$ Functions for the Optimal Huber Estimators E3 in the Contamination Model with Squared Error Loss$^a$

<table>
<thead>
<tr>
<th>Boundary correlation vector</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{AR}(1)p_1^B = 0.0$</td>
<td>0.5</td>
<td>1.636</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.971</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.634</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.1$</td>
<td>0.5</td>
<td>1.636</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.971</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.3$</td>
<td>0.5</td>
<td>1.624</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.937</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.5$</td>
<td>0.5</td>
<td>1.563</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.841</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.785</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.7$</td>
<td>0.5</td>
<td>1.409</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.529</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.496</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.048</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.9$</td>
<td>0.5</td>
<td>1.161</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.535</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>2.374</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>4.150</td>
</tr>
<tr>
<td>$\text{AR}(1)p_1^B = 0.95$</td>
<td>0.5</td>
<td>1.093</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.287</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.636</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>2.141</td>
</tr>
<tr>
<td>$\text{MA}(1)p_1^B = 0.1$</td>
<td>0.5</td>
<td>1.636</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.971</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
<tr>
<td>$\text{MA}(1)p_1^B = 0.3$</td>
<td>0.5</td>
<td>1.624</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.971</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
<tr>
<td>$\text{MA}(1)p_1^B = 0.499.99$</td>
<td>0.5</td>
<td>1.580</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.872</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.953</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.638</td>
</tr>
</tbody>
</table>

$^a$The $\psi$ functions are odd, and on $\mathbb{R}^a$ are of the form

$$\psi(y) = \begin{cases} ay & \text{for } y \in [0, b) \\ b & \text{for } y \in [b, \infty) \end{cases}$$
difference in risks is less than 0.1%. Hence, the class of Huber $\psi$ functions yields approximately optimal $\psi$ functions for the problem defined in (3) that are virtually as good as the strictly optimal $\psi$ functions. As mentioned above, it is much easier to calculate the optimal truncation point $\hat{\gamma}$ for the Huber $\psi$ function than it is to calculate the strictly optimal $\psi$ function using the expansion in terms of orthonormal polynomials. Thus, Huber $\psi$ functions (with truncation points depending on $\rho^b$ and $t$) are recommended for use in the dependent robust models developed above.

Column 1 of Table 3 assesses the performance of the optimal robust estimator for dependent processes $E_1$ in comparison to that of the optimal robust estimator for independent processes $E_2$ on the neighborhoods of independent processes (i.e., using the ratio of their $R_2$ risks). Note that when $R_2$ risk is considered, different $\rho^b$ values in the table do not reflect different underlying neighborhoods of distributions; they reflect different estimators $E_1$. The table shows that the $E_1$ estimators do remarkably well, especially if $\rho^b \leq 0.7$ for AR(1) boundary correlation vectors. (If $\rho^b \leq 0.7$, the ratio of $R_2$ risks is greater than 0.99.) Hence, the specification of a large set of covariance matrices $S$ still yields an estimator that is quite good when the data actually are independent. Virtually identical results hold for the performance of the optimal Huber estimators $E_3$ with respect to the estimator $E_2$ using $R_2$ risk values.

Table 3 also considers the converse situation (see column 2) and assesses the performance of the optimal robust estimator for independent processes $E_2$ with respect to the optimal robust estimator for dependent processes $E_1$ on neighborhoods of independent processes (i.e., using the ratio of their $R_1$ risks) for different boundary correlation vectors $\rho^b$ and different values of $t$. Column 2 of the table gives the comforting result that the optimal robust estimator for independent processes performs extremely well on dependent data, and, hence, is quite robust against positive dependence.

Note that columns 1 and 2 of Table 3 exhibit an asymmetry when very large amounts of dependence are considered, i.e., when $\rho^b \geq 0.9$ for an AR(1) boundary correlation vector. In particular, the optimal robust estimator for independent processes does better (relatively speaking) on dependent distribution neighborhoods than the optimal Huber estimator for dependent processes does on independent distribution neighborhoods. While this asymmetry also exists for $\rho^b < 0.9$, it is not of significant magnitude.

Suppose the data actually are generated by the parametric distribution, i.e., a stationary Gaussian distribution with boundary correlation
vector $\rho^B$. One might ask: How much is lost by using the optimal robust estimator for dependent processes E1 rather than the asymptotically efficient estimator for Gaussian processes E4, i.e., $\bar{y}_n$? Column 3 of Table 3 answers this question by exhibiting the ratio of the R3 risk of E4 to that of E1. For given $\rho^B$, the relative efficiency of E1 increases as $t$ (the size of the neighborhoods) decreases. For given $t$, the relative efficiency of E1 increases as $\rho^B$ increases for MA(1) and AR(1) boundary correlation vectors. This follows because (1) the truncation height of $\psi$ increases as $\rho^B$ increases, and the higher the


<table>
<thead>
<tr>
<th>Boundary correlation vector</th>
<th>$t$</th>
<th>$\text{R2 risk of E}^b_1$</th>
<th>$\text{R2 risk of E}^b_3$</th>
<th>$\text{R1 risk of E}^b_1$</th>
<th>$\text{R1 risk of E}^b_2$</th>
<th>$\text{R3 risk of E}^b_4$</th>
<th>$\text{R3 risk of E}^b_3$</th>
<th>$\text{R4 risk of E}^b_4$</th>
<th>$\text{R4 risk of E}^b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)$\rho^b_1 = 0.9$</td>
<td>0.5</td>
<td>0.921</td>
<td>0.985</td>
<td>0.988</td>
<td>0.959</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.959</td>
<td>0.988</td>
<td>0.966</td>
<td>0.885</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.971</td>
<td>0.995</td>
<td>0.942</td>
<td>0.806</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.993</td>
<td>0.999</td>
<td>0.926</td>
<td>0.735</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1)$\rho^b_1 = 0.95$</td>
<td>0.5</td>
<td>0.858</td>
<td>0.978</td>
<td>0.993</td>
<td>0.979</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.840</td>
<td>0.974</td>
<td>0.979</td>
<td>0.935</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.891</td>
<td>0.984</td>
<td>0.961</td>
<td>0.878</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.933</td>
<td>0.993</td>
<td>0.947</td>
<td>0.824</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Estimators E1 and E3 are so similar that the above table is virtually the same with E1 replaced by E3.

*When R2 or R4 risk is considered, the underlying neighborhoods of distributions are the same for all $\rho^b$ cases considered, only the estimators E1, E2, and E3 vary with $\rho^b$.

*These columns give the same risk ratios as the columns above except that the estimator in the denominator is the optimal Huber estimator E3 rather than the strictly optimal robust estimator E1. This change in the table is made because the estimator E1 has not been calculated for the AR(1), $\rho^b_1 = 0.90$, 0.95 boundary correlation vectors.

Truncation height, the closer is the estimator E1 to $\tilde{y}_n$ and (2) the ratio of R3 risk of $\tilde{y}_n$ to that of an $M$-estimator with given $\psi$ function increases as $\rho^b$ increases. For the case where $t = 1.0$, the loss declines from 19% for an AR(1) least-favorable correlation vector with $\rho^b_1 = 0.1$ to 2% for an AR(1) vector with $\rho^b_1 = 0.95$.

The same question as above can be posed for the case where the data are generated by an independent Gaussian distribution. Column 4 of Table 3 gives the relevant risk ratios, viz. the ratio of R4 risk for the estimator E4 to that of estimator E1. (In this case, for different $\rho^b$ in the table, the underlying distributions used for risk calculations remain unchanged, but the estimators E1 vary with $\rho^b$.) Again, the relative performance of E1 improves as $t$ decreases or as $\rho^b$ increases. And as expected, E1 does better for the case of dependent Gaussian data than for independent Gaussian data.

The converse of the last two cases considered is the assessment of the performance of $\tilde{y}_n$ compared to those of E1 or E2 when the distribution generating the data is arbitrary within the neighborhoods of the dependent or the independent parametric distributions. These are given by the ratios of R1 and R2 risks of the estimators E1 and E4. For any given $\rho^b$, these ratios are zero. That is, $\tilde{y}_n$ performs so poorly for some distributions in the neighborhoods that its R1 and R2 risks are infinite.
V. CONCLUSION

This chapter considers the problem of optimal robust estimation of location in a model with stationary Gaussian parametric distributions. For optimal robustness, a Huber-type minimax criterion is used where minimaxing takes place over neighborhoods of the parametric Gaussian distributions. Four different neighborhood systems are considered.

For the smallest neighborhoods—Hellinger neighborhoods—an estimator is found that is within $\epsilon$ of being asymptotically efficient at the Gaussian parametric model and is within $\epsilon$ of being optimally robust. These results show the extent to which deviations from a Gaussian process can be allowed while still maintaining the optimality of an estimator that essentially is the sample mean.

For neighborhoods of contamination, variational, and Kolmogorov type, the optimal robust estimators have $\psi$ functions very close in shape to Huber $\psi$ functions. In fact, for all practical purposes a Huber $\psi$ function with truncation point adjusted depending on $t$ and $\rho^B$ is as good as the optimal function. Further, for AR(1) and MA(1) boundary correlation vectors with $\rho^B_1 \leq 0.7$, the truncation point is not very sensitive to the value of $\rho^B_1$.

The optimal robust estimators for the dependent model are robust against lack of dependence and the optimal robust estimator for the independent model is robust against positive dependence. When the data are truly Gaussian, the optimal robust estimator for the dependent model is less efficient than $\hat{y}_n$, but decreasingly so as the size ($t$) of the neighborhoods decreases and as the amount of dependence (as measured by $\rho^B$) increases. On the other hand, $\hat{y}_n$ performs disastrously for some distributions in the neighborhoods, and, hence, has infinite maximum risk over the neighborhoods. The specifications of $t$ and $\rho^B$ determine how closely the optimal robust estimator for dependent processes resembles $\hat{y}_n$ and how closely it resembles the sample median $\hat{y}_n$ (which has the minimum possible asymptotic bias for C, V, and K neighborhoods and is optimal for these neighborhoods as $t \to \infty$).

VI. APPENDIX

PROOF OF THEOREM 2: Since Theorem 1 holds $\text{unif}(\theta, F, \Omega)$, the limit as $n \to \infty$ and the supremum over $\Omega \in S$ and $\{F_{\Omega \circ n}\}$ can be exchanged in $R_\psi(R, \theta, t, S)$. Thus, for $G_n = 1/n \sum_{i=1}^n d_{\sigma_i}(\psi, \theta, F_{\Omega_i})$ and
$Z \sim N(0, 1),$

$$R(\psi, d, \theta, t, S) = \sup_{\alpha > 0} \sup_{\Omega \in S} \sup_{\{f_{\psi(d)}\}} E_N \left[ \sigma(\psi, d, \theta, \Omega) \cdot Z + \lim_{n \to \infty} \sqrt{n}G_n \right]$$

$$= \sup_{\Omega \in S} \sup_{\alpha > 0} E_N \left[ \sigma(\psi, d, \theta, \Omega) \cdot Z + \lim_{n \to \infty} \sqrt{n}G_n \right]$$

$$= E_N \left[ \sup_{\Omega \in S} \sigma(\psi, d, \theta, \Omega) \cdot Z + \lim_{n \to \infty} \sqrt{n}G_n \right]. \quad (A.1)$$

The second equality of (A.1) holds because for constants $b, c,$ and $\sigma,$

$$|b + \sigma Z| \leq |c + \sigma Z|$$

if $|b| \leq |c|,$ so if $L$ satisfies assumption L1 we get $L_a(|\cdot|)$ is nondecreasing on $R^+$ and the result of Lehmann (1959, p. 112) implies that $E_N L_a(b + \sigma Z) \leq E_N L_a(c + \sigma Z)$ (also see Bickel, 1981, p. 19). The third equality of Eq. (A.1) holds because the monotone convergence theorem applies with respect to the supremum over $a > 0,$ the second term in parentheses is independent of $\Omega,$ and the function $V(\sigma) \equiv E_N L(\sigma Z + b),$ for “$b$” a constant, is increasing in $\sigma$ for $L$ convex [since $L$ convex implies $V(\sigma)$ is convex and $Z$ symmetrically distributed about zero implies $V(\sigma)$ is an even function].

By Lemma A-1 below, $\sup_{\{f_{\psi(d)}\}} \lim_{n \to \infty} \sqrt{n}G_n = b_i(\psi, d, \theta)$ for $i = H, C, V,$ and $K,$ which yields the desired result.

**Lemma A-1:** If $\{\psi, d\}$ corresponds to some $T \in H_i,$ then

$$\sup_{\{f_{\psi(d)}\}} \lim_{n \to \infty} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d_{aj} \hat{\lambda}(\psi, \theta, F_{in}) \right| = b_i(\psi, d, \theta) \quad \text{for } i = H, C, V, \text{ and } K. \quad (A.2)$$

**Proof of Lemma A-1:** Using $E_{\psi(d)} \psi(\theta, Y) = 0$ and the Cauchy–Schwarz inequality, for all sequences $\{F_{\psi(d)}\}$ in the Hellinger neighborhoods,

$$\lim_{n \to \infty} \sup_{j \leq n} \sqrt{n}A(\psi, \theta, F_{in}) = \lim_{n \to \infty} \sup_{j \leq n} \sqrt{n} \left\{ \int \psi(\theta, y)[f_{in}(y) - \phi(\theta, y)] dy \right\}$$

$$\leq \lim_{n \to \infty} \sup_{j \leq n} \sqrt{n} \left\{ \int \left[ \sqrt{f_{in}(y)} - \sqrt{\phi(\theta, y)} \right]^2 dy \right\}^{1/2}$$

$$\times \left\{ \int \psi(\theta, y)^2 \phi(\theta, y) dy \right\}^{1/2}$$

$$\leq 2t \cdot \left\{ \int \psi(\theta, y)^2 \phi(\theta, y) dy \right\}^{1/2} \quad \text{(A.3)}$$
where the last inequality uses the bounded convergence theorem and the fact that $F_{\infty}^{n} \in \mathcal{F}_{\infty}^{n} \theta$. Thus (for $G_{\infty}$ as defined above),

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} d_{j} \sqrt{n} |\hat{a}(\psi, \theta, F_{\infty})| = b_{k}(\psi, d, \theta). \quad \text{(A.4)}$$

For any function $g(y)$, let $g(y)^+ = g(y) \vee 0 (= \max\{g(y), 0\})$ and $g(y)^- = [-g(y)] \vee 0$. For all sequences $\{F_{\infty}\}$ in the variational neighborhoods,

$$|\hat{a}(\psi, \theta, F_{\infty})| = \left| \int \psi(\theta, y) [f_{0}^{\infty}(y) - \varphi_{0}(y)]^{+} d\mu(y) \right|$$

$$- \int \psi(\theta, y) [f_{0}^{\infty}(y) - \varphi_{0}(y)]^{-} d\mu(y) \right|$$

$$\leq \left[ \text{ess sup}_{y \in R} \psi(\theta, y) - \text{ess inf}_{y \in R} \psi(\theta, y) \right]$$

$$\times \int |f_{0}^{\infty}(y) - \varphi_{0}(y)| d\mu(y)/2$$

$$< (\text{ess sup}_{y \in R} \psi - \text{ess inf}_{y \in R} \psi) \cdot t/\sqrt{n}, \quad \text{(A.5)}$$

using the fact that $\int g^{+} d\mu = \int g^{-} d\mu = 1/2 \int |g| d\mu$ for $g(y) = f_{0}^{\infty}(y) - \varphi_{0}(y)$, and so,

$$\lim_{n \to \infty} \sqrt{n} G_{\infty} \leq \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} |d_{j}| \cdot (\text{ess sup}_{y \in R} \psi - \text{ess inf}_{y \in R} \psi) = b_{k}(\psi, d, \theta). \quad \text{(A.6)}$$

For all sequences $\{F_{\infty}\}$ in the Kolmogorov neighborhoods, assumption A2' and Fubini's Theorem give

$$|\hat{a}(\psi, \theta, F_{\infty})| = \left| \int [\Phi_{\infty}(y) - F_{\infty}(y)] d\nu(y) \right| \leq \|v\|_{V} \cdot t/\sqrt{n}. \quad \text{(A.7)}$$

and so,

$$\lim_{n \to \infty} \sqrt{n} G_{\infty} \leq \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \|v_{j}\|_{V} \cdot t = b_{k}(\psi, d, \theta). \quad \text{(A.8)}$$

For sequences $\{F_{\infty}\}$ in the contamination neighborhoods, each $dF_{\infty}(y)$ can be written as $F_{\infty}(y) = (1 - t/\sqrt{n})\Phi_{\infty}(y) + Q_{\infty}(y)t/\sqrt{n}$, for some $dF_{\infty}$. Thus,

$$\hat{a}(\psi, \theta, F_{\infty}) = \int \psi(\theta, y) dQ_{\infty}(y) \cdot t/\sqrt{n} \leq \text{ess sup}_{y \in R} \psi(\theta, y) \cdot t/\sqrt{n}$$

and

$$\hat{a}(\psi, \theta, F_{\infty}) \geq \text{ess inf}_{y \in R} \psi(\theta, y) \cdot t/\sqrt{n}. \quad \text{(A.9)}$$
Let \( d_{nj}^+ = d_{nj} \vee 0 \) and \( d_{nj}^- = (-d_{nj}) \vee 0 \). Using assumption D2, 
\[
(1/n) \sum_{j=1}^{n} d_{nj}^{+} \rightarrow \text{ess sup } \psi - d_{nj}^- \text{ ess inf } \psi \text{ at } 0
\]
and 
\[
(1/n) \sum_{j=1}^{n} d_{nj}^{-} \rightarrow \text{ess sup } \psi - d_{nj}^+ \text{ ess inf } \psi \text{ at } 0
\]
Thus, we get 
\[
\lim_{n \to \infty} \sqrt{n} G_n \leq \left[ \lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (d_{nj}^{+} \text{ ess sup } \psi - d_{nj}^- \text{ ess inf } \psi) \right]^{1/2}
\]
\[
\vee \left[ \lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (d_{nj}^{-} \text{ ess sup } \psi - d_{nj}^+ \text{ ess inf } \psi) \right]^{1/2}
\]
\[
= t \cdot [(\eta_1 \text{ ess sup } \psi - \eta_2 \text{ ess inf } \psi) \vee (\eta_2 \text{ ess sup } \psi - \eta_1 \text{ ess inf } \psi)]
\]
\[
= b_{\psi}(\psi, d, \theta).
\]
(A.10)

Next, for \( i = H, V, \) and \( K \), consider sequences \( \{F_{\theta, n}^{(i)}\} \) of distributions that are in \( \{F_{\theta, n}^{(i)}(\theta), \Omega_n\} \), for all \( n \) sufficiently large and all \( \Omega \in S \) and whose univariate marginal densities with respect to Lebesgue measure are given by
\[
f_{\theta, n}^{(i)}(y) = \varphi_{\psi}(y) \cdot \exp[\text{sgn}(d_{nj})h(y)/\sqrt{n} - c_n] \quad \text{for } j = 1, 2, \ldots,
\]
(A.11)

where \( c_n \) is a constant defined to make this a density, \( h \in H(\theta) \).

\( H_{\psi}(\theta) = \left\{ h \in L^\infty : \int h(y)\varphi_{\psi}(y)dy = 0 \quad \text{and} \quad \int h^2(y)\varphi_{\psi}(y)dy < 4t^2 \right\} \),
\[
H_{\psi}(\theta) = \left\{ h \in L^\infty : \int h(y)\varphi_{\psi}(y)dy = 0 \quad \text{and} \quad \int |h(y)|\varphi_{\psi}(y) dy < 2t \right\},
\]
and
\[
H_{\psi}(\theta) = \left\{ h \in L^\infty : \int h(y)\varphi_{\psi}(y)dy = 0 \quad \text{and} \quad \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^{x} h(y)\varphi_{\psi}(y) dy \right| < t \right\}.
\]

Such distributions exist since they can be created by point transformations of a sequence of rv's with distribution \( \Phi_{\Omega} \) (see Andrews, 1982). Bickel (1981, Lemma 2) shows that for \( n \) sufficiently large the univariate distributions \( F_{\theta, n}^{(i)} \) are in \( F_{\psi, n}^{(i)}(\theta) \) for all \( h \in H(\theta) \), for \( i = H, V, \) and \( K \). This implies \( F_{\theta, n}^{(i)} \in \{F_{\psi, n}^{(i)}(\theta), \Omega_n\} \) for \( n \) sufficiently large, for all \( h \in H(\theta) \), for \( i = H, V, \) and \( K \). Also, by Bickel (1981, Theorem 2),
\[
\sup_{h \in H(\theta)} \int \psi(\theta, y)h(y)\varphi_{\psi}(y)dy = \begin{cases} 
2t \cdot \left[ \int \psi^2(\theta, y)\varphi_{\psi}(y)dy \right]^{1/2} & \text{for } i = H \\
2t \cdot (\text{ess sup } \psi - \text{ess inf } \psi) & \text{for } i = V \\
2t \cdot \|\psi\|_{\psi} & \text{for } i = K.
\end{cases}
\]
(A.12)
Now, exp[c_n] = 1 + O(1/n) as n → ∞, and so, exp[sgn(d_n) · h(y)/\sqrt{n} - c_n] = 1 + sgn(d_n) · h(y)/\sqrt{n} + O(1/n) as n → ∞, where O(1/n) holds uniformly for y ∈ R, since h is bounded. Hence,

\lim_{n → ∞} (\sqrt{n} \lambda(\psi, \theta, F^{(0)}_{\Phi_0}) - sgn(d_n) · \int \psi(\theta, y)h(y)\varphi_0(y) dy) = 0

and

\sup_{\{F^{(0)}_{\Phi_0}\}} \lim_{n → ∞} |\sqrt{n} G_n| \geq \sup_{\{F^{(0)}_{\Phi_0}\}} \lim_{n → ∞} |\sqrt{n} G_n|

= \sup_{\{F^{(0)}_{\Phi_0}\}} \lim_{n → ∞} \frac{1}{n} \sum_{j=1}^{n} |d_n| \cdot \int \psi(\theta, y)h(y)\varphi_0(y) dy

= b_{\psi}(\psi, d, \theta),

(A.13)

using Eq. (A.12), for i = H, V, and K.

Contamination neighborhoods contain sequences of distributions \{\Phi^{(0)}_{\Phi_0}\}, which are created by point transformations of rv’s with distribution \Phi_0 (see Andrews, 1982), and whose univariate marginal df’s are given by \((1 - t/\sqrt{n})\Phi_0 + Q_j t/\sqrt{n}\) for arbitrary df’s \(Q_j, j = 1, 2, \ldots\). Thus, using Eqs. (A.9) and (A.10),

\sup_{\{F^{(0)}_{\Phi_0}\}} \lim_{n → ∞} |\sqrt{n} G_n| \geq \sup_{\{F^{(0)}_{\Phi_0}\}} \lim_{n → ∞} \frac{1}{n} \sum_{j=1}^{n} d_n \int \psi(\theta, y) dQ_j(y) \cdot t = b_{\psi}(\psi, d, \theta),

(A.14)

Since the inequalities of Eqs. (A.4) and (A.13), (A.6) and (A.13), (A.8) and (A.13), and (A.10) and (A.14) hold in opposite directions, they hold as equalities.

□

**Proof of Theorem 3:** Since \(\eta \geq 1\), \(b_{\psi}(\psi, d, \theta)\) is minimized by taking all \(d_n\) of the same sign. \(\sup_{\Omega \in S} \sigma^2(\psi, d, \theta, \Omega)\) can be written as \(\sup_{\Omega \in S}(Q_1 + Q_2)\), where

\[ Q_2 = \sum_{j=-\infty}^{\infty} \omega_{|j|} E_{\Phi_0} \psi(\theta, Y_j)\psi(\theta, Y_{j+1}) \quad \text{and} \quad Q_1 \text{ is defined analogously with } \omega_{|j|} \text{ replaced by } 1. \]

Using some algebra, assumption D2, a mixing inequality (see Herrndorf, 1984, Lemma 2.1), and the bounded convergence theorem, we get

\[ Q_2 = \lim_{n → \infty} E_{\Phi_0} \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (d_n - d_j) \psi(\theta, Y_j) \right]^2 \geq 0. \]

Since \(Q_2 = 0\) if \(\omega_{|j|} = 0\) for all \(j\), \(\sup_{\Omega \in S} \sigma^2(\psi, d, \theta, \Omega)\) is minimized by taking all \(d_n\) equal. Thus, \(d_n = 1\) for all \(j\) minimizes \(b_{\psi}(\psi, d, \theta)\) and \(\sup_{\Omega \in S} \sigma^2(\psi, d, \theta, \Omega)\) by the same argument as in the proof of Theorem 2, \(R(\psi, d, \theta, t, S)\) as well. □
PROOF OF THEOREM 4: The bivariate normal density \( \varphi_{\theta}(y, x, \rho) \) can be expanded in terms of the Hermite polynomials \( H_k(\cdot), k = 0, 1, \ldots \), as

\[
\varphi_{\theta}(y, x, \rho) = \sum_{k=0}^{\infty} \frac{1}{k!} H_k\left(\frac{y - \theta}{\sigma}\right) H_k\left(\frac{x - \theta}{\sigma}\right) \rho^k \varphi_{\theta}(y) \varphi_{\theta}(x),
\]

where equality is the sense of mean square convergence. For \( j = \pm 1, \pm 2, \ldots \),

\[
E_{4\phi_{\theta}} \psi(\theta, Y_1) \psi(\theta, Y_{1+j}) = \sum_{k=0}^{\infty} \left[ \int_{\mathbb{R}} \psi(\theta, y) H_k\left(\frac{y - \theta}{\sigma}\right) \varphi_{\theta}(y) dy \right]^2 \rho^{j+k}/k!
\leq \sum_{k=0}^{\infty} \left[ \int_{\mathbb{R}} \psi(\theta, y) H_k\left(\frac{y - \theta}{\sigma}\right) \varphi_{\theta}(y) dy \right]^2 \rho^{j+k}/k!
= E_{4\phi_{\theta}} \psi(\theta, Y_1) \psi(\theta, Y_{1+j}),
\]

where the two equalities follow using Eq. (A.15), the boundedness of \( \psi \), and a Fourier series result (e.g., see Andrews, 1982, Lemma 11). Summing over \( j = 0, \pm 1, \pm 2, \ldots \) gives the desired result. \( \square \)

PROOF OF THEOREM 5: Any function \( \psi \) that satisfies A1(a) and A1(b) can be written as \( \psi(\theta, y) = (y - \theta) + g(\theta, y) \), where \( g \) satisfies \( \langle g, 1 \rangle \equiv E_{4\phi_{\theta}} g(\theta, Y) \cdot 1 = 0 \) and \( \langle g, Y - \theta \rangle \equiv E_{4\phi_{\theta}} g(\theta, Y) \cdot (Y - \theta) = 0 \). Without loss of generality we can assume \( g \in L^2[\Phi_{\theta}] \), otherwise \( H(Y - \theta + g, \theta, t, S) = \infty \). We want to show that \( D \equiv \varphi^2(Y - \theta + g, \theta, \Omega^B) - \sigma^2(Y - \theta, \theta, \Omega^B) \geq 0 \) for all \( g \). Using \( \langle g, Y - \theta \rangle = 0 \), we have

\[
D \geq 2 \sum_{j=-\infty}^{\infty} E_{4\phi_{\theta}} (Y_1 - \theta) g(\theta, Y_{1+j})
= 4 \sum_{j=-\infty}^{\infty} \int \int (y - \theta) g(\theta, x) \varphi_{\theta}(y, x, \rho) dy dx.
\]

The bivariate normal density \( \varphi_{\theta}(y, x, \rho^B) \) can be expanded as

\[
\varphi_{\theta}(y, x, \rho^B) = \sum_{i=0}^{\infty} H_{\theta^2}(y) H_{\theta^2}(x) \varphi_{\theta}(y) \varphi_{\theta}(x) (\rho^B)^i/\Gamma(i),
\]

where \( H_{\theta^2}(y) \equiv H_{\theta^2}(y)/\sigma \), and equality is in the sense of mean square convergence. Using Eq. (A.17), this expansion, the fact that \( y - \theta \) and \( g \) are in \( L^2[\Phi_{\theta}] \), and a Fourier series result (e.g., see Andrews, 1982, Lemma 11) we get

\[
D \geq 4 \sum_{j=-\infty}^{\infty} \sum_{i=0}^{\infty} \langle Y - \theta, H_{\theta^2} \rangle \cdot \langle g, H_{\theta^2} \rangle \cdot (\rho^B)^i/\Gamma(i) = 0,
\]
where the equality follows because $y - \theta$ is proportional to $H_{\theta}(y)$, and hence, $y - \theta$ is orthogonal to $H_{\theta}(y), \forall l \neq 1$, and $\langle g, H_{\theta} \rangle = 0$. Thus, $y - \theta$ minimizes $\sigma^2(\cdot, \theta, \Omega^b)$.

For all functions $\psi$ that satisfy A1(b),

$$1 = E_{\theta Y}(\psi(\theta, Y) \cdot (Y - \theta))/\sigma^2 \leq [E_{\theta Y}(\psi^2(\theta, Y))]^{1/2} \cdot [E_{\theta Y}(Y^2 - \theta^2)/\sigma^4]^{1/2} \cdot [E_{\theta Y}(Y - \theta)/\sigma^2]^{1/2} = b_{\theta Y}(\psi, \theta)/(2t\sigma). \quad (A.19)$$

Since $b_{\theta Y}(Y - \theta, \theta) = 2t\sigma$, this implies that $y - \theta$ minimizes $b_{\theta Y}(\psi, \theta)$ subject to A1(b).

Thus, $\xi(\theta, y) = y - \theta$ minimizes $b_{\theta Y}(\psi, \theta)$ and $\sigma^2(\psi, \theta, \Omega^b)$. By the argument used in the proof of Theorem 2, it also minimizes the function $H(\psi, \theta, t, S)$ of these two terms (provided $L$ satisfies assumptions L1 and L2).

**Proof of Lemma 1:** As $\gamma \to \infty$, $\beta_\gamma \to 0$ and $\alpha_\gamma \to 1$. Hence, $r_\gamma(y) \to r(y)$ as $\gamma \to \infty \forall y$, and given $\varepsilon > 0 \forall y$ sufficiently large $|r_\gamma(y)| \leq (1 + \varepsilon)|r(y)| + 1 \equiv \tilde{r}(y)$. So, for $\gamma$ sufficiently large, $|\tilde{r}(y)\tilde{r}(x)|$ is a dominating function of $r_\gamma(y)r_\gamma(x)$ and $\tilde{r}(Y_i)\tilde{r}(Y_{i+1|j}) \in L^2[\Phi_{\Omega}]$, $\forall j$. The dominated convergence theorem then implies that

$$\lim_{\gamma \to \infty} E_{\theta Y}r_\gamma(Y_i)r_\gamma(Y_{i+1|j}) = E_{\theta Y}r(Y_i)r(Y_{i+1|j}), \quad \text{for } j = 0, 1, \ldots$$

(A.20)

For $j = 0$, this gives $\lim_{\gamma \to \infty} b_{\theta Y}(r_\gamma, \theta) = b_{\theta Y}(r, \theta)$.

By changing the order of the infinite sum and the limit and using Eq. (A.20), we get

$$\lim_{\gamma \to \infty} \sum_{j = -\infty}^{\infty} E_{\theta Y}r_\gamma(Y_i)r_\gamma(Y_{i+1|j}) = \sigma^2(r, \theta, t, \Omega^b).$$

The change in order is justified by showing that given $\varepsilon > 0$

$$\left| \sum_{j = -J}^{\infty} E_{\theta Y}r_\gamma(Y_i)r_\gamma(Y_{i+1|j}) \right| < \varepsilon \quad \text{and} \quad \left| \sum_{j = -\infty}^{-J} E_{\theta Y}r_\gamma(Y_i)r_\gamma(Y_{i+1|j}) \right| < \varepsilon,$$

provided $\gamma$ and $J$ are sufficiently large. This can be shown (see Andrews, 1982) using a mixing inequality (see Herrndorf, 1984, Lemma 2.1).

Now, using the assumptions on the loss function $L$, the argument of the proof of Theorem 2 gives Eq. (15).

**Proof of Corollary 1:** It is not hard to show that $\xi_\gamma$ satisfies the Lipschitz assumption A3. Hence, $\xi_\gamma$ and $d^*$ satisfy the assumptions of
Section 2.7. Also,
\[
\lim_{\gamma \to \infty} R_R(\xi, d^*, \theta, t, S) = \lim_{\gamma \to \infty} H(\xi, \theta, t, S) = H(\xi, \theta, t, S)
\]
\[
\leq \inf_{\psi_{A_1} A_2, \text{ and A3 hold}} H(\psi, \theta, t, S) = \inf_{\psi_{A_1} A_2, \text{ and A3 hold}} R_R(\psi, \theta, t, S)
\]
\[
\leq \inf_{(\psi, d)(\psi, d) \text{ corresponds to some } T \in \mathcal{M}_d} R_R(\psi, d, \theta, t, S) \quad (A.21)
\]
by Theorems 2 and 4, Lemma 1, Theorem 5, Theorems 2 and 4, and Theorem 3, respectively.

Proof of Theorem 6: For any function \(\psi(\theta, y)\) that satisfies assumptions A1–A4, consider the odd version of \(\psi(\theta, y)\) defined by
\[
\bar{\psi}(\theta, y) = [\psi(\theta, y) - \psi(-\theta, -y)]/2.
\] (A.22)
\(\bar{\psi}\) satisfies assumptions A1–A4, since \(\phi_0\) is symmetric about \(\theta\). Without loss of generality assume \(\theta = 0\). We have
\[
J(\bar{\psi}) = \sigma^2(\bar{\psi}, 0, \Omega^B) \leq \sigma^2(\psi, 0, \Omega^B) \equiv J(\psi),
\]
since \(J(\psi)\) is a convex function. The latter follows because
\[
J[c \psi_1 + (1 - c) \psi_2] = \lim_{n \to \infty} E_{\mathcal{Q}_{\mathcal{B}}} \left[ c \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_1(Y_j) + (1 - c) \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_2(Y_j) \right]^2
\]
\[
\leq \lim_{n \to \infty} E_{\mathcal{Q}_{\mathcal{B}}} \left\{ c \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_1(Y_j) \right]^2 \right. \\
+ (1 - c) \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_2(Y_j) \right]^2 \right\}
\]
\[
= c J(\psi_1) + (1 - c) J(\psi_2), \quad (A.23)
\]
where the inequality follows because \(g(y) = y^2\) is convex. Also,
\[
b_c(\bar{\psi}, 0) \leq b_c(\psi, 0) \quad \text{by the triangle inequality,}
\]
\[
b_r(\bar{\psi}, 0) \leq t \cdot (\text{ess sup } \psi - \text{ess inf } \psi)/2 - (\text{ess inf } \psi - \text{ess sup } \psi)/2)
\]
\[
= b_r(\psi, 0), \quad (A.24)
\]
and
\[
b_k(\bar{\psi}, 0) = t \cdot \frac{1}{2} \int_K d_k^{\frac{1}{2}} |v_0(y) - v_0(-y)| \leq b_k(\psi, 0)
\]
by the triangle inequality. Thus, by the argument of the proof of Theorem 2, for \(L\) satisfying assumptions L1 and L2, and for \(i = C,\)
V, and K,

\[ E_N L_h(\psi, \theta) + \sigma(\tilde{\psi}, \theta, \Omega^b) \cdot Z \leq E_N L_h(\psi, \theta) + \sigma(\psi, \theta, \Omega^b) \cdot Z. \]

The case for sequences \{\psi_n\} is proved analogously, term by term. \(\square\)

**Proof of Lemma 2:** Given any \(\varepsilon > 0\), for \(M\) sufficiently large,

\[ 2 \cdot \sum_{M+1}^{\infty} E_{\text{mix}} \psi(Y) \psi(Y_{i+1}) \leq 24 C_1(0)^2 \sum_{j=M+1}^{\infty} \alpha(j)^2 < \varepsilon, \quad (A.25) \]

using MIX, assumption A2, and a mixing inequality (see Herrndorf, 1984, Lemma 2.1). This gives the first inequality of Lemma 2. Note that \(\sigma^2_M(\psi, 0, \Omega^b)\) is necessarily positive, since Eq. (A.16) of the proof of Theorem 4 and \(\rho_j^b \geq 0, \forall j\), yield

\[
\sigma^2_M(\psi, 0, \Omega^b) = \int \psi^2(y) \varphi_n(y) dy + 2 \sum_{j=1}^{M} \sum_{k=1}^{\infty} \left[ \int \psi(y) H_k(\frac{y}{\sigma}) \varphi_n(y) dy \right]^2 (\rho_j^b)^2/k! > 0.
\]

(A.26)

The second inequality of Lemma 2 holds because \(E_N L[h_c(\psi, 0) + \sigma \cdot Z]\) is continuous in \(\sigma\). \(\square\)

**Proof of Lemma 3:** The proof is sketched here; details can be found in the proof of Lemma 12 of Andrews (1982). Define

\[ B = \{ \psi \in L^2[\Phi_0]; E_{\Phi_0} \psi(Y) = 0 \quad \text{and} \quad E_{\Phi_0} \psi(Y)/\sigma^2 = 1 \} \]

and

\[ B_\gamma = \{ \psi \in L^2[\Phi_0]; \sup_{y \in R} |\psi(y)| \leq \gamma \}. \]

Consider the weak* topology on the space of \(L^2[\Phi_0]\) functions. In this topology, \(B\) is closed and \(B_\gamma\) is closed and compact. Hence \(B \cap B_\gamma\) is compact. Using Reed and Simon (1981, Theorem 5.7), it can be shown that \(J_M(\psi) \equiv \sigma^2_M(\psi, 0, \Omega^b)\) is lower semicontinuous with respect to the weak* topology on \(B \cap B_\gamma\). This, coupled with the compactness result, implies that \(J_M(\psi)\) attains its infimum on \(B \cap B_\gamma\).

If the function \(\psi^*_\gamma\) which attains the infimum of \(J_M(\psi)\) on \(B \cap B_\gamma\) does not satisfy assumption A3, then take

\[
\psi^{**}_\gamma(y) = \int_R \psi^*_\gamma(y - z) s_\gamma(z) dz, \quad (A.27)
\]

where \(s_\gamma(\cdot)\) is a close (but smooth) approximation of point mass at zero. \(\square\)
PROOF OF LEMMA 4: Since \( \rho_j^B \geq 0 \), \( \forall j \), implies that \( K_M(y, x) < 0 \), \( \forall y \in R^+ \), the right-hand side of Eq. (34), denoted \( W'_\psi(y) \), satisfies
\[
W'_\psi(y) \geq - \frac{\lambda y}{2\sigma^2} + \gamma \int_{R^+} K_M(y, x) dx \geq - \frac{\lambda y}{2\sigma^2} - 2M y > \gamma
\]
for \( y \) sufficiently large, since \( \lambda > 0 \). Hence, for \( y \) sufficiently large, Eq. (34) does not hold and \( \psi_\tau^* \) must be on the upper boundary. The corresponding result for \( y \) sufficiently small follows since \( \psi \) is odd. \( \square \)

PROOF OF THEOREM 7: Substituting Eq. (40) in (34), multiplying (34) by \( \xi_m(y) \phi_\delta(y) \), and integrating (34) over \( D_\tau \), gives
\[
\langle \psi_\tau, \xi_m \rangle = - \frac{\lambda}{2} \left\langle Y, \frac{\xi_m}{\sigma} \right\rangle + \left\langle g_\tau, \frac{\xi_m}{\sigma} \right\rangle
\]
\[
-2 \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \left\langle \xi_k, \frac{\xi_m}{\sigma} \right\rangle \cdot \left\langle \psi_\tau, \frac{\xi_l}{\sigma} \right\rangle \cdot \tau_{kl}
\]
\[
= - \frac{\lambda}{2} \left\langle Y, \frac{\xi_m}{\sigma} \right\rangle + \left\langle g_\tau, \frac{\xi_m}{\sigma} \right\rangle - 2 \sum_{l=1}^{\infty} \left\langle \psi_\tau, \frac{\xi_l}{\sigma} \right\rangle \cdot \tau_{ml},
\]
where the exchange of integral and summation signs necessary for the first equality is justified by a Fourier series result (e.g., see Andrews, 1982, Lemma 11), since \( \xi_\tau, \xi_l \in L^2[\Phi_\delta] \), and the second equality follows by the orthonormality of \( \xi_\tau, l = 1, 2, \ldots \). Equation (A.29) for \( m = 1, 2, \ldots \) can be written as
\[
\psi_\tau = (I + 2Y)^{-1}(-\lambda Y/2 + g_\tau).
\]
The vector of Fourier coefficients \( \psi_\tau \) gives the solution to (34):
\[
\psi_\tau(y) = \sum_{l=1}^{\infty} \left\langle \psi_\tau, \frac{\xi_l}{\sigma} \right\rangle \xi_l(y) = \psi_\tau \xi(y) = (-\lambda Y/2 + g_\tau)(I + 2Y)^{-1} \xi(y),
\]
where the first equality is in the sense of mean square convergence.

The constraint \( E_{\Phi_\delta} \psi_\tau(Y) Y/\sigma^2 = 1 \) is satisfied when
\[
1 = 2 \int_{D_\tau} \left( -\frac{\lambda}{2} Y + g_\tau \right) (I + 2Y)^{-1} \xi(y) \frac{Y}{\sigma^2} \phi_\delta(y) dy
\]
\[
+ \gamma \int_{D_\tau^+} \phi_\delta(y) dy - \gamma \int_{D_\tau^+} \phi_\delta(y) dy
\]
\[
= (-\lambda Y + 2g_\tau)(I + 2Y)^{-1} Y + \gamma[\Phi_\delta(D_\tau^+) - \Phi_\delta(D_\tau^-)].
\]
This yields \( \lambda \) as defined in the theorem. \( \square \)
NOTES

1. I am greatly indebted to Peter J. Bickel and Thomas J. Rotherberg for their help and encouragement.


3. See Atkinson (1976) for a brief description of this method and references to the literature concerning it. Theoretical error bounds can be calculated for this method but often are too large for practical use, and, hence, are not reported here.

4. Note that once the $\psi$ function has been determined it is relatively easy to compute an estimate given actual data, whether one uses the optimal $\psi$ function or the Huber $\psi$ function.

5. This assertion is proved using the argument of Gastwirth and Rubin (1975, Proposition 4.2, Lemma 4.2, and Corollary 4.1).

6. This is a general result that can be proved using the argument of Gastwirth and Rubin (1975, Lemma 4.1 and Theorem 4.1).

REFERENCES


Robust Estimation of Location


