The Annals of Statistics
1994, Vol. 22, No. 2, 1045–1080

ROBUST RECURSIVE ESTIMATION IN THE PRESENCE OF HEAVY-TAILED OBSERVATION NOISE[1]

BY IRVIN C. SCHICK AND SANJOY K. MITTER

Massachusetts Institute of Technology

Under the usual assumptions of normality, the recursive estimator known as the Kalman filter gives excellent results and has found an extremely broad field of application—not only for estimating the state of a stochastic dynamic system, but also for estimating model parameters as well as detecting abrupt changes in the states or the parameters. It is well known, however, that significantly nonnormal noise, and particularly the presence of outliers, severely degrades the performance of the Kalman filter. This results in poor state estimates, nonwhite residuals and invalid inference.

A first-order approximation is derived for the conditional prior distribution of the state of a discrete-time stochastic linear dynamic system in the presence of ε-contaminated normal observation noise. This distribution is then used to derive a first-order approximation of the conditional mean (minimum-variance) estimator. If the observation noise distribution can be represented as a member of the ε-contaminated normal neighborhood, then the conditional prior is also, to first order, an analogous perturbation from a normal distribution whose first two moments are given by the Kalman filter. Moreover, the perturbation is itself of a special form, combining distributions whose parameters are given by banks of parallel Kalman filters and optimal smoothers.

1. Introduction. Time-dependent data are often modeled by linear dynamic systems. Such representations assume that the data contain a deterministic component which may be described by a difference or differential equation. Deviations from this component are assumed to be random and to have certain known distributional properties. These models may be used to estimate the "true" values of the data uncorrupted by measurement error, and possibly also to draw inference on the source generating the data.

A method that has found an exceptionally broad range of applications—not only for estimating the state of a dynamic system, but also for simultaneously estimating model parameters, choosing among several competing models and detecting abrupt changes in the states, the parameters or the form of the model—is the recursive estimator known as the Kalman filter [Kalman (1960), Kalman and Bucy (1961)]. While it has so far enjoyed greater popularity within

Received May 1991; revised May 1993.

[1] Research supported in part by U.S. Army Research Office fellowship ARO-DAAL03-86-G-0017 and by the U.S. Air Force Office of Scientific Research under contracts AFOSR-85-0227 and AFOSR-89-0276.

AMS 1991 *subject classifications.* Primary 62F35, 62M20; secondary 62C20, 93E11.

Key words and phrases. Robust recursive estimation, robust Kalman filtering, outliers in time series, robust conditional mean estimation, asymptotic expansion, ε-contaminated normal distribution.

the engineering community than among statisticians, this versatile technique deserves more attention. Originally derived via orthogonal projections as a generalization of the Wiener filter to nonstationary processes, the Kalman filter has been shown to be optimal in a variety of settings [e.g., Jazwinski (1970), pages 200–218]. It has been derived as the weighted least-squares solution to a regression problem, without regard to distributional assumptions [e.g., Duncan and Horn (1972), Bryson and Ho (1975), pages 349–364]; as the Bayes estimator assuming Gaussian noise, without regard to the cost functional [e.g., Harrison and Stevens (1971), Meinhold and Singpurwalla (1983)]; and as the solution to various game theoretic and other problems. Indeed, Morris (1976) is led to conclude that the Kalman filter is therefore "a robust estimator" and proceeds to demonstrate its minimax optimality "against a wide class of driving noise, measurement noise, and initial state distributions for a linear system model and the expected squared-error cost function."

One condition under which the Kalman filter is not robust is heavy-tailed noise, that is, the presence of outliers: even rare occurrences of unusually large observations severely degrade the performance of the Kalman filter, resulting in poor state estimates, nonwhite residuals and invalid inference. There is no contradiction between this fact and the findings of Morris and others. It is well known that the squared-error criterion is extremely sensitive to outliers [Tukey (1960), Huber (1964)], for reasons that are intuitively easy to grasp. Squaring a large number makes it even larger, so that an outlier entering the cost functional linearly is likely to dominate all other observations. In other words, optimality relative to the "linear system model and the expected squared-error cost function" must *not* be sought when the noise distribution is heavy-tailed.

Statisticians and engineers often confront the problem of dealing with outliers in the course of model building and validation. Routinely ignoring unusual observations is neither wise nor statistically sound, since such observations may contain valuable information as to unmodeled system characteristics, model degradation or breakdown, measurement errors, and so forth. However, detecting unusual observations is only possible by comparison with the underlying trends and behavior; yet it is precisely these that nonrobust methods fail to capture when outliers are present. The purpose of robust estimators is thus twofold: to be as nearly optimal as possible when there are no outliers, that is, under "nominal" conditions; and to be resistent to outliers when they do occur, that is, to be able to extract the underlying system behavior without being unduly affected by spurious values.

Past efforts to mitigate the effects of outliers on the Kalman filter range from ad hoc practices, such as simply discarding observations for which residuals are "too large," to more formal approaches based on nonparametric statistics, Bayesian methods or minimax theory. Many, however, include heuristic approximations with ill-understood characteristics. While some of these techniques have been empirically found to work well, their theoretical justifications have remained scanty at best. Their nonlinear forms, coupled with the difficulties inherent in dealing with nonnormal distributions, have resulted in a strong preference in the literature for Monte Carlo simulations over analytical rigor.

In this paper, a robust recursive estimator is derived formally, in an effort to bridge the gap between appealing heuristics and sound theory. An asymptotic expansion is used to derive a nonlinear filter that approximates the conditional mean estimator. The resulting estimator has good performance characteristics both under nominal conditions and in the presence of outliers. Since its distributional properties are known (approximately), it is also possible to use this estimator for statistical inference, such as failure detection and identification.

The paper is organized as follows. The problem is formally stated in Section 2, and a survey of the literature is offered in Section 3. In Section 4, a first-order approximation of the conditional prior distribution of the state given past observations is derived. This distribution is used to derive a first-order approximation of the conditional mean estimator of the state given past and present observations, in Section 5. Minimax issues and the choice of noise distribution are addressed in Section 6, followed in Section 7 by some simulation results, and in Section 8 by a brief summary.

2. Problem statement. Let $(\mathbb{R}^d, \mathcal{B}, \lambda)$ be a measure space, where $\mathbb{R}$ denotes the real line, $\mathcal{B}$ the Borel σ-algebra and λ the Lebesgue measure. Below, the notation $\mathcal{L}(\mathbf{x})$ denotes the probability law of the random vector $\mathbf{x}$ taking values in $\mathbb{R}^d$; $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance Σ; and $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ denotes its Radon–Nikodym derivative with respect to the Lebesgue measure. Finally, the notation $\mathbf{p}_{\mathbf{x}}(\mathbf{X})$ denotes the probability distribution function of the random variable $\mathbf{x} \in \mathbb{R}^d$ evaluated at $\mathbf{X}$, although the subscript will be dropped wherever there is no ambiguity.

Consider the model

$$\mathbf{z}_n = H_n \theta_n + D_n \mathbf{v}_n, \tag{2.1}$$

where

$$\theta_{n+1} = F_n \theta_n + \mathbf{w}_n, \tag{2.2}$$

$n = 0, 1, \ldots$ denotes discrete time; $\theta_n \in \mathbb{R}^q$ is the system state, with a random initial value distributed as $\mathcal{L}(\theta_0) = \mathcal{N}(\overline{\theta}_0, \Sigma_0)$; $\mathbf{z}_n \in \mathbb{R}^p$ is the observation or measurement; $\mathbf{w}_n \in \mathbb{R}^q$ is a random variable (the process or plant noise) distributed as $\mathcal{L}(\mathbf{w}_n) = \mathcal{N}(\mathbf{0}, Q_n)$; $\mathbf{v}_n \in \mathbb{R}^p$ is a random variable (the observation or measurement noise) distributed as $\mathcal{L}(\mathbf{v}_n) = \mathcal{F}$, a given distribution that is absolutely continuous with respect to the Lebesgue measure with $\mathbb{E}[\mathbf{v}_n] = \mathbf{0}$ and $\mathbb{E}[\mathbf{v}_n \mathbf{v}_n^{\mathrm{T}}] = R$; $\{F_n\}$, $\{H_n\}$, $\{D_n\}$, $\{Q_n\}$, Σ_0 and R are known matrices or sequences of matrices with appropriate dimensions; $\overline{\theta}_0 \in \mathbb{R}^q$ is a known vector; and finally θ_0, $\mathbf{w}_n$ and $\mathbf{v}_n$ are mutually independent for all n.

A well-known estimator $\widehat{\theta}_n$ of the state θ_n given the observations $\mathcal{Z}_n = \{\mathbf{z}_0, \ldots, \mathbf{z}_n\}$ is the Kalman filter, given by the recursion

$$\widehat{\theta}_n = \overline{\theta}_n + K_n \gamma_n, \tag{2.3}$$

where

$$\overline{\theta}_n = F_{n-1} \widehat{\theta}_{n-1} \tag{2.4}$$

is the predicted (a priori) estimate of the state at time n (i.e., before updating by the observation $\mathbf{z}_n$) and

$$M_n = F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} + Q_{n-1} \tag{2.5}$$

is the prediction error covariance at time n;

$$\gamma_n = \mathbf{z}_n - H_n \overline{\theta}_n \tag{2.6}$$

is the innovation at time n and

$$\Gamma_n = H_n M_n H_n^{\mathrm{T}} + D_n R D_n^{\mathrm{T}} \tag{2.7}$$

is its covariance;

$$K_n = M_n H_n^{\mathrm{T}} \Gamma_n^{-1} \tag{2.8}$$

is the gain; and

$$P_n = M_n - K_n \Gamma_n K_n^{\mathrm{T}} \tag{2.9}$$

is the a posteriori estimation error covariance at time n (i.e., after updating). The initial condition $\overline{\theta}_0$ is given.

As is clear from equations (2.3) and (2.6), the estimate is a linear function of the observation, a characteristic that is optimal only in the case of normally distributed noise [Goel and DeGroot (1980)] or elliptical processes (samplepathwise mixtures of normal processes). Similarly, equations (2.5) and (2.8)–(2.9) show that the gain and covariance are independent of the data, a property related once again to the assumption of normality. Finally, in the Gaussian case $\mathcal{F} = \mathcal{N}(\mathbf{0}, R)$, the residual (innovation) sequence $\{\gamma_1, \dots, \gamma_n\}$ is white and is distributed as $\mathcal{L}(\gamma_i) = \mathcal{N}(\mathbf{0}, \Gamma_i)$.

When $\mathcal{F}$ is a heavy-tailed distribution, on the other hand, the state estimation error can grow without bound (since the estimate is a linear function of the observation noise), the residual sequence becomes colored and residuals become nonnormal. Thus, not only is the estimate poor, but furthermore invalid inference would result from utilizing the residual sequence when significant excursions from normality occur. A robust estimator should at the very least have the following characteristics: the state estimation error must remain bounded as a single observation outlier grows arbitrarily; the effect of a single observation outlier must not be spread out over time by the filter dynamics, that is, a single outlier in the observation noise sequence must result in a single outlier in the residual sequence; and the residual sequence should remain nearly white when the observation noise is normally distributed except for an occasional outlier.

If $\mathcal{F}$ is unknown but can be expressed as a member of a class of distributions, it makes sense to seek the optimal estimator $\widehat{\theta}_n$ of θ_n given $\mathcal{Z}_n$ in a *minimax* sense [Huber (1964)]. Huber shows for the static case $\theta_n = \theta$ that under fairly mild conditions, the minimax optimal estimator is in fact the maximum likelihood estimator for the least favorable member of the class, that is, for the distribution with minimum Fisher information.

In choosing a class containing $\mathcal{F}$, a convenient model of indeterminacy similar to that of Huber (1964) is the *ε-contaminated normal neighborhood*

$$\mathcal{P}_{\varepsilon,R} = \{(1-\varepsilon)\mathcal{N}(0,R) + \varepsilon H : H \in \mathcal{S}\}, \tag{2.10}$$

where $\mathcal{S}$ is the set of all suitably regular probability distributions, and $0 \le \varepsilon \ll 1$ is the known fraction of "contamination." It is assumed in the sequel that $\mathcal{F} \in \mathcal{P}_{\varepsilon,R}$ for some appropriately chosen ε and R. The form of the observation noise distribution is exploited in an *asymptotic expansion*, in order to obtain a first-order approximation of the conditional prior distribution $\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1})$ of the state variable θ_n given the observations $\mathcal{Z}_{n-1}$. A key property that ensures the finite dimensionality of this approximation is the *exponential stability* of the Kalman filter, that is, the fact that the effects of past observations decay fast enough. The resulting distribution is a perturbation from the normal, and all the pertinent parameters are given by various Kalman filters and optimal smoothers that each make a specific assumption on the distribution of the noise at each point in time. The relationship between Huber's estimator of a location parameter, its recursive versions proposed by Martin (1972), Martin and Masreliez (1975), Nevel'son (1975) and Price and Vandelinde (1979), and the estimator derived here, is discussed in greater detail in Mitter and Schick (1992).

It is assumed that the observation noise is white, that is, that outliers occur independently. While this assumption may be seen as limiting [other models have been proposed, e.g., by Martin and Yohai (1986)], the principal goal of this effort is to derive a recursive estimator that can be used for inference on the linear dynamic model in the presence of heavy-tailed noise: if outliers were allowed to occur in "patches," the distinction between model changes and sequences of outliers would become arbitrary, or might have to be based upon prior probabilities for patch duration. This is not to say that patchy outliers do not constitute a problem worthy of study—time series outliers can occur in patches, and an approach to that case based upon time-scaling is currently under study.

It is also assumed that outliers only occur in the observation noise: process noise outliers (also known as "innovational" outliers, as opposed to observation or "additive" outliers) would cause abrupt state changes that would not immediately be distinguishable from failures (except by observation of the subsequent behavior of the model, i.e., noncausally). Nevertheless, dealing with process noise outliers in real time is a problem for which satisfactory solutions remain unavailable.

The first-order approximation of the conditional prior distribution $\mathbf{p}(\theta_n | \mathcal{Z}_{n-1})$ is next used to obtain a first-order approximation of the conditional mean of the state variable θ_n given the observations $\mathcal{Z}_n$ (i.e., to update the predicted estimate by the current observation $\mathbf{z}_n$). This step uses a generalization of a proof due to Masreliez (1975) and Masreliez and Martin (1977), made possible by a change in the order of integration. A similar derivation also yields the conditional covariance.

3. Literature survey. Engineers have long had recourse to ad hoc methods aimed at downweighting the influence of outliers on the Kalman filter. The simplest way employed is just to discard observations for which the residual is "too large" [e.g., Meyr and Spies (1984)]. Thus, the predicted estimate $\overline{\boldsymbol{\theta}}_n$ of the state $\boldsymbol{\theta}_n$ would not be updated by $\mathbf{z}_n$ if, for example,

$$\left|[\gamma_n]_i\right| > \alpha\sqrt{[\Gamma_n]_{ii}}, \tag{3.1}$$

for some i (where $[\cdot]_i$ and $[\cdot]_{ij}$ denote elements of a vector and a matrix, respectively), or if

$$\gamma_n^{\mathrm{T}}\Gamma_n^{-1}\gamma_n > \beta, \tag{3.2}$$

for some positive thresholds α and β. This is equivalent to rewriting the Kalman filter in equation (2.3) as

$$\widehat{\boldsymbol{\theta}}_n = \overline{\boldsymbol{\theta}}_n + K_n\psi_n(\gamma_n), \tag{3.3}$$

where ψ_n is an *influence-bounding function* that is linear between some possibly time-dependent (e.g., as a function of the covariance) thresholds, and zero elsewhere. There are several disadvantages to this approach, notably the absence of a firm theoretical basis or justification, as well as the lack of a rigorous way to choose the thresholds. (Three standard deviations is sometimes used, but more for historical reasons than due to statistical considerations.) Moreover, no use is made of information contained in the observations if they fall outside the thresholds, which may in some cases result in decreased efficiency: if something is known about the statistical properties of the outliers, then it might be possible to extract some information from outlying observations as well, and discarding them outright may not be appropriate. Finally, sharply redescending influence-bounding functions of this type lead to a lack of continuity in the estimates as functions of the data, giving rise to nonrobust covariances [see Huber (1981), page 103].

Somewhat more sophisticated approaches have also been advanced to preprocess the data prior to its use in updating the Kalman filter estimate. Thus, for instance, Kirlin and Moghaddamjoo (1986) use the median, while Hewer, Martin and Zeh (1987) use Huber's M-estimator. Both papers report on applications to real data (target tracking in the former, glint noise in the latter), where outliers were found to affect adversely the performance of the Kalman filter.

In recent years, a great deal of work has been published, investigating more formal techniques for "robustifying" recursive estimators. Broadly speaking, these methods can be grouped in three categories:

1. *Bayesian methods*—When the noise is non-Gaussian, but its statistical properties are known and not excessively complex, estimators can be derived in a Bayesian framework, whereby observations are used to update modeled prior information. The parameters of these estimators are often chosen in accordance with some performance criterion, such as the risk.

2. *Nonparametric methods*—There are cases of practical importance where the statistical properties of the noise are either entirely unknown or known only partially, or possibly known but very complex. In such cases, distribution-free estimators are sometimes sought that remain valid in a relatively broad class of situations.
3. *Minimax methods*—Another way of dealing with incomplete or absent knowledge of the statistical properties of the noise is to choose a class of distributions and derive the estimator whose worst-case performance is optimal. If a saddle-point property can be shown to hold, such estimators are referred to as minimax robust.

A review of the literature follows. It is worth noting that the recent literature on robust statistics is vast, and a broad survey is not attempted here. Indeed, even indirectly related works, such as those on robust regression or outlier detection, are not discussed, except when they specifically focus on the robust estimation of the state of a dynamic system. Published reviews include Ershov (1978b), Stockinger and Dutter (1983, 1987), Kassam and Poor (1985) and Martin and Raftery (1987).

McGarty (1975) proposes a method to maximize the Bayes risk, eliminating outliers and concurrently computing the estimate. His model assumes that the state is totally absent from the observation when an outlier occurs, that is, that observations are occasionally pure noise and contain no information at all. That differs from the model assumed here, where the state is always observed, although the noise may occasionally contain outliers. Moreover, McGarty's method is nonrecursive, as well as computationally burdensome.

A Bayesian setting is also employed by Sorenson and Alspach (1971), Alspach (1974) and Agee and Dunn (1980), who use a Gaussian-sums approximation for the prior distributions. There is some similarity between this approach and the derivation of the conditional prior in this paper. However, while the number of components in the approximating sum grows exponentially with time in these papers, the formulation adopted here (which exploits the exponential asymptotic stability of the Kalman filter) results in a bounded number of terms. Although the option of truncating the mixture sums to reduce complexity has been raised in the literature, little is known about the consequences of such a move in the general case. Tanaka and Katayama (1987) use maximum a posteriori (MAP) estimation to determine of which component of the sum the noise was a realization. Their method is noncausal, but that is because they assume both the process and the observation noise to be distributed according to Gaussian sums. They also make the questionable assumption that the conditional distribution of the state (given all past and present observations) is normal. Peña and Guttman (1989) propose to replace the posterior mixture by a single normal distribution whose mean and variance are equal to those of the mixture, and they show that this "collapsing by moments" procedure minimizes the Kullback–Leibler distance between the two distributions. While this method is insensitive to outliers, the resulting loss of efficiency under nominal conditions (normal noise) is unclear.

Meinhold and Singpurwalla (1989) also use mixture of distributions, with Student-t rather than normal components. This assumption yields the rather elegant property that the posterior reduces to the prior at the limit as an observation tends to infinity; for finite observations, however, Student-t prior and noise distributions would not result in a Student-t posterior, necessitating some ad hoc manipulations, both to ensure that the posterior distribution can be represented as a mixture of Student-t distributions and to limit the number of components in the mixture. Furthermore, the results only hold for the scalar case.

A simple way to decrease the influence of outliers is to adjust the noise covariance matrix used in the filter to reflect the greater variance due to them. Suppose, for instance, that outliers occur with probability ε and that the covariances of the nominal (underlying) and outlier models are denoted by R_{nom} and R_{out}, respectively. Then, using the inflated covariance

$$R = (1 - \varepsilon)R_{\text{nom}} + \varepsilon R_{\text{out}} \tag{3.4}$$

in the Kalman filter recursion results in the deflation of the gain K_n and hence a reduction in the influence of outliers. Unfortunately, of course, this results in a reduction of the influence of all other observations as well, with the consequence that very inefficient use is made of measurement information when no outliers are present.

Guttman and Peña (1984, 1985) and Peña and Guttman (1988) propose a more refined version of (3.4): they assume a distributional model for the observation noise and compute a posterior observation noise convariance by using the posterior probability that an outlier has occurred, conditioned on the measurement. Similar approaches are discussed by Harrison and Stevens (1971, 1976) and Kliokis (1987), as well as by Athans, Whiting and Gruber (1977), who assume that measurements are occasionally independent of the state, that is, pure noise. Athans, Whiting and Gruber also offer a comparison between their Bayesian estimator and a simple outlier-rejection scheme based on a χ^2 test. One problem with this method is the need for an explicit model for the noise: Guttman and Peña use a two-component Gaussian mixture (scale contamination) model, which is somewhat limiting—although frequently used in the literature. Another problem is that inflated covariances and poor performance at the nominal model may result when the "outlier" distribution contains significant mass in the "center," as is the case with the Gaussian mixture.

A related method is proposed by Ershov and Lipster (1978) and Ershov (1978a), whose framework is very similar to that of Guttman and Peña, but who make a *hard decision* at each step as to whether or not the observation is an outlier. This approach has the distinct advantage of superior performance at the nominal model, since the effective covariance is either R_{nom}, or R_{out} but not a weighted combination of the two. Furthermore, although the published derivation is for the scalar case, the multivariate extension is straightforward. The difficulty with this formulation is that the problem of choosing an outlier model remains: Ershov and Lipster only consider the Gaussian mixture case. In addition, it is probable that such hard decisions result in nonrobust covariances, in view of the fact that small deviations in the neighborhood of thresholds can yield large differences in the value of the estimate. Indeed, abrupt switching of covariances introduces transients in the filter dynamics which have apparently not been the object of study.

It is worth noting that both the Guttman–Peña and the Ershov–Lipster filters can also be formulated in the form of equation (3.3)—the first with a smooth and the latter with a piecewise linear ψ-function. Neither function is bounded, implying that the performance of these estimators is poor when the observation noise is very heavy-tailed.

Mixture models are also used by West, Harrison and Migon (1985) in the context of generalized linear models for nonlinear time series in the presence of outliers. Their discussion is brief, however, and their proposal rather sketchy.

A Bayesian framework is also used by Kitagawa (1987), who proposes to approximate non-Gaussian distributions by piecewise linear functions and to select the best among a set of competing models by means of the Akaike information criterion (AIC). This method is computation-intensive. Furthermore, there is little theoretical justification for using AIC in this context, although different considerations, such as minimax optimality, could be used for choosing among the competing models.

Another attempt at representing a distribution by simpler functions is that of Tsai and Kurz (1983), where a piecewise polynomial approximation is used in adaptively deriving the influence-bounding function. Some connections between this approach and AIC are discussed in Tsai and Kurz (1982). While adaptive methods are very appealing when modeling information is incomplete, this particular application raises a problem: since outliers are rare occurrences by definition, large samples are likely to be required for even moderate levels of confidence, particularly in the tails. Furthermore, the derivation presented in the paper is for the scalar case only (or, more precisely, for the case where the elements of each observation vector are uncorrelated), and the multivariate extension is quite arbitrary; yet, such correlation could provide crucial information in the event of an outlier that affects some measurements more than others.

The need to select probabilistic models for the noise is entirely circumvented by the use of nonparametric, distribution-free estimators such as the median [Nevel'son (1975), Evans, Kersten and Kurz (1976), Guilbo (1979), Gebski and McNeil (1984)]. Medians and other quantiles have very useful properties, such as strong resistance to transients (like outliers) but perfect tracking of abrupt changes (like step inputs or slope changes). Furthermore, the development of recursive methods for estimating them has eliminated the computational burden and memory requirements commonly associated with such statistics. However, their performance remains ill-understood, as do their statistical properties.

A final class of robust filters is based on a minimax approach. Here, a class or neighborhood of situations (e.g., noise distributions) is selected, and the estimator with the best performance under the least favorable member of that class is sought—where best and worst are defined in a certain sense. This paradigm is very appealing, since, in view of the absence of precise knowledge of the noise distribution, the essence of robust estimation is a quest for methods that perform satisfactorily under a relatively broad range of conditions. Since the least favorable situation may in fact not represent reality, and estimators could conceivably be found that perform better under some other conditions, this approach is necessarily conservative. However, it has the important advantage of providing a lower bound on the performance of the estimator.

One group of papers [VandeLinde, Doraiswami and Yurtseven (1972), Doraiswami (1976), Yurtseven and Sinha (1978), Yurtseven (1979)] assumes bounds on covariances and obtains a minimax estimator under various conditions. Unfortunately, these papers are opaque and not always consistent with each other, making their complicated methods somewhat inaccessible. Moreover, their nonrecursive nature makes them unsuitable for the present problem.

The literature most pertinent to this paper [Masreliez (1974, 1975), Masreliez and Martin (1974, 1977), Tollet (1976), Stanković and Kovačević (1979, 1986), West (1981), Stepiński (1982), Kovačević and Stanković (1986, 1988)] uses *stochastic approximation* of the Robbins–Monro type to get a recursive approximate conditional mean (minimum variance) estimator having the form of (3.3), with the influence-bounding function ψ_n given by the score of the conditional distribution of the observation $\mathbf{z}_n$, that is,

$$\psi_n(\mathbf{z}) = -\nabla_{\mathbf{z}} \log \mathbf{p}_{\mathbf{z}_n}(\mathbf{z} \mid Z_{n-1}) \tag{3.5}$$

$$= -\frac{\nabla_{\mathbf{z}} \mathbf{p}_{\mathbf{z}_n}(\mathbf{z} \mid Z_{n-1})}{\mathbf{p}_{\mathbf{z}_n}(\mathbf{z} \mid Z_{n-1})}. \tag{3.6}$$

This estimator has been found to perform well in simulation studies [e.g., Martin and DeBow (1976)], as well as with real data [e.g., Çetin and Tekalp (1990)], but its theoretical basis has remained inadequate. Moreover, a crucial assumption, that of a normal conditional prior for the state at each time step, is insufficiently justified and remains controversial. [For a continuity theorem regarding the near-normality of the conditional prior, see Martin (1979).] Finally, the one-step estimator is converted into a recursion in an ad hoc manner that contradicts the assumption of conditional normality.

Similar filters are investigated by Agee and Turner (1979) and Agee, Turner and Gomez (1979), who eliminate the explicit relationship between the influence function and distributional assumptions in the interest of versatility. As a result, however, these filters are not minimax and the choice of influence-bounding function remains arbitrary. Mataušek and Stanković (1980) also study related filters for the case of nonlinear, continuous-time, discretely sampled systems; their discussion of influence-bounding functions does not appear to be statistically motivated either. Shirazi, Sannomiya and Nishikawa (1988) consider models where both the process and the observation noises contain outliers; they, too, make the questionable assumption of Gaussian conditional prior and only offer simulation results to support their algorithm. Levin (1980) investigates methods for analyzing the accuracy of filters of the form (3.3) with bounded ψ-functions, including notably the minimax robust estimators described above.

Tsaknakis and Papantoni-Kazakos (1988) start out from a rather different definition of robustness, based on the Prokhorov distance and on what they call "asymptotic outlier resistance," and construct a minimax robust estimator that is insensitive to bursty outliers of fixed duration. While the scalar estimator is minimax, however, its multivariate generalization is ad hoc and does not obviously share this property.

Boncelet and Dickinson (1983) describe a minimax filter obtained by applying M-estimation techniques to the Kalman filter reformulated as a regression problem. However, the results are incomplete, and the crucial problem of updating the covariance is not addressed; further results do not appear to have been published as of this writing. Cipra and Romera (1991) similarly reformulate the one-step update of the state as a least-squares estimation problem and apply M-estimation techniques to it. Some approximations allow them to obtain recursions for both the state and its covariance.

4. The conditional prior distribution.

Before deriving a robust estimator of the state θ_n given the observations $\mathcal{Z}_n$, it is necessary to define the sense in which optimality will be sought. The often-used linear-estimator least-squares criterion is not robust in the presence of outliers, as mentioned earlier, while Huber's asymptotic variance (or, alternatively, the Fisher information) criterion is not meaningful in the time-varying case of equation (2.2).

The *conditional mean estimator* is well known to have several desirable properties, such as unbiasedness and minimum error variance [see e.g., Anderson and Moore (1979), pages 26–28], and is chosen to be the optimality criterion here. The first derivation of a robust approximate conditional mean estimator in the present context is due to Masreliez and Martin (1974, 1977) and is based on Masreliez (1974, 1975); some generalizations are provided by West (1981).

A key assumption made by these and other authors is that at each n the conditional probability distribution of the state θ_n given past observations $\mathcal{Z}_{n-1}$ is normal. This assumption allows some elegant algebraic manipulations that yield a *stochastic approximation*-like estimator. However, while the assumption of conditional normality has been shown in simulation studies to be a good approximation of the true density, it is only strictly correct for finite n in the special case where $\mathcal{F} = \mathcal{N}(\mathbf{0}, R)$ [see Spall and Wall (1984)], which is clearly of no interest here.

In this section, a first-order approximation of the conditional distribution prior to updating, $\mathbf{p}(\theta_n|\mathcal{Z}_{n-1})$, is derived for the case where $\mathcal{F}$ is known and belongs to the ε-contaminated normal family $\mathcal{P}_{\varepsilon,R}$ defined in equation (2.10). While conditional normality is never exactly satisfied in the presence of nonnormal noise, it is shown that the zeroeth-order term in a Taylor series representation of the distribution is normal. The small parameter around which the Taylor series is constructed involves ε, the fraction of "contamination," as well as a measure of the dynamic stability of the model. This approximation is then used, in an extension of Masreliez's theorem, to derive a first-order approximation of a robust conditional mean estimator. It is initially assumed that the "contaminating" distribution H is known. The choice of H in practice, a problem whose solution remains incomplete, is further discussed in Section 6.

It is first noted that the Kalman filter recursions are exponentially asymptotically stable under certain conditions. This property ensures that the effects of past outliers are attenuated rapidly enough as new observations become available. The stability of the Kalman filter recursions has been studied by several researchers, notably Deyst and Price (1968), Caines and Mayne (1970), Jazwinski [(1970), pages 234–243], Hager and Horowitz (1976) and Moore and Anderson (1980). Hager and Horowitz (1976) relax the conditions of *controllability* and *observability*, used below, in certain cases. See also Anderson and Moore (1981) and Anderson (1982).

The following stability theorem is stated without proof.

THEOREM 4.1. *Let the matrix sequences* $\{F_n\}, \{H_n\}, \{Q_n\}$ *and* $\{D_n\}$ *be bounded above, and let* $\{D_n\}$ *also be bounded below. Let there exist positive integers t and s and positive real numbers* α *and* β *such that, for all* n,

$$\sum_{i=n}^{n+t} \left(\prod_{j=n}^{i-1} F_j \right)^{\mathrm{T}} H_i^{\mathrm{T}} (D_i R D_i^{\mathrm{T}})^{-1} H_i \left(\prod_{j=n}^{i-1} F_j \right) > \alpha I \tag{4.1}$$

(i.e., the system is completely observable) and

$$\sum_{i=n-s}^{n} \left(\prod_{j=i+1}^{n} F_j \right) Q_i \left(\prod_{j=i+1}^{n} F_j \right)^{\mathrm{T}} > \beta I \tag{4.2}$$

(i.e., the system is completely controllable).

Then, given any $\tilde{\theta}_0 < \infty$ *and defining the closed-loop recursion*

$$\tilde{\theta}_{n+1} = (I - K_{n+1} H_{n+1}) F_n \tilde{\theta}_n \tag{4.3}$$

[where K_n *is the Kalman gain defined in equation* (2.8)], *there exist* $\lambda > 0$ *and* $0 < \delta < 1$ *such that*

$$\|\tilde{\theta}_n\| < \lambda \delta^n \tag{4.4}$$

(i.e., the filter is exponentially asymptotically stable).

PROOF. See Moore and Anderson (1980). This result is used in the following, slightly different form.

COROLLARY 4.1. *Let the conditions of Theorem* 4.1 *be satisfied, and let a* $0 < \phi < \infty$ *exist such that, for all* n,

$$\left\| \prod_{j=1}^{n} F_j \right\| < \phi \tag{4.5}$$

(i.e., the system is uniformly stable). For $i = 1, 2$, *let* $\bar{\theta}_n^i$ *and* M_n^i, *respectively, denote the estimators and error covariances of two Kalman filters tracking a*

dynamic system of the form (2.1)–(2.2), *with respective initial conditions* $\overline{\theta}_0^i$ *and* M_0^i. *Then there is a* $0 < \delta < 1$ *such that, for any finite* θ,

$$\mathcal{N}(\theta; \overline{\theta}_n^1, M_n^1) = \mathcal{N}(\theta; \overline{\theta}_n^2, M_n^2) + O(\delta^n). \tag{4.6}$$

PROOF. Theorem 4.1 leads in straightforward fashion to

$$\|M_n^1 - M_n^2\| = O(\delta^{2n}) \tag{4.7}$$

and

$$\|\overline{\theta}_n^1 - \overline{\theta}_n^2\| = O_p(\delta^n). \tag{4.8}$$

Now, $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ is everywhere continuously differentiable with respect to μ and Σ except at $\Sigma = 0$. However, it can be shown [see Moore and Anderson (1980)] that M_n^i is bounded away from 0 for all n, so that it is possible to write a first-order Taylor series expansion of $\mathcal{N}(\theta; \overline{\theta}_n^1, M_n^1)$ around the point $(\overline{\theta}_n^2, M_n^2)$. Using (4.7) and (4.8) concludes the proof. [For a detailed proof, see Schick (1989), pages 122–124.] □

Define

$$\Xi(\mathbf{x}; \mu, \Sigma) = \int \mathcal{N}(\mathbf{x} - \xi; \mu, \Sigma)\, dH(\xi). \tag{4.9}$$

Note that this convolution integral yields the distribution of the sum of two random variables, of which one is normal and the other obeys the "contaminating" distribution H. Note also that, from (2.10), one can write

$$\mathbf{v}_n = (1 - \eta_n)\mathbf{v}_n^N + \eta_n \mathbf{v}_n^H, \tag{4.10}$$

where η_n is a random variable independent of θ_0 and $\{\mathbf{w}_n\}$ obeying

$$\eta_n = \begin{cases} 0, & \text{w.p. } (1 - \varepsilon), \\ 1, & \text{w.p. } \varepsilon, \end{cases} \tag{4.11}$$

and $\{\mathbf{v}_n^N\}$ and $\{\mathbf{v}_n^H\}$ are random variables independent of $\{\eta_n\}$, θ_0 and $\{\mathbf{w}_n\}$ with $\mathcal{L}(\mathbf{v}_n^N) = \mathcal{N}(\mathbf{0}, R)$, for some $R > 0$, and $\mathcal{L}(\mathbf{v}_n^H) = H$. Finally, loosely defining a random variable distributed as H as an "outlier," denote the event "there has been no outlier among the first n observations" by $\mathcal{H}_n = \{\eta_0 = 0, \dots, \eta_n = 0\}$, and the event "there has been exactly one outlier among the first n observations, at time $i - 1$" by $\mathcal{H}_n^i = \{\eta_0 = 0, \dots, \eta_{i-2} = 0, \eta_{i-1} = 1, \eta_i = 0, \dots, \eta_n = 0\}$. Then it is easy to verify that

$$\begin{aligned} &\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1})\mathbf{p}(\mathcal{Z}_{n-1}) \\ &\quad = \mathbf{p}(\mathcal{H}_{n-1})\mathbf{p}(\mathcal{Z}_{n-1} \mid \mathcal{H}_{n-1})\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1}, \mathcal{H}_{n-1}) \\ &\qquad + \sum_{i=1}^{n} \mathbf{p}(\mathcal{H}_{n-1}^i)\mathbf{p}(\mathcal{Z}_{n-1} \mid \mathcal{H}_{n-1}^i)\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1}, \mathcal{H}_{n-1}^i) \\ &\qquad + \text{higher-order terms.} \end{aligned} \tag{4.12}$$

Clearly, the first term on the right-hand side of (4.12) is the distribution conditioned on the event that there were no outliers; each term in the summation to the event that there was exactly one outlier; and the higher-order terms to the occurrence of two or more outliers. Moreover, defining $\mathcal{Z}_n^i = \{\mathbf{z}_0, \dots, \mathbf{z}_{i-2}, \mathbf{z}_i, \dots, \mathbf{z}_n\}$, it follows that

$$\begin{aligned} &\mathbf{p}\big(\mathcal{Z}_{n-1} \mid \mathcal{H}_{n-1}^i\big)\, \mathbf{p}\big(\theta_n \mid \mathcal{Z}_{n-1}, \mathcal{H}_{n-1}^i\big) \\ &\quad = \mathbf{p}\big(\mathcal{Z}_{n-1}^i \mid \mathcal{H}_{n-1}^i\big)\, \mathbf{p}\big(\theta_n \mid \mathcal{Z}_{n-1}^i, \mathcal{H}_{n-1}^i\big)\, \mathbf{p}\big(\underline{z}_{i-1} \mid \theta_n, \mathcal{Z}_{n-1}^i, \mathcal{H}_{n-1}^i\big). \end{aligned} \tag{4.13}$$

Note that the *only* nonnormal term on the right-hand side of (4.13) is the last one. It is shown in the sequel that this term corresponds to a convolution of the form (4.9). Furthermore, since the distribution of a past event is expressed here conditioned on subsequent observations, this corresponds to a *smoother*. The second term on the right-hand side of (4.13), on the other hand, is the distribution of a normal random variable (the state θ_n) conditioned on normally distributed observations $\mathcal{Z}_{n-1}^i$. It is therefore a normal distribution, whose mean and variance are given by a Kalman filter that skips the observation $\mathbf{z}_{i-1}$. These remarks are formalized below.

A first-order approximation of the conditional probability distribution $\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1})$ is given by the following theorem.

THEOREM 4.2. *Let the conditions of Theorem* 4.1 *and Corollary* 4.1 *be satisfied for the system given by equations* (2.1) *and* (2.2), *and let* δ *be a real number for which* (4.4) *holds. Let* ω *be the smallest integer such that*

$$\delta^\omega \le \varepsilon \tag{4.14}$$

(*or, alternatively*, $\omega \ge \log\varepsilon / \log\delta$). *If*

$$\omega\varepsilon < 1 \tag{4.15}$$

and if the distribution H has s, then

$$\begin{aligned} \mathbf{p}\big(\theta_n \mid \mathcal{Z}_{n-1}\big) = {} & (1-\varepsilon)^m \kappa_n \kappa_n^0 \mathcal{N}\big(\theta_n; \overline{\theta}_n^0, M_n^0\big) \\ & + \varepsilon(1-\varepsilon)^{m-1} \kappa_n \sum_{i=\ell}^{n} \kappa_n^i \mathcal{N}\big(\theta_n; \overline{\theta}_n^i, M_n^i\big) \Xi\big(H_{i-1} V_n^i \theta_n; \zeta_n^i, Z_n^i\big) \\ & + O_p\big(m^2\varepsilon^2\big), \end{aligned} \tag{4.16}$$

with $m = \min(n, \omega)$ *and* $\ell = \max(1, n - \omega + 1)$, *where, for* $i = 0, 1, \dots$ *and* $n > i$,

$$\overline{\theta}_n^i = F_{n-1}\theta_{n-1}^i, \tag{4.17}$$

$$\theta_n^i = \overline{\theta}_n^i + K_n^i \gamma_n^i, \tag{4.18}$$

$$M_n^i = F_{n-1} P_{n-1}^i F_{n-1}^{\mathrm{T}} + Q_{n-1}, \tag{4.19}$$

$$\gamma_n^i = \mathbf{z}_n - H_n \overline{\theta}_n^i, \tag{4.20}$$

$$\Gamma_n^i = H_n M_n^i H_n^{\mathrm{T}} + D_n R D_n^{\mathrm{T}}, \tag{4.21}$$

$$K_n^i = M_n^i H_n^{\mathrm{T}} (\Gamma_n^i)^{-1}, \tag{4.22}$$

$$P_n^i = M_n^i - K_n^i \Gamma_n^i K_n^{i\,\mathrm{T}} \tag{4.23}$$

and

$$\kappa_n^i = \kappa_{n-1}^i \mathcal{N}(\gamma_{n-1}^i; \mathbf{0}, \Gamma_{n-1}^i), \tag{4.24}$$

and, for $i = 1, 2, \ldots$ and $n > i$,

$$V_n^i = V_{n-1}^i P_{n-1}^i F_{n-1}^{\mathrm{T}} (M_n^i)^{-1}, \tag{4.25}$$

$$\nu_n^i = \nu_{n-1}^i + V_{n-1}^i K_{n-1}^i \gamma_{n-1}^i, \tag{4.26}$$

$$W_n^i = W_{n-1}^i - V_{n-1}^i K_{n-1}^i \Gamma_{n-1}^i K_{n-1}^{i\,\mathrm{T}} V_{n-1}^{i\,\mathrm{T}}, \tag{4.27}$$

$$\tau_n^i = \mathbf{z}_{i-1} - H_{i-1} \nu_n^i, \tag{4.28}$$

$$\zeta_n^i = H_{i-1} V_n^i \overline{\theta}_n^i + \tau_n^i, \tag{4.29}$$

$$Z_n^i = H_{i-1} \big(W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \big) H_{i-1}^{\mathrm{T}}, \tag{4.30}$$

subject to the initial conditions

$$\overline{\theta}_i^i = F_{i-1} \overline{\theta}_{i-1}^0 \tag{4.31}$$

$$M_i^i = F_{i-1} M_{i-1}^0 F_{i-1}^{\mathrm{T}} + Q_{i-1}, \tag{4.32}$$

$$V_i^i = M_{i-1}^0 F_{i-1}^{\mathrm{T}} (M_n^i)^{-1}, \tag{4.33}$$

$$\nu_i^i = \overline{\theta}_{i-1}^0, \tag{4.34}$$

$$W_i^i = M_{i-1}^0, \tag{4.35}$$

$$\kappa_i^i = \kappa_{i-1}^0, \tag{4.36}$$

for $i > 0$, and

$$\theta_0^0 = \overline{\theta}_0, \tag{4.37}$$

$$M_0^0 = M_0, \tag{4.38}$$

$$\kappa_0^0 = 1. \tag{4.39}$$

The normalization constant satisfies

$$\kappa_n^{-1} = (1-\varepsilon)^m \kappa_n^0 + \varepsilon (1-\varepsilon)^{m-1} \sum_{i=\ell}^{n} \kappa_n^i \Xi\big(\tau_n^i; \mathbf{0}, H_{i-1} W_n^i H_{i-1}^{\mathrm{T}} \big). \tag{4.40}$$

PROOF. See Appendix A.

COMMENTS. Some comments on Theorem 4.2 are as follows:

(i) Equations (4.17)–(4.23) are a bank of Kalman filters, each starting at a different point in time $i = 0, 1, \ldots$. Because of the way in which they are initialized, the cases $i > 0$ correspond to Kalman filters that start at time $n = 0$ and skip the $(i-1)$st observation. In other words, $\overline{\theta}_n^i$ is the estimate of θ_n based on the observations $\mathcal{Z}_{n-1}^i$ and the event $\mathcal{H}_{n-1}^i$. The case $i = 0$ is based on all observations, that is, $\overline{\theta}_n^0$ estimates θ_n based on the observations $\mathcal{Z}_{n-1}$ and the event $\mathcal{H}_{n-1}$.

(ii) Equations (4.25)–(4.27) are a bank of optimal fixed-point smoothers [see, e.g., Anderson and Moore (1979), pages 170–175], each estimating the state at a different point in time $i-1$ based on all preceding and subsequent observations. In other words, ν_n^i is the estimate of θ_{i-1} based on the observations $\mathcal{Z}_{n-1}^i$ and the event $\mathcal{H}_{n-1}^i$.

(iii) Thus, each term in the summation on the right-hand side of (4.16) is a Kalman filter that skips one observation, coupled with an optimal smoother that estimates the state at the time the observation is skipped. Some general results pertaining to conditional probability distributions of the form (4.16) are given in Di Masi, Runggaldier and Barazzi (1983).

(iv) Evidently, as $n \to \infty$, the probability of the event that only a finite number of outliers occur vanishes for any $\varepsilon > 0$. That the density can nevertheless be approximated by the first-order expression in (4.16) is due to the exponential asymptotic stability of the Kalman filter: ω represents a "window size" beyond which the effects of older observations have sufficiently attenuated. Compare Martin and Yohai [(1986), Theorem 4.2] and its discussion in Künsch (1986), where weak dependence on temporally distant observations is exploited in the context of influence curves for time series.

(v) Sample values of the "window size" ω appearing in Table 1 can give an idea of the dimensionality involved. These examples are for the scalar time-invariant case, with $H_n = D_n = Q_n = R = 1$ for all n. A δ is computed for each value of $F_n = F$ (by fitting a straight line to $\log \widetilde{\theta}_n$ for large n), and an ω is calculated for each pair of values F and ε [using (4.14)]. As the table indicates, for a given ε, a smaller F (i.e., faster system dynamics) implies that a smaller "window" ω is enough to guarantee sufficient attenuation. Conversely, for a given F, a smaller ε implies that a longer "window" ω is needed.

(vi) It is easy to show that

$$(4.41)\qquad (1-\varepsilon)^n \kappa_n \kappa_n^0 = \frac{\mathbf{p}(\mathcal{H}_{n-1})\mathbf{p}(\mathcal{Z}_{n-1} \mid \mathcal{H}_{n-1})}{\mathbf{p}(\mathcal{Z}_{n-1})}$$

$$(4.42)\qquad = \mathbf{p}(\mathcal{H}_{n-1} \mid \mathcal{Z}_{n-1})$$

is the posterior probability, conditioned on all past observations $\mathcal{Z}_{n-1}$, that no outliers have occurred among the first n observations. Similarly, it is easy to show that

$$(4.43)\qquad \varepsilon(1-\varepsilon)^{n-1}\kappa_n\kappa_n^i \Xi(\tau_n^i; \mathbf{0}, H_{i-1}W_n^i H_{i-1}^{\mathrm{T}}) = \mathbf{p}(\mathcal{H}_{n-1}^i \mid \mathcal{Z}_{n-1})$$

is the posterior probability that exactly one outlier occurred, at time $i-1$. Thus, equation (4.16) is a sum of conditional distributions similar to (4.12).

TABLE 1
Sample values for ω

F	δ	ε					
		0.001	**0.005**	**0.01**	**0.05**	**0.1**	**0.5**
1.00	0.382	8	6	5	4	3	1
0.90	0.362	7	6	5	3	3	1
0.50	0.234	5	4	4	3	2	1
0.10	0.050	3	2	2	1	1	1
0.05	0.025	2	2	2	1	1	1
0.01	0.005	2	1	1	1	1	1

5. The conditional mean estimator. The approximate conditional prior probability distribution of the state θ_n given the observations $\mathcal{Z}_{n-1}$ is now used in an extension of a theorem due to Masreliez. This results in a first-order approximation of the conditional mean (i.e., minimum-variance) estimator.

The following notation is used, respectively, for the a posteriori conditional mean and conditional variance of θ_n :

$$\mathbf{T}_n = \mathbb{E}\left[\theta_n \mid \mathcal{Z}_n\right]; \tag{5.1}$$

$$\Sigma_n = \mathbb{E}\left[(\theta_n - \mathbf{T}_n)(\theta_n - \mathbf{T}_n)^{\mathrm{T}} \mid \mathcal{Z}_n\right]. \tag{5.2}$$

In addition, the functional

$$\psi_n^0(\mathbf{z}) = -\frac{1}{\mathbf{p}_{\mathbf{z}_n},\left(\mathbf{z} \mid \mathcal{Z}_{n-1}, \mathcal{H}_{n-1}\right)} \nabla_{\mathbf{z}}\, \mathbf{p}_{\mathbf{z}_n}\left(\mathbf{z} \mid \mathcal{Z}_{n-1}, \mathcal{H}_{n-1}\right) \tag{5.3}$$

denotes the score function for the conditional probability of $\mathbf{z}_n$ given that no outliers occurred during the first $n-1$ observations. [Compare with equation (3.6).] Similarly, for $i = 1, 2, \ldots$ and all $n \geq i$,

$$\psi_n^i(\mathbf{z}) = -\frac{1}{\mathbf{p}_{\mathbf{z}_{i-1}}(\mathbf{z} \mid \mathcal{Z}_n^i, \mathcal{H}_n^i)} \nabla_{\mathbf{z}} \mathbf{p}_{\mathbf{z}_{i-1}}(\mathbf{z} \mid \mathcal{Z}_n^i, \mathcal{H}_n^i) \tag{5.4}$$

denotes the score function for the conditional probability of $\mathbf{z}_{i-1}$ given that it was an outlier and that no outliers occurred among the remaining $n-1$ observations. For $i = 0, 1, 2, \ldots$ and all $n \geq i$,

$$\Psi_n^i(\mathbf{z}) = \nabla_{\mathbf{z}} \psi_n^{i^{\mathrm{T}}}(\mathbf{z}) \tag{5.5}$$

denotes the Jacobian of ψ_n^i. Finally, let h denote the Radon-Nikodym derivative of the "contaminating" distribution H with respect to the Lebesgue measure, provided that it exists.

THEOREM 5.1. *Let the conditions of Theorem* 4.1, *Corollary* 4.1 *and Theorem* 4.2 *be satisfied for the system given by equations* (2.1) *and* (2.2). *If h exists and* *is bounded and differentiable a.e., then*

$$\mathbf{T}_n = (1-\varepsilon)^m \kappa_{n+1} \pi_n^0 \mathbf{T}_n^0 + \varepsilon (1-\varepsilon)^{m-1} \kappa_{n+1} \sum_{i=\ell}^{n} \pi_n^i \mathbf{T}_n^i + O_{\mathbf{p}}(m^2 \varepsilon^2) \tag{5.6}$$

with $m = \min(n, \omega)$ *and* $\ell = \max(1, n - \omega + 1)$, *where, for* $i = 0, 1, \ldots,$ *and* $n > i$,

$$\mathbf{T}_n^0 = \overline{\theta}_n^0 + M_n^0 H_n^{\mathrm{T}} \psi_n^0(\gamma_n^0), \tag{5.7}$$

$$\mathbf{T}_n^i = \theta_n^i + P_n^i V_n^{i^{\mathrm{T}}} H_{i-1}^{\mathrm{T}} \psi_n^i(\tau_{n+1}^i), \tag{5.8}$$

$$\pi_n^0 = (1-\varepsilon)\kappa_{n+1}^0 + \varepsilon \kappa_n^0 \Xi(\gamma_n^0; \mathbf{0}, H_n M_n^0 H_n^{\mathrm{T}}), \tag{5.9}$$

$$\pi_n^i = (1-\varepsilon)\kappa_{n+1}^i \Xi(\tau_{n+1}^i; 0, H_{i-1} W_{n+1}^i H_{i-1}^{\mathrm{T}}), \tag{5.10}$$

and the score (or influence-bounding) functions are given by

$$\psi_n^0(\gamma) = \frac{(1-\varepsilon)\mathcal{N}(\gamma; \mathbf{0}, \Gamma_n^0)(\Gamma_n^0)^{-1}\gamma - \varepsilon \nabla_\gamma \Xi(\gamma; \mathbf{0}, H_n M_n^0 H_n^{\mathrm{T}})}{(1-\varepsilon)\mathcal{N}(\gamma; \mathbf{0}, \Gamma_n^0) + \varepsilon \Xi(\gamma; \mathbf{0}, H_n M_n^0 H_n^{\mathrm{T}})}, \tag{5.11}$$

$$\psi_n^i(\tau) = \frac{\nabla_\tau \Xi(\tau; \mathbf{0}, H_{i-1} W_{n+1}^i H_{i-1})}{\Xi(\tau; \mathbf{0}, H_{i-1} W_{n+1}^i H_{i-1})}, \tag{5.12}$$

with $\overline{\theta}_n^i$, θ_n^i, γ_n^i, τ_n^i, M_n^i, P_n^i, Γ_n^i, V_n^i, W_n^i, κ_n^i *and* κ_n *as defined in equations* (4.17)–(4.28) *and* (4.40), *subject to the initial conditions* (4.31)–(4.39). *Furthermore,*

$$\Sigma_n = (1-\varepsilon)^m \kappa_{n+1} \pi_n^0 \Sigma_n^0 + \varepsilon (1-\varepsilon)^{m-1} \kappa_{n+1} \sum_{i=\ell}^{n} \pi_n^i \Sigma_n^i + O_{\mathbf{p}}(m^2 \varepsilon^2), \tag{5.13}$$

where, for $i = 0, 1, \ldots$ *and* $n > i$,

$$\Sigma_n^0 = M_n^0 - M_n^0 H_n^{\mathrm{T}} \Psi_n^0(\gamma_n^0) H_n M_n^0 + (\mathbf{T}_n - \mathbf{T}_n^0)(\mathbf{T}_n - \mathbf{T}_n^0)^{\mathrm{T}}, \tag{5.14}$$

$$\Sigma_n^i = P_n^i - P_n^i V_n^{i^{\mathrm{T}}} H_{i-1}^{\mathrm{T}} \Psi_n^i(\tau_{n+1}^i) H_{i-1} V_n^i P_n^i + (\mathbf{T}_n - \mathbf{T}_n^i)(\mathbf{T}_n - \mathbf{T}_n^i)^{\mathrm{T}} \tag{5.15}$$

and Ψ_n^i *is given by equation* (5.5), *subject to* (5.11) *and* (5.12).

Proof. See Appendix B.

Comments. Some comments on Theorem 5.1 are as follows:

(i) Both Theorem 4.2 and Theorem 5.1 are based on the assumption that outliers occur rarely relative to the dynamics of the filter. In the unlikely event that two outliers occur within less than ω time steps of each other, equation (5.8)—which shows that $\mathbf{T}_n^i$ is linear in θ_n^i and therefore [by (4.17) and (4.18)] in $\mathbf{z}_n$—suggests that the estimate would be strongly affected. This implies that the estimator developed here is robust in the presence of rare and isolated outliers, but not when outliers occur in batches. This important limitation is further discussed in Section 8.

(ii) It is easy to see that

$$(1-\varepsilon)^n \kappa_{n+1} \pi_n^0 = \mathbf{p}(\mathcal{H}_{n-1} \mid \mathcal{Z}_n) \tag{5.16}$$

and

$$\varepsilon(1-\varepsilon)^{n-1} \kappa_{n+1} \pi_n^i = \mathbf{p}(\mathcal{H}_n^i \mid \mathcal{Z}_n), \tag{5.17}$$

that is, the estimator is a weighted sum of *stochastic approximation*-like estimators, with weights equal to the posterior probabilities of each outlier configuration. These probabilities are conditioned on all the observations, including the current one.

(iii) Unlike the Kalman filter, the estimation error covariance Σ_n is a function of the observations. Indeed, the Gaussian case is the only one where the error covariance is independent of the observations. Note, however, that the covariance is a function of a set of matrices $\{M_n^i\}$, $\{P_n^i\}$, $\{\Gamma_n^i\}$, $\{V_n^i\}$ and $\{W_n^i\}$, which are themselves independent of the observations. Thus, they can be precomputed and stored, as is sometimes done with the Kalman filter. This would drastically reduce the on-line computational burden.

(iv) The estimate of Theorem 5.1, as well as its error covariance, are both fairly complex. In all but the simplest cases, obtaining them will be computation-intensive. However, the structure given in Theorems 4.2 and 5.1 includes banks of parallel filters and smoothers that are entirely independent of each other. This suggests that the estimator derived here is well suited to parallel computation.

(v) The error covariance Σ_n includes a weighted sum of quadratic terms of the form $(\mathbf{T}_n - \mathbf{T}_n^i)(\mathbf{T}_n - \mathbf{T}_n^i)^{\mathrm{T}}$. In some sense, this sum measures the disagreement among the parallel estimators, weighted by the posterior probabilities of each outlier configuration, and can be regarded as a price paid for analytical redundancy.

(vi) The "robust Kalman filter" of Masreliez and Martin (1974, 1977) is approximately equivalent to the zeroeth-order term in equation (5.6), that is, to $\mathbf{T}_n^0$ as given in (5.7). This may explain its good empirical performance, as reported in the literature, despite the questionable assumption of normal conditional prior on which it is based. It is also instructive to compare $\mathbf{T}_n^i$ with the robust smoother of Martin (1979).

(vii) It is easy to verify that, for $\varepsilon = 0$,

$$\psi_n^0(\gamma_n^0) = (\Gamma_n^0)^{-1}\gamma_n^0, \tag{5.18}$$

so that $\mathbf{T}_n$ reduces to the Kalman filter in the Gaussian case.

6. The noise distribution. The significance of the functional ψ lies in the fact that it processes the innovation so as to mitigate the effects of observation outliers. "Overprocessing" the data results in loss of efficiency at the nominal model, while "underprocessing" makes the estimator excessively sensitive to outliers, that is, nonrobust.

In the case of Huber's M-estimator of location [Huber (1964, 1969, 1972, 1977), (1981), Chapter 4] and its recursive versions [Martin (1972), Martin and Masreliez (1975), Schick (1989), Chapter 3], the goal is to estimate a deterministic parameter—either a time-invariant location parameter or one that changes in a known and deterministic fashion—given observations corrupted by heavy-tailed noise. Since the parameter itself is deterministic, asymptotic performance measures are used, following the lead of Huber. Estimators are designed to minimize the asymptotic estimation error covariance under the least favorable noise distribution, and these are shown to be saddle-points, that is, optimal in the minimax sense.

In this paper, however, the goal is to estimate the state of a stochastic time-variant linear dynamic system. In other words, the parameter to be estimated is itself randomly changing, and the problem consists of optimally tracking it, rather than achieving minimum asymptotic estimation error. Thus, approximations of a conditional mean estimator are sought, since such estimators are known to achieve minimum error variance at each point in time. In Sections 4 and 5, the "contaminating" noise distribution H is treated as known. In other words, the results of Sections 4 and 5 are better characterized as *non-Gaussian* (or, more generally, *Bayesian*) filters than as *robust* ones. To achieve minimax robustness in this case as well, it is necessary to choose a least favorable distribution H and show that the solution satisfies a saddle-point property. [Of related interest is Berger and Berliner (1983, 1986), who investigate Bayes robustness in the presence of ε-contaminated noise, although not in a minimax framework.]

It is clear from equations (5.13)–(5.15) that the estimation error variance Σ_n depends crucially on the distributions of the innovation and residual terms. The relationship between these distributions and Σ_n is complicated, as is fairly evident from these equations, but there is an additional factor that makes this problem especially difficult: the innovation and residual terms are clearly sums of normally distributed random variables and random variables distributed according to a member of the ε-contaminated normal neighborhood of distributions. The main difference between Huber's formulation and this one is thus that the former involves the neighborhood $\mathcal{P}_{\varepsilon,R}$ defined in equation (2.10), whereas the corresponding neighborhood in the latter case is

$$\mathcal{P}_{\varepsilon,R_1,R_2} = \left\{(1-\varepsilon)\mathcal{N}(\mathbf{0},R_1) + \varepsilon\big(\mathcal{N}(\mathbf{0},R_2) \otimes H\big):\ H \in \mathcal{S}\right\} \tag{6.1}$$

where R_1 and R_2 are given positive-definite matrices, and $\otimes$ denotes the convolution operator. To appreciate the distinction, note that when $R_1 = R_2 = R$, Huber's case involves *replacing* outliers, and (6.1) *additive* ones.

The problem of minimizing the Fisher information for the location parameter of neighborhoods of the form (6.1) was first posed by Mallows (1978), who postulated that the minimizing H concentrates its mass on a set of isolated points, and that it has a geometric form; Donoho (1978) proposes a slight variant, also of a basically geometric form, and offers some numerical results supporting his choice. Marazzi (1985) also presents numerical results and proposes some approximations to the form of the least favorable distribution. This issue has been widely discussed in the literature, particularly in a Bayesian setting where either the prior or the noise distribution is normal and the other distribution is sought to maximize the expected risk. Since it has been shown [see Brown (1971)] that the minimum Bayes risk is a linear function of the Fisher information, the problems are equivalent. This connection was used in the present context by Bickel (1981, 1983), Levit (1979, 1980) and Marazzi (1980, 1985, 1990).

Mallows (1980) states without reference that B. F. Logan demonstrated that the least favorable H cannot have a continuous density, but that "after much effort I have been unable to determine" the distribution in question. Casella and Strawderman (1981) show that if the least favorable distribution is constrained to place all its mass within some interval $[-m, m]$ then, for small m, it concentrates on the endpoints. Bickel (1981) investigates this case for large m and derives a cosine-shaped density that is a second-order approximation of the least favorable one. Bickel and Collins (1983) prove under certain regularity conditions that the least favorable density concentrates its mass on a countable subset of isolated points, possibly including $\{\pm\infty\}$. Marazzi (1980) also provides a proof that the least favorable distribution is discrete. None of these authors, however, is able to derive exactly the distribution minimizing the Fisher information in this case.

Another difficulty in deriving a least favorable distribution for the present problem is due to its multivariate nature. The usual ordering of matrices (given $X, Y \in \mathbb{R}^{d\times d}$, $Y > X$ if and only if $Y - X > 0$, i.e., their difference is positive definite) is not a lattice ordering. Thus, finding the member of a class of distributions that maximizes the error covariance is not generally possible in the multivariate case. In the special case of *spherically symmetric* distributions, the multivariate extension is of course trivial: if the least favorable distributions and influence-bounding functions can be found coordinatewise, everything else follows immediately.

Huber touches on the multivariate case only very briefly [Huber (1972), (1977), page 35; (1981), pages 211 and 222–223]. He proposes to consider spherically symmetric distributions and to apply nondegenerate affine transformations to obtain parametric families of "elliptic" distributions. This, however, brings forth the problem of determining the scaling parameter when, as is usually the case, it is not known a priori. Huber addresses the issue of simultaneous location and scale estimation in the scalar case and also offers some methods for estimating the scaling parameter [see Huber (1981), pages 215–223]. In the present case, the scaling matrix is simply the covariance of the innovation and residual terms and can be found analytically. Thus, if the observation noise distribution $\mathcal{F}$ is spherically symmetric, the multivariate extension is straightforward. However, the difficulty with finding the least favorable distribution componentwise remains.

It is clear from the literature discussed above that the least favorable distribution in the neighborhood $\mathcal{P}_{\varepsilon,R_1,R_2}$ is of a highly complex shape and extremely difficult to derive. Moreover, even if such a least favorable distribution were found, it is not clear a priori that the resulting estimator could be shown to satisfy a saddle-point condition. Since the very choice of neighborhood is to a large extent arbitrary, all this effort is perhaps unwarranted in the present case.

An approximation [see also Marazzi (1990)] consists of the following: since $\mathcal{P}_{\varepsilon,R_1,R_2} \subset \mathcal{P}_{\varepsilon,R_1}$, the least favorable distribution in $\mathcal{P}_{\varepsilon,R_1}$, clearly has Fisher information no greater than that in $\mathcal{P}_{\varepsilon,R_1,R_2}$. Indeed, the least favorable distribution in $\mathcal{P}_{\varepsilon,R_1}$ (derived by Huber) can easily be shown not to be a member of $\mathcal{P}_{\varepsilon,R_1,R_2}$, by noting that the support of the minimizing H distribution is not $\mathbb{R}$, so that it cannot be the result of a convolution with a normal distribution $\mathcal{N}(\mathbf{0}, R_2)$. Thus, since it was shown to be unique, its Fisher information is in fact strictly less than that of the least favorable distribution in $\mathcal{P}_{\varepsilon,R_1,R_2}$. Consequently, a *conservative* approach to approximating a minimax solution is simply to use the least favorable distribution in $\mathcal{P}_{\varepsilon,R}$ for given ε and R; this has also the additional advantage of simplicity.

The well-known least favorable distribution of Huber [see, e.g., Huber (1969), pages 87–89, (1981), pages 84–85] is given by

$$f^*(x) = \begin{cases} (1-\varepsilon)\mathcal{N}(k;\,0,1)\exp\bigl(kx + k^2\bigr), & x < -k, \\ (1-\varepsilon)\mathcal{N}(x;\,0,1), & -k \le x \le k, \\ (1-\varepsilon)\mathcal{N}(k;\,0,1)\exp\bigl(-kx + k^2\bigr), & k < x, \end{cases} \tag{6.2}$$

where k is related to the fraction of "contamination" ε by

$$2\left(\frac{\mathcal{N}(k;0,1)}{k} - \int_{-\infty}^{-k} \mathcal{N}(x;0,1)\,dx\right) = \frac{\varepsilon}{1-\varepsilon} \tag{6.3}$$

For this distribution, it follows from (5.3) that the score (influence-bounding) function is

$$\psi_\varepsilon(x) = \begin{cases} -k, & x < -k, \\ x, & -k \le x \le k, \\ k, & k < x. \end{cases} \tag{6.4}$$

Thus, the transformation $\psi_\varepsilon(x)$ leaves its argument unaffected if it is within some predefined range, and truncates it if it goes beyond that range. This function illustrates well the concept of *bounded-influence* estimation. Since wild observations are truncated, no single data point can totally dominate the others; this contrasts with the Kalman filter, for instance, where any data point may have arbitrarily large influence on the estimate of the parameter. There is, nevertheless, a problem with this choice of observation noise distribution in the present problem: ψ_n^{i} is unbounded at $\pm k$. Although a simple step-function approximation of this function has performed well in simulations, there is, at this writing, no firm justification for such a substitution.

Deriving a least favorable distribution for the neighborhood $\mathcal{P}_{\varepsilon,R_1,R_2}$ seems to be destined to remain an open problem for a while longer. However, the estimator derived in this paper may be used when there is sufficient prior information to support the choice of a particular "contaminating" distribution H, or with a suitable approximation to the score functions.

7. Numerical examples. This section presents the results of some Monte Carlo simulation experiments, comparing the performance of several published robust recursive estimators for a number of different observation noise distributions. Since individual estimators could be "tuned" to function better in particular situations, these results are only intended to enable a rough and primarily qualitative comparison. The following recursive estimators were tested: the Kalman filter; the Guttman–Peña estimator; the Ershov–Lipster estimator; the Masreliez–Martin estimator; and the first-order approximation of the conditional mean estimator derived in this paper (with the approximation discussed earlier). For further details on these estimators or the numerical experiments, see Schick [(1989), Chapter 5].

At a minimum, a good robust estimator should be resistant to outliers but lose minimal efficiency at the nominal model. To verify these properties, the noise distributions used in the simulations range from the very light- to the very heavy-tailed, following the well-known Princeton robustness study [see Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972), pages 67–68]. They include the normal distribution, the scale-contaminated normal (or Gaussian mixture) distribution, the Laplace (double exponential) distribution, Tukey's "Slash" distribution (the ratio of a normally distributed random variable to a [0, 1] uniformly distributed random variable) and the Cauchy distribution.

Each simulation experiment described here consists of 200 runs of 50 time steps each. For simplicity, only the scalar time-invariant case is considered. Model parameters are $F_n = 0.1$ or 0.5 and $H_n = D_n = Q_n = R = 1$ for all n, with initial conditions $\overline{\theta}_0 = 0$ and $M_0 = 1$. Assumed outlier variances $\widetilde{R}_{\text{out}} = 4, 6.25$ and 9, and assumed fractions of "contamination" $\widetilde{\varepsilon} = 0.01, 0.05$ and 0.10 are used in constructing the recursive estimators. The contaminating normal distribution (in the scale-contaminated normal case), as well as the Laplace distribution, both have variances $R_{\text{out}} = 9$.

Table 2 illustrates the performance of various estimators when no outliers are present, by showing the percentage by which their respective mean-squared errors (MSE) exceed the optimal value given by the Kalman filter. As expected, the Guttman–Peña estimator is very close to the Kalman filter for small $\widetilde{R}_{\text{out}}$ and $\widetilde{\varepsilon}$; however, its MSE increases markedly with both parameters. The Masreliez–Martin estimator has a slightly higher MSE than the first-order approximation, and the difference between the two increases with $\widetilde{\varepsilon}$.

Table 3 also illustrates the performance of the estimators under nominal conditions, by showing the degree to which the residuals deviate from whiteness. Although the Kalman filter theoretically results in white residuals, that was not exactly true here due to the finite number of experiments and possibly shortcomings of the pseudorandom number generator. Thus, the fractions by which the lag-1 serial correlations for robust estimators exceed that of the Kalman filter are shown in the table. Here again the performance of the Guttman–Peña estimator degrades with increasing $\widetilde{R}_{\text{out}}$ and $\widetilde{\varepsilon}$. The difference between the Masreliez–Martin estimator, which truncates observations, and the first-order approximation, which does so selectively, is clear, particularly for higher

TABLE 2
Percentage by which the MSEs of robust estimators exceed that of the Kalman filter, for the nominal (no contamination) case

	$F = 0.1$			$F = 0.5$		
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
Guttman–Peña						
$\tilde{R}_{\mathrm{out}} = 4$	0.34	2.02	4.49	0.32	2.02	4.55
$\tilde{R}_{\mathrm{out}} = 6.25$	0.94	4.97	9.83	0.98	5.19	10.43
$\tilde{R}_{\mathrm{out}} = 9$	1.78	8.29	15.30	1.90	8.90	16.81
Ershov–Lipster						
$\tilde{R}_{\mathrm{out}} = 4$		9.79			10.88	
$\tilde{R}_{\mathrm{out}} = 6.25$		14.32			16.67	
$\tilde{R}_{\mathrm{out}} = 9$		17.51			20.78	
Masreliez–Martin						
	1.48	6.11	11.20	1.66	6.94	12.88
First-order approximation						
	1.46	5.67	9.97	1.49	5.60	9.96

TABLE 3
*Fraction by which the lag-*1 *serial correlations of robust estimators exceed that of the Kalman filter, for the nominal (no contamination) case*

	$F = 0.1$			$F = 0.5$		
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
Guttman–Peña						
$\tilde{R}_{\mathrm{out}} = 4$	0.26	0.96	1.60	0.88	3.41	5.68
$\tilde{R}_{\mathrm{out}} = 6.25$	0.47	1.54	2.39	1.65	5.56	8.69
$\tilde{R}_{\mathrm{out}} = 9$	0.67	2.04	3.02	2.42	7.46	11.18
Ershov–Lipster						
$\tilde{R}_{\mathrm{out}} = 4$		1.35			5.42	
$\tilde{R}_{\mathrm{out}} = 6.25$		1.63			6.71	
$\tilde{R}_{\mathrm{out}} = 9$		1.81			7.49	
Masreliez–Martin						
	0.44	1.35	2.14	1.59	5.00	8.03
First-order approximation						
	0.42	1.28	1.96	1.55	4.41	6.85

TABLE 4

MSEs of robust estimators as percentages of that of the Kalman filter, at the time an "outlier" occurs

Mixture			Laplace		
$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
Guttman–Peña					
53.91	45.98	43.65	81.50	77.97	78.58
Ershov–Lipster					
	76.40			84.10	
Masreliez–Martin					
56.93	48.58	45.61	81.50	77.26	76.67
First-order approximation					
57.28	49.09	46.13	81.33	76.46	76.44

Slash			Cauchy		
$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
Guttman–Peña					
8.52	8.04	7.91	40.16	40.14	40.17
Ershov–Lipster					
	8.05			40.11	
Masreliez–Martin					
5.67	4.72	4.41	1.12	0.94	0.90
First-order approximation					
5.69	4.76	4.47	1.12	0.95	0.90

assumed values of $\tilde{\varepsilon}$ (i.e., lower truncation levels).

Table 4 illustrates the respective performances of the four robust recursive estimators in the presence of an outlier. For ease of comparison, all MSEs are presented as percentages of that of the Kalman filter. Also, for economy of space, only the $F_n = 0.1$, $\tilde{R}_{\text{out}} = 9$ case is shown. Since the Guttman–Peña estimator models exactly the scale-contaminated (mixture) distribution, its performance is particularly good in that case. However, it degrades severely as the outlier distribution becomes more heavy-tailed. The performance of the Masreliez–Martin and first-order approximation estimators are very similar in most cases, and the two diverge most from the others when the outliers obey the "Slash" and Cauchy distributions.

Given its relatively modest loss of efficiency under nominal conditions, and its good performance when the observation noise is very heavy-tailed, the first-order approximation of the conditional mean estimator derived here compares favorably with other robust recursive estimators. However, the choice of estimator must depend upon the particular problem at hand.

8. Conclusion. This paper follows and extends the work of Martin and Masreliez in combining the robust *location estimation* ideas of Huber with the *stochastic approximation* method of Robbins and Monro to develop a robust recursive estimator of the state of a linear dynamic system. Both point estimation and filtering seek to obtain estimates of parameters based on observations contaminated by noise, but while the parameters to be estimated are fixed in the former case, they vary according to some (possibly stochastic) model in the latter. When the "location parameter" varies randomly, that is, when process noise is present, the stochastic approximation technique cannot be used to obtain a consistent recursive estimator. Moreover, asymptotic performance measures make little sense in ths case, and a *conditional mean* estimator is sought instead.

Using an asymptotic expansion around a small parameter involving the fraction of "contamination" ε, a first-order approximation is obtained for the conditional prior distribution of the state (given all past observations) for the case where the observation noise belongs to the ε-contaminated Guassian neighborhood. This approximation makes use of the exponential stability of the Kalman filter, which ensures that the effects of past outliers attenuate fast enough. The first-order approximation to the conditional prior distribution is then used in a theorem that generalizes a result due to Masreliez, to derive a first-order approximation to the conditional mean of the state (given all past observations and the current one). This non-Gaussian estimator has the form of banks of Kalman filters and optimal smoothers weighted by the posterior probabilities that each observation was an outlier. It performs well in the presence of heavy-tailed observation noise, but whether or not its added complexity (relative to the estimator of Masreliez and Martin) is warranted depends on the particular application for which it is to be used.

The principal limitations of the robust recursive estimator derived here are the following:

1. Theorem 5.1 describes an approximate estimator that is not robust when two or more outliers occur within less than ω time intervals of each other. This is a limitation due to the fact that the approximations are of first order. Using a second-order approximation would eliminate the nonrobustness of the estimator against pairs of outliers, but not against three or more outliers. Higher-order approximations to the conditional prior and conditional mean are thus one potential direction for future research. How much they would complicate the estimator, and whether or not the result will be of any practical value, remains to be seen.
2. The estimator and approximate estimation error covariance matrix of Theorem 5.1 are defined to within $O_p(\omega^2\varepsilon^2)$. Such probabilistic bounds are not of much practical use in determining the performance of the estimator, and better measures of the estimation error constitute an important direction for future research.
3. Because the derivation of a least favorable distribution in this case remains an open problem, the estimator derived here is not minimax. Indeed, even if

the least favorable distribution could be found, there is no guarantee that it and the corresponding estimator would be a saddle-point and thus a solution to the minimax problem.

Other topics for future research include patchy outliers, process noise outliers, the continuous-time case, simultaneous estimation of model parameters, failure (jump) detection and nonlinear models.

APPENDIX A

Proof of Theorem 4.2. This is a much-abridged outline of the proof of Theorem 4.2. Details may be found in Schick [(1989), pages 130–144]. For simplicity, the case where the "contaminating" distribution H admits the Radon–Nikodym derivative h, and where $H_n = D_n = I$ for all n, is treated below. The extension to the general case is immediate.

The proof proceeds by induction. Note first that

$$\begin{aligned}\mathbf{p}(\theta_{n+1} \mid \mathcal{Z}_n)\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})\\ = \int \mathbf{p}(\theta_{n+1} \mid \theta_n)\mathbf{p}(\mathbf{z}_n \mid \theta_n)\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1})\, d\theta_n\end{aligned} \tag{A.1}$$

from the definition of the conditional and marginal probabilities, as well as the independence of $\{\mathbf{w_n}\}$ and $\{\mathbf{v_n}\}$. Moreover, some tedious manipulation, repeated completions of the square and use of the Sherman–Morrison–Woodbury theorem yield that

$$\begin{aligned}\int \mathcal{N}(\theta_{n+1}; F_n\theta_n, Q_n)\mathcal{N}(\mathbf{z}_n - \theta_n; \mathbf{0}, R)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\, d\theta_n\\ = \mathcal{N}(\theta_{n+1}; \overline{\theta}_{n+1}^0, M_{n+1}^0)\mathcal{N}(\gamma_n^0; \mathbf{0}, \Gamma_n^0),\end{aligned} \tag{A.2}$$

and, similarly, it can be shown that

$$\begin{aligned}\int \mathcal{N}(\theta_{n+1}; F_n\theta_n, Q_n)h(\mathbf{z}_n - \theta_n)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\, d\theta_n\\ = \mathcal{N}(\theta_{n+1}; \overline{\theta}_{n+1}^{n+1}, M_{n+1}^{n+1})\Xi(V_{n+1}^{n+1}\theta_{n+1};\ \zeta_{n+1}^{n+1}, Z_{n+1}^{n+1}).\end{aligned} \tag{A.3}$$

Finally, it is easy to show by integration (including an order change) that

$$\mathbf{p}(z_0) = (1-\varepsilon)\mathcal{N}(\gamma_0^0; 0, \Gamma_0^0) + \varepsilon\Xi(\tau_1^1; \mathbf{0}, W_1^1) \tag{A.4}$$

$$= \kappa_0^{-1}. \tag{A.5}$$

For the case $n = 0$, therefore, combining (A.1) with (A.2) and (A.3) yields

$$\begin{aligned}
&\mathbf{p}(\theta_1 \mid \mathbf{z}_0)\mathbf{p}(\mathbf{z}_0) \\
&\quad = \int \mathcal{N}(\theta_1; F_0\theta_0, Q_0) \tag{A.6}\\
&\qquad \times \big((1-\varepsilon)\mathcal{N}(\mathbf{z}_0 - \theta_0; 0, R) + \varepsilon h(\mathbf{z}_0 - \theta_0)\big)\mathcal{N}(\theta_0; \overline{\theta}_0, M_0)\, d\theta_0 \\
&\quad = (1-\varepsilon)\mathcal{N}(\theta_1; \overline{\theta}_1^0, M_1^0)\mathcal{N}(\gamma_0^0; \mathbf{0}, \Gamma_0^0) \\
&\qquad + \varepsilon\mathcal{N}(\theta_1; \overline{\theta}_1^1, M_1^1)\Xi(V_1^1\theta_1, \zeta_1^1, Z_1^1), \tag{A.7}
\end{aligned}$$

which, together with (A.5), establishes (4.16) for the case $n = 0$. Next, assuming by the induction argument that (4.16) holds for some n (with $m = n$ and $\ell = 1$, i.e., assuming for now that $n \leq \omega$) and once again using (A.1) yields

$$\begin{aligned}
&\mathbf{p}(\theta_{n+1} \mid \mathcal{Z}_n)\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1}) \\
&\quad = \int \mathcal{N}(\theta_{n+1}; F_n\theta_n, Q_n) \\
&\qquad \times \big((1-\varepsilon)\mathcal{N}(\mathbf{z}_n - \theta_n; 0, R) + \varepsilon h(\mathbf{z}_n - \theta_n)\big) \\
&\qquad \times \Bigg((1-\varepsilon)^n \kappa_n \kappa_n^0 \mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0) \tag{A.8}\\
&\qquad\quad + \varepsilon(1-\varepsilon)^{n-1}\kappa_n \sum_{i=1}^{n} \kappa_n^i \mathcal{N}(\theta_n; \overline{\theta}_n^i, M_n^i)\Xi(V_n^i\theta_n; \zeta_n^i, Z_n^i) \\
&\qquad\quad + O_p\left(n^2\varepsilon^2\right)\Bigg)\, d\theta_n.
\end{aligned}$$

Considerable algebraic manipulation establishes in the same fashion the validity of (4.16) (with $m = n$ and $\ell = 1$) for all n.

There remains to show that the error term remains bounded as $n \to \infty$. This proof exploits the exponential asymptotic stability of the Kalman filter, established by Theorem 4.1. Using Corollary 4.1, it can be shown that

$$\mathcal{N}(\theta_n; \overline{\theta}_n^i, M_n^i) = \mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0) + O_p(\delta^{n-i}) \tag{A.9}$$

and that

$$\Xi(V_n^i\theta_n; \zeta_n^i, Z_n^i) = \Xi(\tau_n^i; \mathbf{0}, W_n^i) + O_p(\delta^{n-i}). \tag{A.10}$$

Moreover, it can be shown that

$$\kappa_n^0 = \kappa_n^i \mathcal{N}(\tau_n^i; \mathbf{0}, W_n^i + R), \tag{A.11}$$

so that each term in the summation in (A.8) may be rewritten as

$$\begin{aligned}
&\varepsilon(1-\varepsilon)^{n-1}\kappa_n\kappa_n^i \mathcal{N}(\theta_n; \overline{\theta}_n^i, M_n^i)\Xi(V_n^i\theta_n; \zeta_n^i, Z_n^i) \\
&\quad = \varepsilon(1-\varepsilon)^{n-1}\kappa_n\kappa_n^o\left(\rho_n^i \mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0) + O_p(\delta^{n-i})\right),
\end{aligned} \tag{A.12}$$

where

$$\rho_n^i = \frac{\Xi(\tau_n^i; \mathbf{0}, W_n^i)}{\mathcal{N}(\tau_n^i, \mathbf{0}, W_n^i + R)} \tag{A.13}$$

is the likelihood ratio for the dual alternatives of whether or not $\mathbf{z}_{i-1}$ was an outlier. Once again using Corollary 4.1, it can be shown that

$$\mathbb{E}[\rho_n^i] = 1 + O_p(\delta^{n-i+1}) \tag{A.14}$$

w.p. 1, and the Chernoff bound implies that deviations from the mean vanish geometrically in probability. Thus, terms corresponding to past outliers are "absorbed" by the "no outlier" term.

To derive the coefficients and error term, suppose first that a finite number k of outliers occurred during the first n time steps. The prior probability of such an event is $\varepsilon^k(1-\varepsilon)^{n-k}$. All the outliers may have occurred during the most recent ω time steps, resulting in

$$\binom{\omega}{k} = \frac{\omega!}{k!(\omega-k)!} \tag{A.15}$$

terms in the corresponding sum. Alternatively, $k-1$ outliers may have occurred during the most recent ω time steps, and one during the earlier $n-\omega$ time steps. In this case, the effects of that early outlier will have attenuated to $O(\varepsilon)$ by (4.14), and the corresponding term will therefore be indistinguishable, to $O(\varepsilon^2)$, from the case where only $k-1$ outliers occurred. Clearly, there are

$$\binom{n-\omega}{1} = n - \omega \tag{A.16}$$

such terms. Analogous arguments can be made for $k-2, \dots, 0$ outliers occurring during the last ω time steps. Now, if no outliers at all occurred during the most recent ω steps, then this case is indistinguishable, to $O(\varepsilon^2)$, from the case where no outliers ever occurred. The same would be true if $k-1$ outliers occurred, neither of which during the most recent ω time steps, and so on. In general, therefore, the "no outlier" term has the coefficient

$$\begin{aligned}(1-\varepsilon)^n + \varepsilon(1-\varepsilon)^{n-1}\binom{n-\omega}{1} + \varepsilon^2(1-\varepsilon)^{n-2}\binom{n-\omega}{2} + \cdots \\ = (1-\varepsilon)^{\omega}.\end{aligned} \tag{A.17}$$

Similarly, the "one outlier" term corresponds to the coefficient

$$\begin{aligned}\varepsilon(1-\varepsilon)^{n-1} + \varepsilon^2(1-\varepsilon)^{n-2}\binom{n-\omega}{1} + \varepsilon^3(1-\varepsilon)^{n-3}\binom{n-\omega}{2} + \cdots \\ = \varepsilon(1-\varepsilon)^{\omega-1}.\end{aligned} \tag{A.18}$$

Similar arguments may be made for higher numbers of outliers, and the order of each term is

$$\varepsilon^k(1-\varepsilon)^{\omega-k}\frac{\omega(\omega-1)\cdots(\omega-k+1)}{k!} = O(\varepsilon^k\omega^k). \tag{A.19}$$

From (4.15), the most significant term is for the smallest possible k, that is, for $k = 2$, concluding the proof. □

APPENDIX B

Proof of Theorem 5.1. This is a much-abridged outline of the proof of Theorem 5.1, which is an extension of a theorem due to Masreliez. Details may be found in Schick [(1989), pages 147–157]. For simplicity, the case $H_n = D_n = I$ for all n is treated below. The extension to the general case is trivial.

Note first that

$$\mathbf{p}(\theta_n \mid \mathcal{Z}_n) = \frac{\mathbf{p}(\mathbf{z}_n \mid \theta_n)\,\mathbf{p}(\theta_n \mid \mathcal{Z}_{n-1})}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}, \tag{B.1}$$

from the definition of the conditional probability and the independence of $\{\mathbf{v}_n\}$. Let f denote the Radon–Nikodym derivative of $\mathcal{F}$ (which exists by hypothesis). From (5.1) and (4.16) (with $m = n$ and $\ell = 1$), it therefore follows that

$$\begin{aligned}\mathbf{T}_n &= \frac{1}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}\\ &\quad\times \int \theta_n f(\mathbf{z}_n - \theta_n)\Bigg((1-\varepsilon)^n \kappa_n \kappa_n^0 \mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\\ &\qquad + \varepsilon(1-\varepsilon)^{n-1}\kappa_n \sum_{i=1}^{n} \kappa_n^i \mathcal{N}(\theta_n; \overline{\theta}_n^i, M_n^i)\,\Xi(V_n^i\theta_n; \zeta_n^i, Z_n^i) + O_p(n^2\varepsilon^2)\Bigg)\,d\theta_n.\end{aligned} \tag{B.2}$$

Consider the first term on the right-hand side of (B.2), that is, the "no outlier among the first n observations" term: this is basically the expression considered by Masreliez. Rewriting it as

$$\begin{aligned}&\frac{(1-\varepsilon)^n \kappa_n \kappa_n^0}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1)})} \int \theta_n f(\mathbf{z}_n - \theta_n)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\,d\theta_n\\ &\qquad = \frac{(1-\varepsilon)^n \kappa_n \kappa_n^0}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}\\ &\qquad\quad \times \Big(M_n^0 \int (M_n^0)^{-1}(\theta_n - \overline{\theta}_n^0) f(\mathbf{z}_n - \theta_n)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\,d\theta_n\\ &\qquad\quad + \overline{\theta}_n^0 \int f(\mathbf{z}_n - \theta_n)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0)\,d\theta_n\Big),\end{aligned} \tag{B.3}$$

and noting that

$$(M_n^0)^{-1}(\theta_n - \overline{\theta}_n^0)\mathcal{N}(\theta_n; \overline{\theta}_n^0, M_n^0) = -\nabla_{\theta}\mathcal{N}(\theta; \overline{\theta}_n^0, M_n^0)\Big|_{\theta = \theta_n}, \tag{B.4}$$

it can be shown, integrating by parts, that

$$\int (M_n^0)^{-1}(\boldsymbol{\theta}_n - \overline{\boldsymbol{\theta}}_n^0)\mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^0,M_n^0)\, f(\mathbf{z}_n - \boldsymbol{\theta}_n)\, d\boldsymbol{\theta}_n$$

$$= \int \mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^0,M_n^0)\nabla_{\boldsymbol{\theta}} f(\mathbf{z}_n - \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} d\boldsymbol{\theta}_n \tag{B.5}$$

$$= -\int \mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^0,M_n^0)\nabla_{\mathbf{z}} f(\mathbf{z} - \boldsymbol{\theta}_n)\Big|_{\mathbf{z}=\mathbf{z}_n} d\boldsymbol{\theta}_n \tag{B.6}$$

$$= -\nabla_{\mathbf{z}}\mathbf{p}(\mathbf{z} \mid \mathcal{Z}_{n-1},\mathcal{H}_{n-1})\big|_{\mathbf{z}=\mathbf{z}_n} \tag{B.7}$$

$$= \mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1},\mathcal{H}_{n-1})\psi_n^0(\gamma_n^0), \tag{B.8}$$

from (5.3). It follows, after some manipulation, that

$$\begin{aligned} &\frac{(1-\varepsilon)^n\kappa_n\kappa_n^0}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}\int \boldsymbol{\theta}_n f(\mathbf{z}_n - \boldsymbol{\theta}_n)\mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^0,M_n^0)\, d\boldsymbol{\theta}_n \\ &\quad = (1-\varepsilon)^n\kappa_{n+1}\pi_n^0\mathbf{T}_n^0. \end{aligned} \tag{B.9}$$

Consider now each term in the summation in (B.2), that is, the "exactly one outlier among the first n observations, at time $i-1$" terms. Since these are $O(\varepsilon)$, the following approximation is used in the sequel:

$$f(\mathbf{z}_n - \boldsymbol{\theta}_n) = (1-\varepsilon)\mathcal{N}(\mathbf{z}_n;\boldsymbol{\theta}_n,R) + O(\varepsilon), \tag{B.10}$$

This permits manipulations similar to those used for the "no outlier" term. Note next that, by (4.9),

$$\Xi(V_n^i\boldsymbol{\theta}_n;\zeta_n^i,Z_n^i) = \Xi(\zeta_n^i - V_n^i\boldsymbol{\theta}_n;\mathbf{0},Z_n^i). \tag{B.11}$$

Then,

$$\begin{aligned} &\frac{\varepsilon(1-\varepsilon)^{n-1}\kappa_n\kappa_n^i}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}\int \boldsymbol{\theta}_n f(\mathbf{z}_n - \boldsymbol{\theta}_n)\mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^i,M_n^i)\Xi(V_n^i\boldsymbol{\theta}_n;\zeta_n^i,Z_n^i)\, d\boldsymbol{\theta}_n \\ &= \frac{\varepsilon(1-\varepsilon)^n\kappa_n\kappa_n^i}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})}\int \boldsymbol{\theta}_n\mathcal{N}(\mathbf{z}_n;\boldsymbol{\theta}_n,R)\mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^i,M_n^i) \\ &\qquad\qquad \times \Xi(\zeta_n^i - V_n^i\boldsymbol{\theta}_n;\mathbf{0},Z_n^i)\, d\boldsymbol{\theta}_n + O(\varepsilon^2), \end{aligned} \tag{B.12}$$

and noting that

$$\mathcal{N}(\mathbf{z}_n;\boldsymbol{\theta}_n,R)\mathcal{N}(\boldsymbol{\theta}_n;\overline{\boldsymbol{\theta}}_n^i,M_n^i) = \mathcal{N}(\gamma_n^i;\mathbf{0},\Gamma_n^i)\mathcal{N}(\boldsymbol{\theta}_n;\boldsymbol{\theta}_n^i,P_n^i) \tag{B.13}$$

and

$$\nabla_{\boldsymbol{\theta}}\Xi(\zeta_n^i - V_n^i\boldsymbol{\theta};\mathbf{0},Z_n^i)\,\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} = -{V_n^i}^{\mathrm{T}}\nabla_{\mathbf{z}}\Xi(\mathbf{z} - V_n^i\boldsymbol{\theta}_n;\mathbf{0},Z_n^i)\big|_{\mathbf{z}=\mathbf{z}_{i-1}} \tag{B.14}$$

[recall from (4.28) and (4.29) that ζ_n^i is linear in $\mathbf{z}_{i-1}$], it can be shown in the same manner as before that

$$\begin{aligned}\frac{\varepsilon(1-\varepsilon)^{n-1}\kappa_n\kappa_n^i}{\mathbf{p}(\mathbf{z}_n \mid \mathcal{Z}_{n-1})} \int \theta_n f(\mathbf{z}_n - \theta_n)\mathcal{N}(\theta_n; \overline{\theta}_n^i, M_n^i)\Xi(V_n^i\theta_n; \zeta_n^i, Z_n^i)\, d\theta_n \\ = \varepsilon(1-\varepsilon)^{n-1}\kappa_{n+1}\pi_n^i\mathbf{T}_n^i + O(\varepsilon^2).\end{aligned} \tag{B.15}$$

This proves the assertion for $m = n$ and $\ell = 1$. Equation (5.6) is obtained as in the proof of Theorem 4.2.

The covariance in (5.13) is obtained in much the same manner, starting with (5.2), rewriting the quadratic product as

$$(\theta_n - \mathbf{T}_n)(\theta_n - \mathbf{T}_n)^{\mathrm{T}} = \left(\theta_n - \overline{\theta}_n^0 + \overline{\theta}_n^0 - \mathbf{T}_n\right)\left(\theta_n - \overline{\theta}_n^0 + \overline{\theta}_n^0 - \mathbf{T}_n\right)^{\mathrm{T}} \tag{B.16}$$

and expanding. □

Acknowledgments. The authors are grateful to Dr. Ofer Zeitouni and Professor Peter J. Huber for their helpful comments, and to an Associate Editor and the referees for suggestions that have considerably improved the legibility of this paper.

REFERENCES

AGEE, W. S. and DUNN, B. A. (1980). Robust filtering and smoothing via Gaussian mixtures. Technical Report 73, Data Sciences Div., U.S. Army White Sands Missile Range, NM.

AGEE, W. S. and TURNER, R. H. (1979). Application of robust regression to trajectory data reduction. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds.) 107–126. Academic, New York.

AGEE, W. S., TURNER, R. H. and GOMEZ, J. E. (1979). Application of robust filtering and smoothing to tracking data. Technical Report 71, Data Sciences Div., U.S. Army White Sands Missile Range, NM.

ALSPACH, D. L. (1974). The use of Gaussian sum approximations in nonlinear filtering. In *Proc. of the Eighth Annual Princeton Conf. on Information Sciences and Systems* (M. E. van Valkenburg, ed.) 479–483. Dept. Electrical Engineering, Princeton Univ.

ANDERSON, B. D. O. (1982). Internal and external stability of linear time-varying systems. *SIAM J. Control Optim.* **20** 408–413.

ANDERSON, B. D. O. and MOORE J. B. (1979). *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, NJ.

ANDERSON, B. D. O. and MOORE, J. B. (1981). Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM J. Control Optim.* **19** 20–32.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location—Survey and Advances.* Princeton Univ. Press.

ATHANS, M., WHITING, R. H. and GRUBER, M. (1977). A suboptimal estimation algorithm with probabilistic editing for false measurements with applications to target tracking with wake phenomena. *IEEE Trans. Automat. Control* **AC-22** 372–384.

BERGER, J. O. and BERLINER, L. M. (1983). Robust Bayes and empirical Bayes, analysis with ε-contaminated priors. Technical Report 83–35, Dept. Statistics, Purdue Univ.

BERGER, J. O. and BERLINER, L. M. (1986). Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *Ann. Statist.* **14** 461–486.

BICKEL, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309.

BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 511–528. Academic, New York.

BICKEL, P. J. and COLLINS, J. R. (1983). Minimizing Fisher information over mixtures of distributions. *Sankhyā Ser. A* **45** 1–19.

BONCELET, C. G., JR. and DICKINSON, B. W. (1983). An approach to robust Kalman filtering. In *Proc.of the 22nd IEEE Conf. on Decision and Control* **1** 304–305. IEEE, New York.

BROWN L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

BRYSON, A. E., JR. and HO, Y.-C. (1975). *Applied Optimal Control: Optimization, Estimation, and Control.* Wiley, New York.

CAINES, P. E. and MAYNE, D. Q. (1970). On the discrete time matrix Riccati equation of optimal control. *Internat. J. Control.* **12** 785–794.

CASELLA, G. and STRAWDERMAN, W. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9** 868–876.

ÇETİN, A. E. and TEKALP, A. M. (1990). Robust reduced update Kalman filtering. *IEEE Trans. Circuits and Systems* **37** 155–156.

CIPRA, T. and ROMERA, R. (1991). Robust Kalman filter and its application in time series analysis. *Kybernetika* **27** 481–494.

DEYST, J. J. and PRICE, C. F. (1968). Conditions for asymptotic stability of the discrete minimum-variance linear estimator. *IEEE Trans. Automat. Control* **AC-13** 702–705.

DI MASI, G. B., RUNGGALDIER, W. J. and BARAZZI, B. (1983). Generalized finite-dimensional filters in discrete time. In *Nonlinear Stochastic Problems. Proc. of the NATO Advanced Study Inst. on Nonlinear Stochastic Problems, Armaçao de Pera, Portugal* (R. S. Bucy and J. M. F. Moura, eds.) 267–277. Reidel, Dordrecht.

DONOHO, D. L. (1978). The asymptotic variance formula and large-sample criteria for the design of robust estimators. Unpublished senior thesis, Dept. Statistics, Princeton Univ.

DORAISWAMI, R. (1976). A decision theoretic approach to parameter estimation. *IEEE Trans. Automat. Control* **AC-21** 860–866.

DUNCAN, D. B. and HORN, S. D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *J. Amer. Statist. Assoc.* **67** 815–821.

ERSHOV, A. A. (1978a). Robust filtering algorithms. *Automat. Remote Control* **39** 992–996.

ERSHOV, A. A. (1978b). Stable methods of estimating parameters. *Automat. Remote Control* **39** 1152–1181.

ERSHOV, A. A. and LIPSTER, R. SH. (1978). Robust Kalman filter in discrete time. *Automat. Remote Control* **39** 359–367.

EVANS, J., KERSTEN, P. and KURZ, L. (1976). Robustized recursive estimation with applications. *Inform. Sci.* **11** 69–92.

GEBSKI, V. and MCNEIL, D. (1984). A refined method of robust smoothing. *J. Amer. Statist. Assoc.* **79** 616–623.

GOEL, P. K. and DEGROOT, M. H. (1980). Only normal distributions have linear posterior expectations in linear regression. *J. Amer. Statist. Assoc.* **75** 895–900.

GUILBO, E. P. (1979). Robust adaptive stochastic approximation-type algorithms. In *A Link Between Science and Applications of Automatic Control. Proc. of the Seventh Triennal World Cong. of the Internat. Federation Automatic Control* (A. Niemi, ed.) **3** 2153–2157. Pergamon Press, Oxford.

GUTTMAN, I. and PEÑA, D. (1984). Robust Kalman filtering and its applications. Technical Report 2766, Mathematics Research Center, Univ. Wisconsin–Madison.

GUTTMAN, I. and PEÑA, D. (1985). Comment on "Dynamic generalized linear models and Bayesian forecasting" by M. West, P. J. Harrison and H. S. Migon. *J. Amer. Statist. Assoc.* **80** 91–92.

HAGER, W. W. and HOROWITZ, L. L. (1976). Convergence and stability properties of the discrete Riccati operator equation and the associated optimal control and filtering problems. *SIAM J. Control Optim.* **14** 295–312.

HARRISON, P. J. and STEVENS, C. F. (1971). A Bayesian approach to short-term forecasting. *Op-*

erations Research Quarterly **22** 341–362.

HARRISON, P. J. and STEVENS, C. F. (1976). Bayesian forecasting. *J. Roy. Statist. Soc. Ser. B* **38** 205–228.

HEWER, G. A., MARTIN, R. D. and ZEH, J. (1987). Robust preprocessing for Kalman filtering of glint noise. *IEEE Trans. Aerospace Electron. Systems* **AES-23** 120–128.

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

HUBER, P. J. (1969). *Théorie de l'inference statistique robuste.* Presses de l'Université de Montréal.

HUBER, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041–1067.

HUBER, P. J. (1977). *Robust Statistical Procedures.* SIAM, Philadelphia.

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

JAZWINSKI, A. H. (1970). *Stochastic Processes and Filtering Theory.* Academic, New York.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering—Transactions of the ASME* **82** 35–45.

KALMAN, R. E. and BUCY, R. S. (1961). New results in filtering and prediction theory. *Journal of Basic Engineering—Transactions of the ASME* **83** 95–108.

KASSAM, S. A. and POOR, H. V. (1985). Robust techniques for signal processing: a survey. *Proc. of the IEEE* **73** 433–481.

KIRLIN, R. L. and MOGHADDAMJOO, A. (1986). Robust adaptive Kalman filtering for systems with unknown step inputs and non-Gaussian measurement errors. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34** 252–263.

KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.* **82** 1032–1050.

KLIOKIS, E. A. (1987). Optimal filtering in discrete-time random structure systems. *Automat. Remote Control* **48** 1460–1467.

KOVAČEVIĆ, B. D. and STANKOVIĆ, S. S. (1986). Robust real-time identification for a class of linear time-varying discrete systems. *Internat. J. Systems Sci.* **17** 1409–1419.

KOVAČEVIĆ, B. D. and STANKOVIĆ, S. S. (1988). Robust real-time algorithms for identification of linear multivariable time-varying systems. *Internat. J. Control* **47** 349–362.

KÜNSCH, H. R. (1986). Discussion of "Influence functionals for time series" by R. D. Martin and V. J. Yohai. *Ann. Statist.* **14** 824–826.

LEVIN, I. K. (1980). Accuracy analysis of a robust filter of a certain type by the method of convex hulls. *Automat. Remote Control.* **5** 660–669.

LEVIT, B. YA. (1979). On the theory of the asymptotic minimax property of second order. *Theory Probab. Appl.* **24** 435–437.

LEVIT, B. YA. (1980). On asymptotic minimax estimates of the second order. *Theory Probab. Appl.* **25** 552–568.

MALLOWS, C. L. (1978). Problem 78-4: minimizing an integral. *SIAM Rev.* **10** 183.

MALLOWS, C. L. (1980). Some theory of nonlinear smoothers. *Ann. Statist.* **8** 695–715.

MARAZZI, A. (1980). Robust Bayesian estimation for the linear model. Research Report 27, Fachgruppe für Statistik, ETH Zürich.

MARAZZI, A (1985). On constrained minimization of the Bayes risk for the linear model. *Statist. Decisions* **3** 277–296.

MARAZZI, A. (1990). Restricted minimax credibility: two special cases. *Mitteilungen der Schweiz. Vereinigung der Versicherungsmathematiker* **1** 101–114.

MARTIN, R. D. (1972). Robust estimation of signal amplitude. *IEEE Trans. Inform. Theory* **IT-18** 596–606.

MARTIN, R. D. (1979). Approximate conditional-mean type smoothers and interpolators. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 117–143. Springer, Berlin.

MARTIN, R. D. and DEBOW, G. (1976). Robust filtering with data-dependent covariance. In *Proc. 1976 Information Sciences and Systems* (G. L. Meyer and W. J. Rugh, eds.). Dept. Electrical Engineering, Johns Hopkins Univ.

MARTIN, R. D. and MASRELIEZ, C. J. (1975). Robust estimation via stochastic approximation. *IEEE Trans. Inform. Theory* **IT-21** 263–271.

MARTIN, R. D. and RAFTERY, A. E. (1987). Comment: robustness, computation, and non-

Euclidean models. *J. Amer. Statist. Assoc.* **82** 1044–1050.

MARTIN, R. D. and YOHAI, V. J. (1986). Influence functionals for time series (with discussion). *Ann. Statist.* **14** 781–855.

MASRELIEZ, C. J. (1974). Approximate non-Gaussian filtering with linear state and observation relations (abstract). In *Proc. Eighth Annual Princeton Conf. Information Sciences and Systems* (M. E. van Valkenburg, ed.) 398. Dept. Electrical Engineering, Princeton Univ.

MASRELIEZ, C. J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Trans. Automat. Control* **AC-20** 107–110.

MASRELIEZ; C. J. and MARTIN, R. D. (1974). Robust Bayesian estimation for the linear model and robustizing the Kalman filter. In *Proc. Eighth Annual Princeton Conf. on Information Sciences and Systems* (M. E. van Valkenburg, ed.) 488–492. Dept. Electrical Engineering, Princeton Univ.

MASRELIEZ, C. J. and MARTIN, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Trans. Automat. Control* **AC-22** 361–371.

MATAUŠEK, M. R. and STANKOVIĆ, S. S. (1980). Robust real-time algorithm for identification of non-linear time-varying systems. *Internat. J. Control.* **31** 79–94.

MCGARTY, T. P. (1975). Bayesian outlier rejection and state estimation. *IEEE Trans. Automat. Control.* **AC-20** 682–687.

MEINHOLD, R. J. and SINGPURWALLA, N. D. (1983). Understanding the Kalman filter. *Amer. Statist.* **37** 123–127.

MEINHOLD, R. J. and SINGPURWALLA, N. D. (1989). Robustification of Kalman filter models. *J. Amer. Statist. Assoc.* **84** 479–486.

MEYR, H. and SPIES, G. (1984). The structure and performance of estimators for real-time estimation of randomly varying time delay. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32** 81–94.

MITTER, S. K. and SCHICK, I. C. (1992). Point estimation, stochastic approximation, and robust Kalman filtering. In *Systems, Models and Feedback: Theory and Applications. Proc. U.S.–Italy Workshop in Honor of Professor Antonio Ruberti* (A. Isidori and T. J. Tarn, eds.) 127–151. Birkhäuser, Boston.

MOORE, J. B. and ANDERSON, B. D. O. (1980). Coping with singular transition matrices in estimation and control stability theory. *Internat. J. Control.* **31** 571–586.

MORRIS, J. M. (1976). The Kalman filter: a robust estimator for some classes of linear quadratic problems. *IEEE Trans. Inform. Theory* **IT-22** 526–534.

NEVEL'SON, M. B. (1975). On the properties of the recursive estimates for a functional of an unknown distribution function. In *Limit Theorems of Probability Theory.* (P. Révész, ed.) 227–251. North-Holland, Amsterdam.

PEÑA, D. and GUTTMAN, I. (1988). Bayesian approach to robustifying the Kalman filter. In *Bayesian Analysis of Time Series and Dyanamic Models* (J. C. Spall, ed.) 227–258. Dekker, New York.

PEÑA, D. and GUTTMAN, I. (1989). Optimal collapsing of mixture distributions in robust recursive estimation. *Comm. Statist. Theory Methods* **18** 817–833.

PRICE, E. L. and VANDELINDE V. D. (1979). Robust estimation using the Robbins–Monro stochastic approximation algorithm. *IEEE Trans. Inform. Theory* **IT-25** 698–704.

SCHICK, I. C. (1989). Robust recursive estimation of the state of a discrete-time stochastic linear dynamic system in the presence of heavy-tailed observation noise. Ph.D. dissertation, Dept. Mathematics, MIT. [Reprinted as Report LIDS-TH-1975, Laboratory for Information and Decision Systems, MIT (1990).]

SHIRAZI, M. N., SANNOMIYA, N. and NISHIKAWA, Y. (1988). Robust ε-contaminated Gaussian filtering of discrete-time linear systems. *Internat. J. Control.* **48** 1967–1977.

SORENSON, H. W. and ALSPACH, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica* **7** 465–479.

SPALL, J. C. and WALL, K. D. (1984). Asymptotic distribution theory for the Kalman filter state estimator. *Comm. Statist. Theory Methods* **13** 1981–2003.

STANKOVIĆ, S. S. and KOVAČEVIĆ, B. (1979). Comparative analysis of a class of robust real-time identification methods. In *Identification and System Parameter Estimation. Proc. Fifth*

IFAC Symp. (R. Isermann, ed.) **1** 763–770. Pergamon, Oxford.

STANKOVIĆ, S. S. and KOVAČEVIĆ, B. D. (1986). Analysis of robust stochastic approximation algorithms for process identification. *Automatica* **22** 483–488.

STEPIŃSKI, T. (1982). Comparative study of robust methods of vehicle state estimation. In *Identification and System Parameter Estimation. Proc. Sixth IFAC Symp.* (G. A. Bekey and G. N. Saridis, eds.) **1** 829–834. Pergamon, Oxford.

STOCKINGER, N. and DUTTER, R. (1983). Robust time series analysis: an overview. Research Report 9, Institut für Statistik, Technische Universität Graz.

STOCKINGER, N. and DUTTER, R. (1987). Robust time series analysis: a survey. *Kybernetika* **23** 1–5 (Supplements) 1–92.

TANAKA, M. and KATAYAMA, T. (1987). Robust Kalman filter for linear discrete-time system with Gaussian sum noises. *Internat. J. Systems Sci.* **18** 1721–1731.

TOLLET, I. H. (1976). Robust forecasting for the linear model with emphasis on robustness toward occasional outliers. In *Proc. IEEE Internat. Conf. Cybernetics and Society* 600–605. IEEE, New York.

TSAI, C. and KURZ, L. (1982). A robustized maximum entropy approach to system identification. In *System Modeling and Optimization. Lecture Notes in Control and Inform. Sci.* **38** 276–284. Springer, Berlin.

TSAI, C. and KURZ, L. (1983). An adaptive robustizing approach to Kalman filtering. *Automatica* **19** 279–288.

TSAKNAKIS, H. and PAPANTONI-KAZAKOS, P. (1988). Outlier resistant filtering and smoothing. *Inform. and Comput.* **79** 163–192.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin et al., eds.) Stanford Univ. Press.

VANDELINDE, V. D., DORAISWAMI, R. and YURTSEVEN, H. O. (1972). Robust filtering for linear systems. In *Proc. IEEE Conf. on Decision and Control* 652–656. IEEE, New York.

WEST, M. (1981). Robust sequential approximate Bayesian estimation. *J. Roy. Statist. Soc. Ser. B* **43** 157–166.

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80** 73–83.

YURTSEVEN, H. Ö. (1979). Multistage robust filtering for linear systems. In *Proc. 18th Conf. Decision and Control* 500–501. IEEE, New York.

YURTSEVEN, H. Ö. and SINHA, A. S. C. (1978). Two-stage exact robust filtering for a single-input single-output system. In *Proc. Joint Automatic Control Conf.* **4** 165–173. Instrument Society of America, Pittsburgh, PA.

LABORATORY FOR INFORMATION
AND DECISION SYSTEMS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139