LIDS-P-2165

Point Estimation, Stochastic Approximation, and Robust Kalman Filtering

SANJOY K. MITTER and IRVIN C. SCHICK

Dedicated to Antonio Ruberti on the occasion of his sixty-fifth birthday.

Abstract. Significantly non-normal noise, and particularly the presence of outliers, severely degrades the performance of the Kalman Filter, resulting in poor state estimates, non-white residuals, and invalid inference. An approach to robustifying the Kalman Filter based on minimax theory is described. The relationship between the minimax robust estimator of location formulated by Huber, its recursive versions based on the stochastic approximation procedure of Robbins and Monro, and an approximate conditional mean filter derived via asymptotic expansion, is shown. Consistency and asymptotic normality results are given for the stochastic approximation recursion in the case of multivariate time-varying stochastic linear dynamic systems with no process noise. A first-order approximation is given for the conditional prior distribution of the state in the presence of ε-contaminated normal observation noise and normal process noise. This distribution is then used to derive a first-order approximation of the conditional mean estimator for the case where both observation and process noise are present.

1. Introduction

Kalman Filtering has found an exceptionally broad range of applications, not only for estimating the state of a dynamic system in the presence of process and observation noise, 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. It is a remarkably versatile estimator, originally derived via orthogonal projections as a generalization of the Wiener filter to non-stationary processes, then shown to be optimal in a variety of settings: as the weighted least-squares solution to a regression problem, without regard to distributional assumptions; as the Bayes estimator assuming Gaussian noise, without regard to the cost functional; and as the solution to various game theoretic problems.

Nevertheless, the Kalman Filter breaks down catastrophically in the presence of heavy-tailed noise, i.e. outliers. Even rare occurrences of unusually large observations severely degrade its performance, resulting in poor state estimates, non-white residuals, and invalid inference. A robust version of the Kalman Filter would have to satisfy two objectives: be as nearly optimal as possible when there are no outliers (under "nominal" conditions); and be resistant to outliers when they do occur (track the underlying trajectory without being unduly affected by spurious observations).

Below, the notation $\mathbf{L}(\underline{x})$ denotes the probability law of the random vector $\underline{x}$, $\mathbf{N}(\underline{\mu}, \Sigma)$ denotes a multivariate normal distribution with mean $\underline{\mu}$ and covariance Σ, and $\mathrm{N}(\underline{x}; \underline{\mu}, \Sigma)$ is the corresponding probability density function.

Consider the model

$$\underline{z}_n = H_n \, \underline{\theta}_n + D_n \, \underline{v}_n, \tag{1.1}$$

where

$$\underline{\theta}_{n+1} = F_n \, \underline{\theta}_n + \underline{w}_n, \tag{1.2}$$

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

The Kalman Filter is the estimator $\hat{\underline{\theta}}_n$ of the state $\underline{\theta}_n$ given the observations $\{ \underline{z}_1, \cdots, \underline{z}_n \}$, and obeys the well-known recursion

$$\hat{\underline{\theta}}_{n+1} = F_n \, \hat{\underline{\theta}}_n + K_{n+1} \, \underline{\gamma}_{n+1}, \tag{1.3}$$

where

$$\underline{\gamma}_{n+1} = \underline{z}_{n+1} - H_{n+1} \, F_n \, \hat{\underline{\theta}}_n \tag{1.4}$$

is the innovation at time $n+1$ and

$$\Gamma_{n+1} = H_{n+1} \, M_{n+1} \, H_{n+1}^{\mathrm{T}} + D_{n+1} \, R \, D_{n+1}^{\mathrm{T}} \tag{1.5}$$

is its covariance,

$$K_{n+1} = M_{n+1} H_{n+1}^{\mathrm{T}} \Gamma_{n+1}^{-1} \tag{1.6}$$

is the gain,

$$M_{n+1} = F_n \Sigma_n F_n^{\mathrm{T}} + Q_n \tag{1.7}$$

is the *a priori* estimation error covariance at time $n+1$ (i.e. before updating by the observation $\underline{z}_{n+1}$), and

$$\Sigma_{n+1} = (I - K_{n+1} H_{n+1}) M_{n+1} \tag{1.8}$$

is the *a posteriori* estimation error covariance at time $n+1$ (i.e. after updating). The inital condition is

$$\underline{\hat{\theta}}_0 = \underline{\overline{\theta}}_0. \tag{1.9}$$

As is clear from (1.3)-(1.4), 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 [4]). Similarly, (1.6)-(1.8) 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 $F = \mathbf{N}(\underline{0}, R)$, the residual (innovation) sequence $\{ \underline{\gamma}_1, \cdots, \underline{\gamma}_n \}$ is white and is distributed as $\mathbf{L}(\underline{\gamma}_i) = \mathbf{N}(\underline{0}, \Gamma_i)$. When F is not normal, 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 non-normal. Thus, not only is the estimate poor, but furthermore invalid inference would result from utilizing the residual sequence in the case of significant excursions from normality.

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 non-parametric statistics, Bayesian methods, or minimax theory. The purpose of this paper is to review robust recursive estimation in the context of Huber's theory of minimax robust estimation. The relationship between robust point estimation, recursive robust estimation by means of stochastic approximation, and approximate conditional mean estimation based on asymptotic expansion, is described. This provides a rigorous basis for sub-optimal filtering in the presence of non-Gaussian noise.

2. Robust Point Estimation

Let $(\mathbf{R}, \mathbf{B}, \lambda)$ be a measure space, where $\mathbf{R}$ is the real line, $\mathbf{B}$ the Borel σ-algebra, and λ the Lebesgue measure. Let F be a zero-mean probability measure on $(\mathbf{R}, \mathbf{B})$ such that F is absolutely continuous with respect to λ and admits the density $f(x) := dF(x) / dx$ a.s. in accordance with the Radon-Nikodym theorem.

For some positive integer n, let $\{ v_1, \cdots, v_n \}$ be a sample of independent random variates taking values in $\mathbf{R}$, with common distribution F. Let $\theta \in \mathbf{\Theta} \subseteq \mathbf{R}$ be a location parameter, and define the observations z_i by

$$z_i := \theta + v_i, \tag{2.1}$$

for $i = 1, \cdots, n$. Let $\mathbf{R}^n$ be the product of n copies of $\mathbf{R}$, and let $T_n : \mathbf{R}^n \to \mathbf{\Theta}$ be an estimator for the parameter θ.

A broad class of such estimators are solutions $T_n(z_1, \cdots, z_n)$ to maximization problems of the form

$$\max_{T \in \mathbf{\Theta}} \sum_{i=1}^{n} \rho(z_i - T), \tag{2.2}$$

for some suitably chosen real-valued function ρ. For instance, if $\rho(x) := \log f(x)$, then the solution of (2.2) is the maximum likelihood estimate; if $\rho(x) := - \| x \|^2$, it is the least squares estimate; if $\rho(x) := - | x |$, it is the minimum modulus estimate, i.e. the median.

Robust estimation answers the need raised by the common situation where the distribution function F is not precisely known. A class of solutions to such problems is based on *minimax theory*: the distribution F is assumed to be a member of some set of distributions, and the best estimator is sought for the least favorable member of that set, in terms of some given measure of performance. While this approach is pessimistic, since the true distribution may well not be the least favorable one, it has the advantage of providing an optimum lower bound on performance. Minimax theory has been used as a conservative approach to hypothesis testing and decision problems in the presence of statistical indeterminacy; the first to formulate a minimax theory of robust estimation was apparently Huber [5]-[9].

A suitable measure of performance for the robust estimation of a location parameter is the *asymptotic variance*. This choice has several advantages. First, as is usually the case, asymptotic analytical results are considerably easier to obtain than small sample results. Furthermore, under certain conditions, the estimator can be shown to be asymptotically normal, which has the added benefit of making possible hypothesis testing and the construction of confidence intervals. Second, the sample variance is strongly dependent on the tails of the distribution; indeed, for any estimator whose value is always contained within the convex hull of the observations, the supremum of its actual variance is infinite. Thus, the asymptotic variance is a better performance measure than the sample variance. Third, the asymptotic variance is related to the Fisher Information through the Cramér-Rao inequality, and the Fisher Information lends itself well to algebraic manipulation.

The procedure, then, is as follows. It is postulated that the unknown distribution function F is a member of a certain set $\mathbf{P}$ of distributions on

(**R**, **B**). The least favorable distribution is that member of **P** leading to the largest asymptotic variance, or, equivalently (provided that the Cramér-Rao lower bound is achieved), the one minimizing the Fisher Information. Since the maximum likelihood estimator is known to achieve the Cramér-Rao lower bound, demonstrating that the least favorable distribution and the maximum likelihood estimator associated with it are a saddle point yields a minimax robust estimator. (For a theorem that provides regularity conditions under which a distribution-estimator pair is a saddle-point solution, see Verdú and Poor [25].)

The existence of a least favorable distribution has been investigated by several researchers; indeed, one of the primary tasks of minimax theory is deriving sufficient conditions for the existence of such distributions. In general, proofs of existence involve some topological restrictions that are problematical since in many cases the sets of probability distributions of interest are not tight, so that their closures are not compact in the weak topology. To circumvent this difficulty, Huber proposes to endow the set **P** with the "vague" topology, defined as the weakest topology such that maps $P \rightarrow \int \psi \, dP$ are continuous for all continuous functions ψ with compact support. Let $I(P)$ denote the Fisher Information for the distribution P, and suppose that every $P \in \mathbf{P}$ admits a density in accordance with the Radon-Nikodym theorem. In this framework, the existence and uniqueness of the least favorable distribution in **P** are established by the following theorem due to Huber:

Theorem 2.1 *If* **P** *is vaguely compact and convex, then there is a* $P_0 \in \mathbf{P}$ *minimizing* $I(P)$. *If, furthermore,* $0 < I(P_0) < \infty$ *and the support of the corresponding density* f_0 *is convex, then* P_0 *is unique.*

Proof See Huber [5:86-90], [6:81-85], [9:79-81]. QED

'Let ρ be a continuous, convex, real-valued function of a real variable, whose derivative ψ exists a.e. and takes both negative and positive values. An alternative way of stating (2.2), provided that Θ is an open set, is

$$\sum_{i=1}^{n} \psi(z_i - T) = 0 \tag{2.3}$$

at $T = T_n(z_1, \cdots, z_n)$, where $\psi(z - T) := \alpha \, \partial\rho(z - T) / \partial T$ a.e., and α is an arbitrary constant. Choosing $\alpha = -1$ for aesthetic reasons, it follows that for the case of the maximum likelihood estimator associated with the least favorable density,

$$\psi(z - T) = -\frac{\partial}{\partial T} \log f_0(z - T) \tag{2.4}$$

$$= \frac{f_0'(z - T)}{f_0(z - T)} \tag{2.5}$$

a.s., provided that the derivatives exist. Let

$$\xi(T) := \int \psi(z-T)\, dF(z) \tag{2.6}$$

denote the expectation of ψ with shift T, provided that it exists. Note the relationship between (2.3) and (2.6). The following lemma, due to Huber, establishes the existence of the expectation in (2.6), and the fact that it crosses zero:

Lemma 2.2 *If there is a T^* such that $\xi(T^*) < \infty$ exists, then $\xi(T)$ exists for all T (though it is not necessarily finite), is monotone decreasing with T, and takes both positive and negative values.*

Proof A proof of existence is suggested in Huber [9:48]; for the rest of the proof, see Huber [5], [6:64-65]. QED

Given the conditions of Lemma 2.2, the following theorem, also due to Huber, establishes the consistency and asymptotic normality of the estimator $T_n : \mathbf{R}^n \to \Theta$ defined above:

Theorem 2.3 *If $\xi(T)$ exists and there is a T^* such that $0 < \xi(T)$ for $T < T^*$ and $\xi(T) < 0$ for $T^* < T$, and if*

$$\int |\psi(z-T)|\, dF(z) < \infty, \tag{2.7}$$

then $T_n(z_1, \cdots, z_n) \to T^*$ as $n \to \infty$ almost surely and in probability (i.e. T_n is consistent).

If, moreover, $\xi(T^*) = 0$, $\xi(T)$ is continuous, differentiable and strictly monotone in a neighborhood of T^*, and if

$$0 < \int \psi^2(z-T)\, dF(z) < \infty \tag{2.8}$$

is continuous in a neighborhood of T^*, then

$$\mathbf{L}(\sqrt{n}\,(T_n - T^*)) \to \mathbf{N}\left[0, \frac{\int \psi^2(z-T^*)\, dF(z)}{(\xi'(T^*))^2}\right] \tag{2.9}$$

as $n \to \infty$ (i.e. T_n is asymptotically normal).

Proof See Huber [5], [6:66-72]; also [9:45-50]. QED

Finally, the relationship of the results of Theorem 2.3 to the true distribution and location parameter is established by the following corollary:

Corollary 2.4 *If the conditions of Theorem 2.3 are satisfied, and if the true underlying distribution is $F = P_0$, then*

$$\mathbf{L}(\sqrt{n}\,(T_n - \theta)) \to \mathbf{N}\left[0, \frac{1}{I(P_0)}\right] \tag{2.10}$$

as $n \to \infty$ (i.e. T_n is asymptotically efficient).

Proof See Huber [5:72-73]; also Schick [22:34-36]. QED

A convenient model of indeterminacy, proposed by Huber [5], is the *ε-contaminated normal neighborhood*

$$\mathbf{P}_\varepsilon := \{ (1 - \varepsilon) \mathrm{N}(0, 1) + \varepsilon H : \; H \in \mathbf{S} \}, \tag{2.11}$$

where S is the set of all probability distributions symmetric with respect to the origin, and $0 \leq \varepsilon < 1$ is the known fraction of "contamination." Note that the presence of outliers in a nominally normal sample can be modeled here by a distribution H with tails heavier than normal. The least favorable distribution in this neighborhood is given by the following theorem:

Theorem 2.5 *For the set $\mathbf{P}_\varepsilon$, the distribution minimizing the Fisher Information is given by*

$$f_\varepsilon(x) := \begin{cases} (1 - \varepsilon) \mathrm{N}(k;0,1)\, e^{kx+k^2} & x < -k \\ (1 - \varepsilon) \mathrm{N}(x;0,1) & -k \leq x \leq k \\ (1 - \varepsilon) \mathrm{N}(k;0,1)\, e^{-kx+k^2} & k < x \end{cases} \tag{2.12}$$

where k is related to the fraction of contamination ε by

$$2 \left[\frac{\mathrm{N}(k;0,1)}{k} - \int_{-\infty}^{-k} \mathrm{N}(x;0,1)\, dx \right] = \frac{\varepsilon}{1-\varepsilon}. \tag{2.13}$$

Proof Outlines of a proof can be found in Huber [6:87-89], [9:84-85]. QED

It follows from (2.5) and (2.12) that

$$\psi_\varepsilon(x) = \begin{cases} -k & x < -k \\ x & -k \leq x \leq k \\ k & k < x \end{cases} \tag{2.14}$$

a.s. 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. It is easy to see by integrating (2.14) that the corresponding ρ_ε is quadratic in the center and linear in the tails, so that the estimator defined by (2.3) and (2.14) represents in some sense a continuum between the sample mean and the sample median. As $\varepsilon \to 0$, (2.13) implies that $k \to \infty$, so that $\rho_\varepsilon(x) \propto x^2$ resulting in the sample mean (the least square estimate). As $\varepsilon \to 1$, on the other hand, $k \to 0$, and for small k, $\rho_\varepsilon(x) \propto |x|$ approximately, corresponding to the sample median (the minimum modulus estimate).

A drawback of this approach is that solving (2.3) involves "batch" processing with some kind of iterative procedure such as the Newton-Raphson method. In other words, it requires all the observations $\{ z_1, \cdots, z_n \}$ at once. Another drawback is that it assumes that the observations are identically distributed, i.e. that θ is constant. The next section describes a recursive method that updates the estimate every time an observation z_i is received, and that allows for a linear time-variant location parameter.

3. Stochastic Approximation

It is possible to recursively maximize a stochastic function like (2.2), or find the root of a stochastic function like (2.3), by means of the *stochastic approximation* procedure based on the work of Robbins and Monro [21] and developed by many others. For general reviews of this methodology, see for instance Wasan [26:8-35], Nevel'son and Has'minskii [19:79-83, 88-94], or Kushner and Clark [10:19-47]. The use of stochastic approximation in the context of robust estimation was first proposed by Martin [11], Martin and Masreliez [12], Nevel'son [18], and Price and Vandelinde [20]. See also Englund, Holst, and Ruppert [2], who investigate the colored noise case.

For $\xi(T)$ defined in (2.6), suppose that there is a T^* such that $0 < \xi(T)$ for $T < T^*$ and $\xi(T) < 0$ for $T^* < T$. Consider the recursion

$$T_{n+1} = T_n + a_n \psi(z_n - T_n), \tag{3.1}$$

where $n = 1, 2, \cdots$, $\{a_n\}$ is a given real-valued sequence, and T_1 is an arbitrary (possibly random) starting point. There is a very considerable literature investigating conditions under which $T_n \to T^*$ as $n \to \infty$, as well as the asymptotic distribution of T_n. Note that since the value of $\psi(z_n - T_n)$ is random, it is necessary for the sequence $\{a_n\}$ to obey certain conditions in order to ensure convergence: it must tend towards zero at a rate sufficient for the error variance to vanish asymptotically, yet must not reach zero for $n < \infty$ since it must be able to compensate for any and all random perturbations due to the observations $\{z_n\}$. Indeed, there must at all times remain "an infinite amount of corrective effort" to converge to the correct limit, no matter where the estimate may have deviated (Young [28:34]).

A rather more general result than those in the literature is proven below, extending consistency and asymptotic normality results to the multivariate, time-varying case where the location parameter does not necessarily approach a limit.

For some integer p, let $(\mathbf{R}^p, \mathbf{B}, \lambda)$ be a measure space, where $\mathbf{R}$ is the real line, $\mathbf{B}$ the Borel σ-algebra, and λ the Lebesgue measure. Let F be a zero-mean probability measure on $(\mathbf{R}^p, \mathbf{B})$ such that F is absolutely continuous with respect to λ and admits the density f in accordance with the Radon-Nikodym theorem.

For some positive integer n, let $\{ \underline{v}_1, \cdots, \underline{v}_n \}$ be a sample of independent random variates taking values in $\mathbf{R}^p$, with common distribution F. Define the transformation

$$\underline{z}_n = H_n \, \underline{\theta}_n + D_n \, \underline{v}_n, \tag{3.2}$$

$n = 1, 2, \cdots$, where $\{H_n\}$ and $\{D_n\}$ are known sequences of matrices with $D_n \in \mathbf{R}^{p \times p}$ and $H_n \in \mathbf{R}^{p \times q}$, and $\underline{\theta}_n \in \mathbf{R}^q$ obeys the recursion

$$\underline{\theta}_{n+1} = F_n \, \underline{\theta}_n, \tag{3.3}$$

$n = 1, 2, \cdots$, where $\{F_n\}$ is a known sequence of matrices with $F_n \in \mathbf{R}^{q \times q}$, and $\underline{\theta}_0$ is an unknown (but finite) parameter.

Consider the recursion

$$\underline{T}_{n+1} = F_n \, \underline{T}_n + \left[(D_{n+1}^{-1} \, H_{n+1})^{\mathrm{T}} (D_{n+1}^{-1} \, H_{n+1}) \right]^{-1} (D_{n+1}^{-1} \, H_{n+1})^{\mathrm{T}} A_n \, \underline{\psi} \left[D_{n+1}^{-1} \, (\underline{z}_{n+1} - H_{n+1} \, F_n \, \underline{T}_n) \right] \tag{3.4}$$

(provided that all inverses exist), where $n = 1, 2, \cdots$, $\underline{T}_n \in \mathbf{R}^q$, $\{A_n\}$ is a given matrix sequence with $A_n \in \mathbf{R}^{q \times q}$, $\underline{T}_0$ is an arbitrary (possibly random) starting point, and $\underline{\psi}$ is related to the least favorable distribution by

$$\underline{\psi}(\underline{z} - \underline{T}) = - \underline{\nabla}_T \log f_0(\underline{z} - \underline{T}) \tag{3.5}$$

$$= - \frac{1}{f_0(\underline{z} - \underline{T})} \, \underline{\nabla}_T \, f_0(\underline{z} - \underline{T}) \tag{3.6}$$

a.s., within an arbitrary multiplicative constant. Furthermore, let

$$\underline{\xi}(\underline{T}) := E \left[\underline{\psi}(\underline{z} - \underline{T}) \right] \tag{3.7}$$

as before. Let

$$\Sigma(\underline{T}) := E \left[(\underline{\psi}(\underline{z} - \underline{T}) - \underline{\xi}(\underline{T})) (\underline{\psi}(\underline{z} - \underline{T}) - \underline{\xi}(\underline{T}))^{\mathrm{T}} \right], \tag{3.8}$$

and define

$$J(\underline{T}) := \left[\frac{\partial}{\partial t_j} \, \xi_i(\underline{t}) \right]_{\underline{t} = \underline{T}} \tag{3.9}$$

to be the Jacobian of $\underline{\xi}(\underline{T})$, provided it exists.

Note that finding the least favorable distribution in the multivariate case is not trivial. The usual ordering of matrices (given $X, Y \in \mathbf{R}^{m \times m}$, $Y > X$ if and only if $Y - X > 0$, i.e. their difference is positive definite) is not a lattice ordering. Practically, this means that (in contrast to numbers on the real line) two non-equal matrices need not have an ordered relationship. Thus, finding the member of a class of distributions that minimizes the Fisher Information is not generally possible in the multivariate case. In the special case of *spherically symmetric* distributions, the multivariate extension is of course trivial: the least favorable distributions and influence-bounding functions are found coordinatewise, and everything else follows immediately.

The asymptotic behavior of the recursion $\underline{T}_n$ is established by the following theorem:

Theorem 3.1 *Let $\underline{\xi}(\underline{T})$ exist for all $\underline{T}$, and for any $\delta > 0$ and all $q \times q$ matrices $M > 0$, let*

$$\sup_{\delta \le \|\underline{T}\|} \underline{T}^{\mathrm{T}} M \, \underline{\xi}(\underline{T}) < 0. \tag{3.10}$$

Assume there exists an $S_0 < \infty$ such that

$$E\left[\, \underline{\psi}(\underline{z}-\underline{T}) \, \underline{\psi}^{\mathrm{T}}(\underline{z}-\underline{T}) \,\right] \le S_0 \tag{3.11}$$

for all $\underline{T}$, and let $\{A_n\}$ be a sequence such that $A_n > 0$ for all n,

$$\sum_{n=1}^{\infty} A_n = \infty, \tag{3.12}$$

and

$$\sum_{n=1}^{\infty} A_n^{\mathrm{T}} A_n < \infty. \tag{3.13}$$

If there is an $\alpha < \infty$ such that for all n and all m, with $0 \le m \le n$,

$$\left[\prod_{j=m}^{n} F_j \right]^{\mathrm{T}} \left[\prod_{j=m}^{n} F_j \right] < \alpha I \tag{3.14}$$

(where products are ordered by descending index), if there is a $\beta_1 > 0$ and a $\beta_2 < \infty$ such that

$$\beta_1 I < D_n < \beta_2 I \tag{3.15}$$

for all n, and finally if there is a $\gamma_1 > 0$ and a $\gamma_2 < \infty$ such that

$$\gamma_1 I < H_n^{\mathrm{T}} H_n < \gamma_2 I \tag{3.16}$$

for all n, then, given any $\underline{T}_0 < \infty$, $\underline{T}_n - \underline{\theta}_n \to 0$ as $n \to \infty$ a.s. (i.e. $\underline{T}_n$ is consistent).

If, moreover, $\underline{\xi}(0) = 0$, $\underline{\xi}(\underline{T})$ is continuous, differentiable and strictly monotone in a neighborhood of 0 with $\|J(0)\| < \infty$, if $\Sigma(0) > 0$, $\Sigma(\underline{T})$ is continuous and bounded in a neighborhood of 0, and finally if

$$\limsup_{n \to \infty} n A_n < \infty, \tag{3.17}$$

then

$$\mathbf{L}(\, \Sigma_n^{-1/2} (\, \underline{T}_n - \underline{\theta}_n \,)\,) \to \mathbf{N}(\, 0, I \,), \tag{3.18}$$

where

$$
\begin{aligned}
\Sigma_n = & \left[(D_n^{-1} H_n)^{\mathrm{T}} (D_n^{-1} H_n) \right]^{-1} (D_n^{-1} H_n)^{\mathrm{T}} \\
& \left[I + A_{n-1} J(0) \right] (D_n^{-1} H_n F_{n-1}) \Sigma_{n-1} \\
& (D_n^{-1} H_n F_{n-1})^{\mathrm{T}} \left[I + A_{n-1} J(0) \right]^{\mathrm{T}} \\
& (D_n^{-1} H_n) \left[(D_n^{-1} H_n)^{\mathrm{T}} (D_n^{-1} H_n) \right]^{-1} \\
+ & \left[(D_n^{-1} H_n)^{\mathrm{T}} (D_n^{-1} H_n) \right]^{-1} A_{n-1} \Sigma(0) A_{n-1}^{\mathrm{T}} \\
& \left[(D_n^{-1} H_n)^{\mathrm{T}} (D_n^{-1} H_n) \right]^{-1} \qquad (3.19)
\end{aligned}
$$

with

$$\Sigma_0 = 0 \qquad (3.20)$$

(i.e. $\underline{T}_n$ is asymptotically normal).

Proof For the sake of legibility, the case $H_n = D_n = I$ for all n is treated below. The extension to the general case is straight-forward.

The proof of consistency is a generalization of Blum [1]. Defining

$$
\begin{aligned}
Y_n := E & \Big[(\underline{T}_{n+1} - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1} - \underline{\theta}_{n+1}) \\
& - (\underline{T}_n - \underline{\theta}_n)^{\mathrm{T}} (\underline{T}_n - \underline{\theta}_n) \,\Big|\, \underline{T}_1, \cdots, \underline{T}_n \Big], \qquad (3.21)
\end{aligned}
$$

it can be shown that the sequence

$$\left\{ (\underline{T}_n - \underline{\theta}_n)^{\mathrm{T}} (\underline{T}_n - \underline{\theta}_n) - \sum_{j=1}^{n-1} Y_j \right\} \qquad (3.22)$$

is a martingale. Establishing first that the expectation of the absolute value of (3.22) is bounded for all n, it follows by virtue of a martingale convergence theorem that the sequence (3.22) converges almost surely. It is then shown that each of the two terms in (3.22) does so as well, by using monotonicity and boundedness to prove that

$$
\begin{aligned}
\lim_{n \to \infty} & \Big[(\underline{T}_{n+1} - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1} - \underline{\theta}_{n+1}) \\
& - (\underline{T}_n - \underline{\theta}_n)^{\mathrm{T}} F_n^{\mathrm{T}} F_n (\underline{T}_n - \underline{\theta}_n) \\
& + 2 (F_n \underline{T}_n - \underline{\theta}_{n+1})^{\mathrm{T}} A_n \underline{\xi}(F_n \underline{T}_n - \underline{\theta}_{n+1}) \Big] \\
& = 0 \qquad (3.23)
\end{aligned}
$$

w.p.1. Some manipulation and the Chebychev inequality then imply that there exists a subsequence $\{n_m\}$ such that

$$\lim_{m \to \infty} (\, F_{n_m} \, \underline{T}_{n_m} - \underline{\theta}_{n_m+1} \,)^{\mathrm{T}} \left[\prod_{k=n_m+1}^{\infty} F_k \right]^{\mathrm{T}} \left[\prod_{k=n_m+1}^{\infty} F_k \right] A_{n_m} \, \underline{\xi}(\, F_{n_m} \, \underline{T}_{n_m} - \underline{\theta}_{n_m+1} \,) = 0 \tag{3.24}$$

w.p.1, whence it follows that

$$\lim_{m \to \infty} (\, \underline{T}_{n_m} - \underline{\theta}_{n_m} \,) = 0 \tag{3.25}$$

w.p.1. Substitution into (3.23) then generalizes (3.25) to all n, proving consistency.

The proof of asymptotic normality is a generalization of Fabian [3], and is based upon the convergence of the characteristic function. From the continuity and differentiability of $\underline{\xi}(\underline{T})$ in a neighborhood of 0 by hypothesis,

$$\underline{\xi}(\underline{T}) = J(0)\, \underline{T} + O(\, \| \underline{T} \|^2 \,) \tag{3.26}$$

for small enough $\| \underline{T} \|$. By virtue of consistency, (3.4) may thus be rewritten as

$$\begin{aligned} \underline{T}_{n+1} - \underline{\theta}_{n+1} = {} & \Big[I + A_n \, J(0) + A_n \, O_p(\, \| F_n \, \underline{T}_n - \underline{\theta}_{n+1} \| \,) \Big] \\ & \qquad F_n \, (\, \underline{T}_n - \underline{\theta}_n \,) \\ & + A_n \Big[\underline{\psi}(\, \underline{z}_{n+1} - F_n \, \underline{T}_n \,) \\ & \qquad - \underline{\xi}(\, F_n \, \underline{T}_n - \underline{\theta}_{n+1} \,) \Big] \end{aligned} \tag{3.27}$$

w.p.1 for large enough n. It is first established that, defining

$$A(\, n \,, \delta_2, \underline{T} \,, \underline{\theta} \,) := \left\{ \underline{z} : \; \| \, \underline{\psi}(\underline{z} - \underline{T}) - \underline{\xi}(\underline{T} - \underline{\theta}) \, \|^2 \geq \delta_2 \, n \right\} \tag{3.28}$$

for some $\delta_2 > 0$,

$$\lim_{n \to \infty} \int_{A(n,\delta_2,F_n \underline{T}_n,\underline{\theta}_{n+1})} \| \, \underline{\psi}(\, \underline{z} - F_n \, \underline{T}_n \,) - \underline{\xi}(\, F_n \, \underline{T}_n - \underline{\theta}_{n+1} \,) \, \|^2 \, dP_n(\underline{z}) = 0 \tag{3.29}$$

w.p.1, which is analogous to Lindeberg's condition for asymptotic normality.

Taylor approximations are then constructed for the characteristic functions of the terms in the recursion (3.27), and it is shown that the characteristic function of $\underline{T}_n - \underline{\theta}_n$ approaches that of a normal distribution as $n \to \infty$. This is aided by constructing the recursion

$$\zeta_{n+1}(\underline{s}) = \zeta_n \left[F_n^{\mathrm{T}} \left(I + A_n \, J(0) + o_p(n^{-1}) \right)^{\mathrm{T}} \underline{s} \right] \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(0) \, A_n^{\mathrm{T}} \underline{s} \right], \tag{3.30}$$

subject to the initial condition

$$\zeta_0(\underline{s}) = e^{\imath \underline{s}^{\mathrm{T}} (\underline{T}_0 - \underline{\theta}_0)}, \tag{3.31}$$

and showing that ζ_n is asymptotically equivalent to the characteristic function of $\underline{T}_n - \underline{\theta}_n$. Finally, it is shown that the limiting variance is given by (3.19) by constructing a recursion that yields the variance in the exponent of the limiting characteristic function, and proving that it is asymptotically equivalent to the asymptotic variance of $\underline{T}_n - \underline{\theta}_n$. (For a more detailed proof, see Schick [22:92-106].) QED

It is clear that one can do no better recursively than in batch mode; in other words, it is not possible to do better by considering the observations one at a time than by considering them all at once. Thus, the asymptotic variance of the recursive estimator is no smaller than that of the Huber estimator of Section 2, but it can be shown that the two are asymptotically equivalent for the right choice of gains $\{A_n\}$. (See for instance Schick [22:67].) Note in passing that if the true distribution is the least favorable one, then this choice of gain sequence results in an asymptotically efficient estimator.

This section shows the relationship between robust point estimation and robust recursive estimation. However, the estimator of Theorem 3.1 corresponds to a linear dynamic model with no process noise. In other words, it is an estimator of a location parameter that varies in a deterministic and known manner. While there may be instances that require such models, the absence of process noise makes this a special case of limited application. Not only is process noise often physically present, but it is also a useful abstraction that compensates for small and unsystematic modeling errors. The following section addresses the case where process noise is present.

4. Conditional Mean Estimation

As before, let

$$\underline{z}_n = H_n \, \underline{\theta}_n + D_n \, \underline{v}_n, \tag{4.1}$$

but let the location parameter now be random and obey the recursion

$$\underline{\theta}_{n+1} = F_n \, \underline{\theta}_n + \underline{w}_n , \tag{4.2}$$

$n = 1, 2, \cdots$, with all parameters and distributions as specified in Section 1. Again, let $\{ \underline{v}_1, \cdots, \underline{v}_n \}$ be a sample of independent random variates taking values in $\mathbf{R}^p$, with common distribution $F \in \mathbf{P}_\varepsilon$ (a multivariate version of (2.11)) having positive and bounded variance R.

In this case, asymptotic variance (or alternatively the Fisher Information) is not a meaningful measure of performance. The *conditional mean* estimator, on the other hand, is well known to have several desirable properties, such as unbiasedness and minimum error variance. The first derivation of a robust approximate conditional mean estimator in the present context is due to Masreliez and Martin [15]-[16], and is based on Masreliez [13]-[14]; some generalizations are provided by West [27].

A key assumption made by these and other authors is that at each n, the conditional probability distribution of the state $\underline{\theta}_n$ given past observations $\{ \underline{z}_0, \cdots, \underline{z}_{n-1} \}$ is normal. This assumption allows some algebraic manipulations that yield an elegant stochastic approximation-like estimator. However, while it has been shown in simulation studies to be a good approximation of the true conditional density, it is only strictly correct for finite n in the special case where F is normal (see Spall and Wall [24]), which is clearly of no interest here.

In this section, a first-order approximation of the conditional distribution prior to updating, $p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})$, is derived for the case where F is known. (The extension of this result to the least favorable distribution remains an open problem at this writing.) Although conditional normality is never exactly satisfied in the presence of non-normal noise, it is shown that the zeroeth-order term in a Taylor series representation of the distribution is indeed normal. The small parameter around which the Taylor series is constructed involves ε, the fraction of contamination. This approximation is then used, in an extension of Masreliez's theorem, to derive a first-order approximation of a robust conditional mean estimator.

Note first that the Kalman Filter recursion is 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; the following theorem is due to Moore and Anderson:

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,

$$\sum_{i=n}^{n+t} \left[\prod_{j=n}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} (D_i \; R \; D_i^{\mathrm{T}})^{-1} H_i \left[\prod_{j=n}^{i-1} F_j \right] > \alpha I \tag{4.3}$$

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

$$\sum_{i=n-s}^{n} \left[\prod_{j=i+1}^{n} F_j \right] Q_i \left[\prod_{j=i+1}^{n} F_j \right]^{\mathrm{T}} > \beta I \tag{4.4}$$

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

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

$$\underline{\tilde{\theta}}_{n+1} = (I - K_{n+1} H_{n+1}) \, F_n \, \underline{\tilde{\theta}}_n , \tag{4.5}$$

(where K_n *is the Kalman gain defined in equation (1.6)), there exist* $\lambda > 0$ *and* $0 < \delta < 1$ *such that*

$$|| \; \underline{\tilde{\theta}}_n \; || < \lambda \, \delta^n , \tag{4.6}$$

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

Proof See Moore and Anderson [17]. QED

This result is used in the following, slightly different form.

Corollary 4.2 *Let the conditions of Theorem 4.1 be satisfied, and let a* $0 < \phi < \infty$ *exist such that for all* n *,*

$$\left|\left| \prod_{j=1}^{n} F_j \right|\right| < \phi \tag{4.7}$$

(i.e. the system is uniformly stable). For $i = 1, 2$*, let*

$$\underline{\theta}^i_{n+1} = F_n \, \underline{\theta}^i_n + K^i_{n+1} (\underline{z}_{n+1} - H_{n+1} F_n \, \underline{\theta}^i_n) \tag{4.8}$$

$$K^i_n = M^i_n H_n^{\mathrm{T}} (H_n M^i_n H_n^{\mathrm{T}} + D_n \; R \; D_n^{\mathrm{T}})^{-1} \tag{4.9}$$

$$M^i_{n+1} = F_n \, P^i_n \, F_n^{\mathrm{T}} + Q_n \tag{4.10}$$

$$P^i_n = (I - K^i_n H_n) M^i_n \tag{4.11}$$

be two Kalman Filters with respective initial state means $\underline{\theta}^i_0$ *and covariances* M^i_0*,* $i = 1, 2$*. Then, there is a* $0 < \delta < 1$ *such that for any finite* $\underline{\theta}$*,*

$$\mathrm{N}(\underline{\theta} ; \underline{\theta}^1_n , M^1_n) = \mathrm{N}(\underline{\theta} ; \underline{\theta}^2_n , M^2_n) + O(\delta^n). \tag{4.12}$$

Proof Since $\mathrm{N}(\underline{x} ; \underline{\mu} , \Sigma)$ is everywhere continuously differentiable with respect to $\underline{\mu}$ and Σ except at $\Sigma = 0$, and moreover since it can be shown that M^i_n is bounded away from 0 for all n, it is possible to write a first-order Taylor series expansion of $\mathrm{N}(\underline{\theta} ; \underline{\theta}^1_n , M^1_n)$ around $\underline{\theta}^2_n$ and M^2_n. But Theorem 4.1 implies that the respective differences between M^i_n and between $\underline{\theta}^i_n$ for $i = 1, 2$ are each $O(\delta^n)$ or less. The result follows immediately. (For a detailed proof, see Schick [22:122-124].) QED

A first-order approximation of the conditional probability distribution $p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})$ of the state $\underline{\theta}_n$ given past observations $\{\underline{z}_0, \cdots, \underline{z}_{n-1}\}$ is given by the following theorem:

Theorem 4.3 *Let the conditions of Theorem 4.1 and Corollary 4.2 be satisfied for the system given by equations (4.1)-(4.2), and let δ be a real number for which (4.6) holds. Let ω be the smallest integer such that*

$$\delta^{\omega} \le \varepsilon. \tag{4.13}$$

If

$$\omega\, \varepsilon < 1 \tag{4.14}$$

and if the distribution H has bounded moments, then

$$\begin{aligned} p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}) &\\ = (1-\varepsilon)^n \kappa_n \kappa_n^0\, \mathbf{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^{\;0}) &\\ + \varepsilon(1-\varepsilon)^{n-1} \kappa_n \sum_{i=1}^{n} \kappa_n^i\, \mathbf{N}(\underline{\theta}_n; \underline{\theta}_n^i, M_n^i) &\\ \int \mathbf{N}(\underline{z}_{i-1} - \underline{\xi};\, H_{i-1}\underline{v}_n^i + H_{i-1}V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), &\\ H_{i-1}W_n^i H_{i-1}^{\mathrm{T}} - H_{i-1}V_n^i M_n^i V_n^{i\,\mathrm{T}} H_{i-1}^{\mathrm{T}}) &\\ h(\underline{\xi})\, d\underline{\xi} &\\ + O_p(n^2\varepsilon^2) & \end{aligned} \tag{4.15}$$

for all $n \le \omega$, and

$$\begin{aligned} p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}) &\\ = (1-\varepsilon)^{\omega} \kappa_n \kappa_n^0\, \mathbf{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^{\;0}) &\\ + \varepsilon(1-\varepsilon)^{\omega-1} \kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\, \mathbf{N}(\underline{\theta}_n; \underline{\theta}_n^i, M_n^i) &\\ \int \mathbf{N}(\underline{z}_{i-1} - \underline{\xi};\, H_{i-1}\underline{v}_n^i + H_{i-1}V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), &\\ H_{i-1}W_n^i H_{i-1}^{\mathrm{T}} - H_{i-1}V_n^i M_n^i V_n^{i\,\mathrm{T}} H_{i-1}^{\mathrm{T}}) &\\ h(\underline{\xi})\, d\underline{\xi} & \end{aligned}$$

$$+ \; O_p(\omega^2 \varepsilon^2) \tag{4.16}$$

for all $n \geq \omega$, *where, for* $i = 1, 2, \cdots$ *and* $n > i$,

$$\underline{\theta}_n^i = F_{n-1}\,\underline{\theta}_{n-1}^i$$

$$+ \; F_{n-1} M_{n-1}^i H_{n-1}^{\mathrm{T}} \Gamma_{n-1}^{i\;-1} (\underline{z}_{n-1} - H_{n-1}\,\underline{\theta}_{n-1}^i) \tag{4.17}$$

$$M_n^i = F_{n-1} P_{n-1}^i F_{n-1}^{\mathrm{T}} + Q_{n-1} \tag{4.18}$$

$$P_n^i = M_n^i - M_n^i H_n^{\mathrm{T}} \Gamma_n^{i\;-1} H_n M_n^i \tag{4.19}$$

$$\Gamma_n^i = H_n M_n^i H_n^{\mathrm{T}} + D_n D_n^{\mathrm{T}} \tag{4.20}$$

$$V_n^i = V_{n-1}^i P_{n-1}^i F_{n-1}^{\mathrm{T}} M_n^{i\;-1} \tag{4.21}$$

$$\underline{\nu}_n^i = \underline{\nu}_{n-1}^i + V_{n-1}^i M_{n-1}^i H_{n-1}^{\mathrm{T}} \Gamma_{n-1}^{i\;-1} (\underline{z}_{n-1} - H_{n-1}\,\underline{\theta}_{n-1}^i) \tag{4.22}$$

$$W_n^i = W_{n-1}^i - V_{n-1}^i M_{n-1}^i H_{n-1}^{\mathrm{T}} \Gamma_{n-1}^{i\;-1} H_{n-1} M_{n-1}^i V_{n-1}^{i\;\mathrm{T}} \tag{4.23}$$

$$\kappa_n^i = \kappa_{n-1}^i \, \mathbf{N}(\underline{z}_{n-1};\, H_{n-1}\,\underline{\theta}_{n-1}^i,\, \Gamma_{n-1}^i) \tag{4.24}$$

subject to the initial conditions

$$\underline{\theta}_i^i = F_{i-1}\,\underline{\theta}_{i-1}^0 \tag{4.25}$$

$$M_i^i = F_{i-1} M_{i-1}^0 F_{i-1}^{\mathrm{T}} + Q_{i-1} \tag{4.26}$$

$$V_i^i = M_{i-1}^0 F_{i-1}^{\mathrm{T}} M_i^{i\;-1} \tag{4.27}$$

$$\underline{\nu}_i^i = \underline{\theta}_{i-1}^0 \tag{4.28}$$

$$W_i^i = M_{i-1}^0 \tag{4.29}$$

$$\kappa_i^i = \kappa_{i-1}^0 \tag{4.30}$$

for $i > 0$, *and*

$$\underline{\theta}_0^0 = \overline{\underline{\theta}}_0 \tag{4.31}$$

$$M_0^0 = M_0 \tag{4.32}$$

$$\kappa_0^0 = 1. \tag{4.33}$$

The normalization constant satisfies

$$\kappa_n^{-1} = (1 - \varepsilon)^n \, \kappa_n^0$$

$$+ \; \varepsilon (1 - \varepsilon)^{n-1} \sum_{i=1}^{n} \kappa_n^i \int \mathbf{N}(\underline{z}_{i-1} - \underline{\xi};\, H_{i-1}\,\underline{\nu}_n^i,$$

$$H_{i-1} W_n^i H_{i-1}^{\mathrm{T}}) \, h(\underline{\xi}) \, d\underline{\xi} \tag{4.34}$$

for all $n \leq \omega$, *and*

$$\begin{aligned}\kappa_n^{-1} &= (1-\varepsilon)^{\omega}\,\kappa_n^0 \\ &\quad + \varepsilon(1-\varepsilon)^{\omega-1} \sum_{i=n-\omega+1}^{n} \kappa_n^i \int \mathbf{N}(\underline{z}_{i-1}-\underline{\xi};\, H_{i-1}\underline{v}_n^i, \\ &\qquad H_{i-1} W_n^i H_{i-1}^{T}\,)\, h(\underline{\xi})\, d\underline{\xi} \qquad (4.35)\end{aligned}$$

for all $n \geq \omega$.

Proof Equation (4.15) is first established by induction. There remains to show that (4.16) holds for $n > \omega$, i.e. that the number of terms in (4.15) does not increase without bound as $n \to \infty$. Corollary 4.2 and the Chernoff bound are used to demonstrate that terms in (4.15) for $j \leq n - \omega$ are "absorbed" into the zeroeth-order term with an exponentially vanishing error term. Finally, a combinatorial argument establishes that the order of the error term is $\omega^2\varepsilon^2$. (For a detailed proof, see Schick [22:130-144]; in addition, an abbreviated proof is given in Schick and Mitter [23].) QED

It is interesting to note that Equations (4.17)-(4.20) are a bank of Kalman Filters, each starting at a different time $i = 0, 1, 2, \cdots$: the cases $i > 0$ correspond to Kalman Filters skipping the ith observation, while the case $i = 0$ is based on all observations. Equations (4.21)-(4.23) are a bank of optimal fixed-point smoothers, each estimating the state at a different time $i = 0, 1, 2, \cdots$, based on all preceeding and subsequent observations. Thus, each term in the summations on the right-hand sides of (4.15)-(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.

Loosely defining a random variable distributed as H as an "outlier," the first term in (4.15)-(4.16) corresponds to the event that "there has been no outlier among the first n observations," and each term in the summation to the event "there has been exactly one outlier among the first n observations, at time $i-1$." Higher-order terms correspond to the occurrence of two or more outliers, and are absorbed into the error term.

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.

The approximate conditional prior probability distribution given by Theorem 4.3 is now used in an extension of a theorem due to Masreliez, resulting in a first-order approximation of the conditional mean estimator.

Let h denote the Radon-Nikodym derivative of the contaminating distribution H with respect to the Lebesgue measure, provided it exists. Let

$$\underline{T}_n := E\,[\,\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_n\,] \tag{4.36}$$

and

$$\Sigma_n := E\,[\,(\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} \mid \underline{z}_0, \cdots, \underline{z}_n\,] \tag{4.37}$$

respectively denote the *a posteriori* conditional mean and conditional variance of $\underline{\theta}_n$. In addition, let the score function (the additive inverse of the gradient of the logarithm) for the conditional probability of $\underline{z}_n$ given that no outliers occurred during the first $n-1$ observations be denoted by

$$\underline{\psi}_n^0(\underline{z}_n) := -\underline{\nabla}_{\underline{z}_n} \log p(\,\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0\,). \tag{4.38}$$

Similarly, for $i = 1, 2, \cdots$ and all $n \geq i$, let

$$\underline{\psi}_n^i(\underline{z}_{i-1}) := -\underline{\nabla}_{\underline{z}_{i-1}} \log p(\,\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_n = 0\,) \tag{4.39}$$

denote the score function for the conditional probability of $\underline{z}_{i-1}$ given that no outliers occurred among the remaining $n-2$ observations. Finally, for $i = 0, 1, 2, \cdots$ and all $n \geq i$, let

$$\Psi_n^i(\underline{z}) := \underline{\nabla}_{\underline{z}}\, \underline{\psi}_n^{i\,\mathrm{T}}(\underline{z}) \tag{4.40}$$

denote the additive inverse of the Hessian of the logarithm of the conditional probability, i.e. the Jacobian of $\underline{\psi}_n^i$.

A first-order approximation of the conditional mean estimator $\underline{T}_n$ of the state $\underline{\theta}_n$ given past and present observations $\{\underline{z}_0, \cdots, \underline{z}_n\}$ is given by the following theorem:

Theorem 4.4 *Let the conditions of Theorem 4.1, Corollary 4.2, and Theorem 4.3 be satisfied for the system given by equations (4.1)-(4.2). If h exists and is bounded and differentiable a.e., then*

$$\underline{T}_n = (1-\varepsilon)^n \kappa_{n+1} \pi_n^0 \underline{T}_n^0 + \varepsilon (1-\varepsilon)^{n-1} \kappa_{n+1} \sum_{i=1}^{n} \pi_n^i \underline{T}_n^i + O_p(n^2 \varepsilon^2) \tag{4.41}$$

for all $n \leq \omega$, and

$$\underline{T}_n = (1-\varepsilon)^{\omega} \kappa_{n+1} \pi_n^0 \underline{T}_n^0 + \varepsilon (1-\varepsilon)^{\omega-1} \kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \underline{T}_n^i + O_p(\omega^2 \varepsilon^2) \tag{4.42}$$

for all $n \geq \omega$, *where*

$$\underline{T}_n^0 = \underline{\theta}_n^0 + M_n^0 H_n^{\mathrm{T}} \underline{\psi}_n^0 (\underline{z}_n - H_n \underline{\theta}_n^0) \tag{4.43}$$

$$\begin{aligned} \underline{T}_n^i = \underline{\theta}_n^i &+ M_n^i H_n^{\mathrm{T}} \Gamma_n^{i\ -1} (\underline{z}_n - H_n \underline{\theta}_n^i) \\ &+ P_n^i V_n^{i\ \mathrm{T}} H_{i-1}^{\mathrm{T}}\ \underline{\psi}_n^i (\underline{z}_{i-1} - H_{i-1} \underline{\nu}_{n+1}^i) \end{aligned} \tag{4.44}$$

$$\begin{aligned} \pi_n^0 = (1 - \varepsilon) \kappa_{n+1}^0 &\\ + \varepsilon\ \kappa_n^0 \int \mathbf{N}(\underline{z}_n - \underline{\xi}; H_n \underline{\theta}_n^0, H_n M_n^0 H_n^{\mathrm{T}}) h(\underline{\xi})\, d\underline{\xi} \end{aligned} \tag{4.45}$$

$$\begin{aligned} \pi_n^i = (1 - \varepsilon) \kappa_{n+1}^i \int \mathbf{N}(\underline{z}_{i-1} - \underline{\xi}; H_{i-1} \underline{\nu}_{n+1}^i, &\\ H_{i-1} W_{n+1}^i H_{i-1}^{\mathrm{T}}) h(\underline{\xi})\, d\underline{\xi} \end{aligned} \tag{4.46}$$

and the score functions are given by

$$\begin{aligned} \underline{\psi}_n^0(\underline{z}_n - H_n \underline{\theta}_n^0) = - \underline{\nabla}_z \log \Big[(1 - \varepsilon) \mathbf{N}(\underline{z}; H_n \underline{\theta}_n^0, \Gamma_n^0) &\\ + \varepsilon \int \mathbf{N}(\underline{z} - \underline{\xi}; H_n \underline{\theta}_n^0, &\\ H_n M_n^0 H_n^{\mathrm{T}}) h(\underline{\xi})\, d\underline{\xi} \Big] \Big|_{\underline{z} = \underline{z}_n} \end{aligned} \tag{4.47}$$

$$\begin{aligned} \underline{\psi}_n^i (\underline{z}_{i-1} - H_{i-1} \underline{\nu}_{n+1}^i) = - \underline{\nabla}_z \log \int \mathbf{N}(\underline{z} - \underline{\xi}; H_{i-1} \underline{\nu}_{n+1}^i, &\\ H_{i-1} W_{n+1}^i H_{i-1}^{\mathrm{T}})\ h(\underline{\xi})\, d\underline{\xi} \Big|_{\underline{z} = \underline{z}_{i-1}} \end{aligned} \tag{4.48}$$

with $\underline{\theta}_n^i$, M_n^i, P_n^i, Γ_n^i, V_n^i, $\underline{\nu}_n^i$, W_n^i, κ_n^i, *and* κ_n *as defined in equations (4.17)-(4.24) and (4.34)-(4.35), subject to the initial conditions (4.25)-(4.33). Furthermore,*

$$\begin{aligned} \Sigma_n = (1 - \varepsilon)^n \kappa_{n+1} \pi_n^0 \Sigma_n^0 + \varepsilon (1 - \varepsilon)^{n-1} \kappa_{n+1} \sum_{i=1}^{n} \pi_n^i \Sigma_n^i &\\ + O_p(n^2 \varepsilon^2) \end{aligned} \tag{4.49}$$

for all $n \leq \omega$, *and*

$$\begin{aligned} \Sigma_n = (1 - \varepsilon)^{\omega} \kappa_{n+1} \pi_n^0 \Sigma_n^0 + \varepsilon (1 - \varepsilon)^{\omega-1} \kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \Sigma_n^i &\\ + O_p(\omega^2 \varepsilon^2) \end{aligned} \tag{4.50}$$

for all $n \geq \omega$, where

$$\begin{aligned} \Sigma_n^0 = M_n^0 - M_n^0 H_n^{\mathrm{T}} \Psi_n^0(\underline{z}_n - H_n \underline{\theta}_n^0) H_n M_n^0 &\\ + (\underline{T}_n - \underline{T}_n^0)(\underline{T}_n - \underline{T}_n^0)^{\mathrm{T}} \end{aligned} \tag{4.51}$$

$$\Sigma_n^i = P_n^i - P_n^i V_n^{i\ \mathrm{T}} H_{i-1}^{\mathrm{T}} \Psi_n^i (\underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i) H_{i-1} V_n^i P_n^i$$
$$+ (\underline{T}_n - \underline{T}_n^i)(\underline{T}_n - \underline{T}_n^i)^{\mathrm{T}}, \qquad (4.52)$$

and Ψ_n^i is given by equation (4.43), subject to (4.47)-(4.48).

Proof From Theorem 4.3, the conditional prior is the sum of an $O_p(1)$ normal distribution and ω $O_p(\varepsilon)$ terms that each involve a normal distribution convolved with the contaminating distribution h. Furthermore, each term can be treated independently by virtue of linearity. Since the zeroeth-order term is normal, the corresponding term in the expression for the conditional mean can be found by direct application of the theorem due to Masreliez [14], and has the form of (4.43). The product of normal distributions in each $O_p(\varepsilon)$ term can be rewritten by grouping together the terms involving $\underline{\theta}_n$, after which Masreliez's theorem can once again be applied by changing the order of integration, yielding terms of the form (4.44). The terms in the expression for the conditional covariance are obtained analogously. (For a detailed proof, see Schick [22:147-157]; in addition, an abbreviated proof is given in Schick and Mitter [23].) QED

The estimator of Theorem 4.4 is a weighted sum of terms having the form of stochastic approximation equations. The robust filter of Masreliez and Martin [15]-[16] is approximately equivalent to the zeroeth-order term in (4.41)-(4.42), i.e. to $\underline{T}_n^0$, although the way in which they transform a one-step estimator into a recursion is *ad hoc* and violates their assumption of conditional normality at the next time step.

Note that the current observation $\underline{z}_n$ is processed by the influence-bounding function $\underline{\psi}_n^0$, i.e. $\underline{T}_n^0$ is robust against an outlier at time n. Similarly, each past observation $\underline{z}_{i-1}$ is processed by an influence-bounding function $\underline{\psi}_n^i$, i.e. $\underline{T}_n^i$ is robust against an outlier at time $i-1$. However, $\underline{T}_n^i$ is linear in $\underline{z}_n$. This is because while the $O_p(1)$ term corresponds to the event that there were no outliers among the most recent ω observations, so that the current observation could be one with probability $O(\varepsilon)$, the $O_p(\varepsilon)$ terms each correspond to the event that there was an outlier among the most recent ω observations, so that the probability that the current observation is an outlier is only $O_p(\varepsilon^2)$.

Both Theorem 4.3 and Theorem 4.4 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, 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. Higher-order approximations for the conditional prior distribution and the conditional mean could be constructed to be robust against pairs, triplets, or higher numbers of outliers.

Unlike the Kalman Filter, the estimation error covariance in Theorem 4.4 is a function 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 pre-computed and stored, as is sometimes done with the Kalman Filter. This would drastically reduce the on-line computational burden. Moreover, the banks of parallel filters and smoothers are entirely independent of each other, so that this estimate appears to be well suited to parallel computation.

Note finally that, as can easily be verified, for $\varepsilon = 0$,

$$\underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0) = -\frac{\underline{\nabla}_{\underline{z}_n} \mathrm{N}(\underline{z}_n; H_n \underline{\theta}_n^0, \Gamma_n^0)}{\mathrm{N}(\underline{z}_n; H_n \underline{\theta}_n^0, \Gamma_n^0)} \tag{4.53}$$

$$= \Gamma_n^{0\ -1}(\underline{z}_n - H_n \underline{\theta}_n^0), \tag{4.54}$$

so that $\underline{T}_n$ reduces to the Kalman Filter when the observation noise is Gaussian.

5. Conclusion

This paper reviews Huber's minimax approach for the robust estimation of a location parameter, as well as its recursive extensions inspired by the stochastic approximation method of Robbins and Monro, and develops an approximate conditional mean estimator by constructing an asymptotic expansion for the conditional prior distribution around a small parameter involving the fraction of contamination in the observation noise.

It underscores the relationship between point estimation and filtering: both 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, i.e. 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 this case, and a conditional mean estimator is sought instead.

The derivation of the least favorable distribution in this context remains an open problem. The estimator presented here is therefore approximately Bayesian but not minimax. An approximation that has been suggested is to replace the convolution terms in Theorems 4.3 and 4.4 with Huber's least favorable distribution given in Theorem 2.5. Although this would result in a conservative estimator, its simplicity is quite appealing, and the results of simulation experiments have been favorable.

The approximate conditional mean estimator derived here is robust when outliers occur in isolation, but not when they occur in patches. Higher-order approximations to the conditional prior and conditional mean would result in estimators that are robust in the presence of patchy outliers, though at the expense of considerable additional complexity.

Other directions for future research include the application of time scaling to the problem of patchy outliers or other colored noise; the continuous-time case, for which the algebra promises to be more tractable; outliers in the process noise; fault detection and identification in the presence of outliers; and the asymptotic behavior of the approximate filters presented in this paper.

REFERENCES

[1] Blum, J.R. (1954) "Multidimensional Stochastic Approximation Methods," *Ann. Math. Stat.*, 25, 4, 737-744.

[2] Englund, J.E., U. Holst, and D. Ruppert (1988) "Recursive M-estimators of Location and Scale for Dependent Sequences," *Scandinavian J. Statistics*, 15, 2, 147-159.

[3] Fabian, V. (1968) "On Asymptotic Normality in Stochastic Approximation," *Ann. Math. Stat.*, 39, 4, 1327-1332.

[4] Goel, P.K. and M.H. DeGroot (1980) "Only Normal Distributions Have Linear Posterior Expectations in Linear Regression," *J.A.S.A.*, 75, 372, 895-900.

[5] Huber, P.J. (1964) "Robust Estimation of a Location Parameter," *Ann. Math. Stat.*, 35, 1, 73-101.

[6] Huber, P.J. (1969) *Thèorie de l'Inférence Statistique Robuste*, Presses de l'Université de Montréal (Montréal).

[7] Huber, P.J. (1972) "The 1972 Wald Lecture. Robust Statistics: a Review," *Ann. Math. Stat.*, 43, 4, 1041-1067.

[8] Huber, P.J. (1977) *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics (Philadelphia, Pennsylvania).

[9] Huber, P.J. (1981) *Robust Statistics*, John Wiley (New York).

[10] Kushner, H.J. and D.S. Clark (1978) *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag (Berlin and New York).

[11] Martin, R.D. (1972) "Robust Estimation of Signal Amplitude," *IEEE Trans. Information Theory*, IT-18, 5, 596-606.

[12] Martin, R.D. and C.J. Masreliez (1975) "Robust Estimation via Stochastic Approximation," *IEEE Trans. Information Theory*, IT-21, 3, 263-271.

[13] Masreliez, C.J. (1974) "Approximate Non-Gaussian Filtering with Linear State and Observation Relations," *Proc. Eighth Annual Princeton Conf. Information Sciences and Systems*, Dept. Electrical Engineering, Princeton University (Princeton, New Jersey), 398 (abstract only).

[14] Masreliez, C.J. (1975) "Approximate Non-Gaussian Filtering with Linear State and Observation Relations," *IEEE Trans. Automatic Control*, AC-20, 1, 107-110.

[15] Masreliez, C.J. and R.D. Martin (1974) "Robust Bayesian Estimation for the Linear Model and Robustizing the Kalman Filter," *Proc. Eighth Annual Princeton Conf. Information Sciences and Systems*, Dept. Electrical Engineering, Princeton University (Princeton, New Jersey), 488-492.

[16] Masreliez, C.J. and R.D. Martin (1977) "Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter," *IEEE Trans. Automatic Control*, AC-22, 3, 361-371.

[17] Moore, J.B. and B.D.O. Anderson (1980) "Coping with Singular Transition Matrices in Estimation and Control Stability Theory," *Int. J. Control*, 31, 3, 571-586.

[18] Nevel'son, M.B. (1975) "On the Properties of the Recursive Estimates for a Functional of an Unknown Distribution Function," in P. Révész (ed.), *Limit Theorems of Probability Theory* (Colloq. Limit Theorems of Probability and Statistics, Keszthely), North-Holland (Amsterdam and London), 227-251.

[19] Nevel'son, M.B. and R.Z. Has'minskii (1973) *Stochastic Approximation and Recursive Estimation*, American Mathematical Society (Providence, Rhode Island).

[20] Price, E.L. and V.D. Vandelinde (1979) "Robust Estimation Using the Robbins-Monro Stochastic Approximation Algorithm," *IEEE Trans. Information Theory*, IT-25, 6, 698-704.

[21] Robbins, H. and S. Monro (1951) "A Stochastic Approximation Method," *Ann. Math. Stat.*, 22, 400-407.

[22] 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. thesis, Department of Mathematics, Massachusetts Institute of Technology (Cambridge, Massachusetts). Reprinted as Report LIDS-TH-1975, Laboratory for

Information and Decision Systems, Massachusetts Institute of Technology, May 1990.

[23] Schick, I.C. and S.K. Mitter (1991) "Robust Recursive Estimation in the Presence of Heavy-Tailed Observation Noise," submitted to *Ann. Stat.*

[24] Spall, J.C. and K.D. Wall (1984) "Asymptotic Distribution Theory for the Kalman Filter State Estimator," *Commun. Statist. Theor. Meth.*, 13, 16, 1981-2003.

[25] Verdú, S. and H.V. Poor (1984) "On Minimax Robustness: a General Approach and Applications," *IEEE Trans. Automatic Control*, AC-30, 2, 328-340.

[26] Wasan, M.T. (1969) *Stochastic Approximation*, Cambridge University Press (Cambridge, U.K.).

[27] West, M. (1981) "Robust Sequential Approximate Bayesian Estimation," *J. Royal Statistical Society*, B, 43, 2, 157-166.

[28] Young, P. (1984) *Recursive Estimation and Time-Series Analysis: an Introduction*, Springer-Verlag (Berlin and New York).

Acknowledgments: This research was supported in part by the U.S. Air Force Office of Scientific Research under contracts AFOSR-85-0227 and AFOSR-89-0276.

Sanjoy K. Mitter, Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Irvin C. Schick, Network Analysis Department, BBN Communications Division, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts 02140.