May 1990 LIDS-TH-1975

ROBUST RECURSIVE ESTIMATION OF THE STATE OF A DISCRETE-TIME STOCHASTIC LINEAR DYNAMIC SYSTEM IN THE PRESENCE OF HEAVY-TAILED OBSERVATION NOISE

by

Irvin C. Schick

This report is based on the unaltered thesis of Irvin C. Schick submitted to the Department of Mathematics in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Applied Mathematics at the Massachusetts Institute of Technology. This research was conducted at the M.I.T. Laboratory for Information and Decision Systems and supported by the U.S. Air Force Office of Scientific Research under contracts AFOSR-85-0227 and AFOSR-89-0276.

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Robust Recursive Estimation of the State of a Discrete-Time Stochastic Linear Dynamic System in the Presence of Heavy-Tailed Observation Noise

by

IRVIN C. SCHICK

B.S., Electrical Engineering
Massachusetts Institute of Technology
(1976)

M.S., Chemical Engineering
Massachusetts Institute of Technology
(1978)

Submitted to the
Department of Mathematics
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy in Applied Mathematics
at the
Massachusetts Institute of Technology
May 1989

Signature of Author: ______________________________
Department of Mathematics
May 5, 1989

Certified by: ______________________________
Professor Sanjoy K. Mitter
Thesis Supervisor

Certified by: Peter J. Huber
Professor Peter J. Huber
Thesis Committee Member

Accepted by: ______________________________
Professor Daniel J. Kleitman, Chairman
Applied Mathematics Committee

Accepted by: ______________________________
Professor Sigurdur Helgason, Chairman
Departmental Graduate Committee
Department of Mathematics

Robust Recursive Estimation of the State of a Discrete-Time Stochastic Linear Dynamic System in the Presence of Heavy-Tailed Observation Noise

by
IRVIN C. SCHICK

Submitted to the Department of Mathematics
on May 5, 1989
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy in Applied Mathematics

Abstract

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 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 non-normal noise, and particularly the presence of outliers, severely degrades the performance of the Kalman Filter. This results in poor state estimates, non-white residuals, and invalid inference.

Several attempts have been made in the literature to mitigate the effects of non-normality on the Kalman Filter. These range from the *ad hoc* practice of routinely discarding observations that yield excessively large residuals, to more formal approaches based on non-parametric statistics, Bayesian methods, or minimax theory. While some of these techniques have been found empirically to work well, their theoretical justifications have remained scanty at best. Many, moreover, are based on heuristic approximations with ill-understood characteristics.

This thesis aims at providing sufficient theoretical foundations for certain robust recursive estimators to justify their use for state estimation as well as inference for linear dynamic systems. It is based on the minimax robustness concept of Huber, and the recursive estimation ideas of Martin and Masreliez. Existing results are first reviewed and standardized, not only ensuring notational consistency but also making modeling assumptions consistent with each other and correcting omissions and errors.

In particular, results pertaining to the existence and derivation of a minimax optimal robust estimator of a location parameter are reviewed in detail. This is followed by a review of stochastic approximation recursions of the Robbins-Monro form, their convergence, asymptotic normality, and asymptotic efficiency. The multivariate and time-varying cases are also described in detail.

The main results of the thesis are a first-order approximation to the conditional prior distribution of the state of a discrete-time stochastic linear dynamic system in the presence of a certain class of heavy-tailed observation noise, and a first-order approximation to the conditional mean (minimum-variance) estimator based on it.

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 the normal distribution whose first two moments are given by the Kalman Filter. Moreover, the perturbation is itself of a special form, combining distributions with moments given by banks of Kalman Filters and optimal smoothers.

This form makes it possible to derive an approximate conditional mean estimator which is a weighted sum of stochastic approximation-like terms. This estimator, while somewhat complex, is very well suited to parallel computation. It also has an intuitively appealing form, the zeroeth-order term of which is shown to be analogous to the filter of Masreliez and Martin.

Some simulation results are also presented, describing the behavior of the robust estimator for several observation noise distributions, and comparing it to that of a standard Kalman Filter as well as other published robust recursive estimators.

Thesis Supervisor: Dr. Sanjoy K. Mitter
Title: Professor of Electrical Engineering

bâtıl hemîşe bâtıl-ı bîhûdedir velî
müşkil budur ki suret-i haktan zuhûr ider

Bâkî

Acknowledgments

I am very grateful to Professor Sanjoy K. Mitter, both for originally suggesting the topic of this thesis, and for his guidance throughout the long process that led to its completion. He has been a mentor and a friend, and I thank him wholeheartedly for both.

My debt to Professor Peter J. Huber should be obvious at a glance. I am grateful to him, as well as to Professor Greta M. Ljung, for their valuable contributions as members of my Doctoral Committee.

I was very fortunate to have the opportunity to benefit from the guidance of Dr. Alexander Samarov and Dr. Ofer Zeitouni. The former, with a healthy dose of skepticism and with probing questions, helped me clarify the issues and get my work off to a good start. The latter's broad knowledge and keen insight, on the other hand, assured that this thesis finally came to a fruitful conclusion. The helpful comments of Professor Herman Chernoff are also gratefully acknowledged.

The untimely passing of the late Professor Edwin Kuh deprived me of his precious advice, and robbed the world of a great mind. I should also like to express my thanks to Professor Stephan Morgenthaler for first introducing me to the field of robust statistics, and sparking an interest in me that somehow abides even today, at the end of this ordeal. Last but not least, it is certain that I would have not been able to complete this effort were it not for two leaves of absence from Bolt Beranek and Newman, for which I shall be eternally grateful. The support and understanding of Dr. Joshua Seeger and of all my colleagues at BBN was invaluable.

Table of Contents

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 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). Originally derived *via* orthogonal projections as a generalization of the Wiener filter to non-stationary processes, the Kalman Filter has been shown to be optimal in a variety of settings (e.g. Jazwinski, 1970, pp.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, pp.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 most assuredly not robust is heavy-tailed noise, i.e. the presence of outliers. It is well known that even rare occurrences of unusually large observations severely degrade the performance of the Kalman Filter, resulting in poor state estimates, non-white residuals, and invalid inference. There is no contradiction between this fact and the findings of Morris and others. It is by now well-established that the mean-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 is likely to dominate all other observations in an algorithm that depends on squaring. In other words, optimality relative to the mean-squared error criterion must *not* be sought when the noise distribution is heavy-tailed.

Past efforts to mitigate the effects of outliers on the Kalman Filter range from *ad hoc* practices such as routinely discarding observations for which residuals are "too large," to more formal approaches based on non-parametric 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 non-normal distributions, have resulted in a strong preference in the literature for Monte Carlo simulations over analytical rigor. It is the goal of this thesis to provide sufficient theoretical foundations for certain types of robust recursive estimators to justify their use for both state estimation and inference.

It is important to bear in mind that routinely ignoring unusual observations is neither wise, nor statistically sound. Such observations may contain valuable information as to unmodeled system characteristics, failures, measurement errors, etc. But detecting unusual observations is only possible by comparison with the underlying trends and behavior; yet, it is precisely these that non-robust 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, i.e. under "nominal" conditions; and to be resistent to outliers when they do occur, i.e. to be able to extract the underlying system behavior without being unduly affected by them.

This thesis is organized as follows. The problem is formally stated in Section 1.1: Equations are given for the linear dynamic system and the Kalman Filter, and a context is proposed for deriving the robust recursive estimator. This is followed by a review of the literature in Section 1.2.

Huber's argument for a theory of robust estimation based upon minimax principles is reconstructed in Section 2.1: The asymptotic variance as measure of performance, as well as its relationship to the Fisher Information, are discussed first, followed by some properties of the Fisher Information, and conditions for the existence of a solution to the minimax problem. In Section 2.2, Huber's minimax robust estimator of location is rederived, to lay the groundwork for the development of recursive estimators of location, in Section 3.

Recursive estimators based upon the stochastic approximation method of Robbins and Monro and others are reviewed in Section 3.1, where proofs are given for convergence, asymptotic normality, and asymptotic efficiency. The multivariate generalization is discussed in Section 3.2, and these results are further generalized in Section 3.3 to the case of a time-variant location parameter, whose evolution is modeled by a deterministic linear dynamic system.

The problem of estimating the state of a stochastic dynamic system is introduced in Section 4. First, a first-order approximation is derived in Section 4.1 for the conditional prior distribution of the state given all past observations. In Section 4.2, this conditional prior is used in a generalization of a theorem due to Masreliez, to derive a first-order approximation to the conditional mean estimator. Further approximations are discussed in Section 4.3, followed by a brief review of the minimax aspects of this problem, in Section 4.4.

Numerical examples are discussed in Section 5, where various robust filters are simulated under different observation noise distributions. The latter are described in Section 5.1, and the former in Section 5.2; performance measures are discussed in Section 5.3, and the simulation results are analyzed in Section 5.4. A brief assessment of these results follows.

A summary is provided in Section 6.1, and some possible directions for future research are suggested in Section 6.2.

The contribution of this thesis is twofold. First, an attempt is made to standardize existing results on minimax robust recursive estimation, not only ensuring notational consistency but also making modeling assumptions consistent with each other and correcting numerous omissions and errors. Much of the past work on robust recursive estimation has been disparate and insufficiently formal: results from minimax robustness, stochastic approximation, and recursive estimation have been used with little regard for consistency. Thus, a self-contained presentation is given of existing results on minimax robust recursive estimation theory within a single unified framework.

Second, a robust recursive estimator is derived formally, in an effort to bridge the gap between appealing heuristics and sound theory. Since its distributional properties are known -- at least approximately -- it is possible to use this estimator for statistical inference, such as fault detection and identification. In this regard, the principal contribution of this thesis is methodological: it is shown how an asymptotic expansion may be used to derive a nonlinear filter that approximates a conditional mean estimator. The resulting estimator is shown to have good performance characteristics both under nominal conditions and in the presence of outliers.

1.1 Problem Statement

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 $\mathbf{N}(\underline{x}; \underline{\mu}, \Sigma)$ is its Radon-Nikodym derivative with respect to the Lebesgue measure.

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 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 $\mathbf{L}(\underline{w}_n) = \mathbf{N}(0, Q_n)$; $\underline{v}_n \in \mathbf{R}^p$ is the observation (measurement) noise distributed as $\mathbf{L}(\underline{v}_n) = P$, with $E[\underline{v}_n] = 0$ 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 independent for all n.

A well known estimator of the state $\underline{\theta}_n$ given the observations $\{\underline{z}_1, \cdots, \underline{z}_n\}$ is the Kalman Filter, given by the 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

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

As is clear from equations (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, 1980). Similarly, equations (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 $P = \mathrm{N}(0, R)$, the residual (innovation) sequence $\{\underline{\gamma}_1, \cdots, \underline{\gamma}_n\}$ is white and is distributed as $\mathrm{L}(\underline{\gamma}_i) = \mathrm{N}(0, \Gamma_i)$.

When P 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.

Figure 1.1 illustrates the behavior of the Kalman Filter in the presence of an outlier: the estimate tracks the state very closely until the outlier, which occurs at time $n = 20$, at which time the estimation error increases sharply; moreover, the effects of the outlier persist for some time.

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, i.e. a single outlier in the observation noise sequence must result in a single outlier in the residual sequence.
- As a corollary, the residual sequence should remain nearly white when the observation noise is normally distributed except for an occasional outlier.

It is assumed in the sequel that P, the distribution of the observation noise, is non-normal but spherically symmetric with respect to the origin, and that it belongs to a neighborhood of perturbations (in a sense to be defined) from the normal distribution. It is also assumed that the observation noise is white, i.e. that outliers occur independently. While this assumption may be seen as limiting (other models have been proposed, e.g. by Martin and Yohai, 1986), it is justified by the principal goal of this effort, which is to derive a recursive estimator that can be used for inference on the linear dynamic model in the presence of heavy-tailed noise: clearly, if outliers were allowed to occur in "patches," the distinction between model changes and sequences of outliers would become rather arbitrary, and might indeed be reduced merely to a decision based on the duration of the excursion from the predicted trajectory. This is not to say that patchy outliers do not constitute a problem worthy of study -- on the contrary, time series outliers do sometimes occur in patches, and this problem is briefly touched upon in

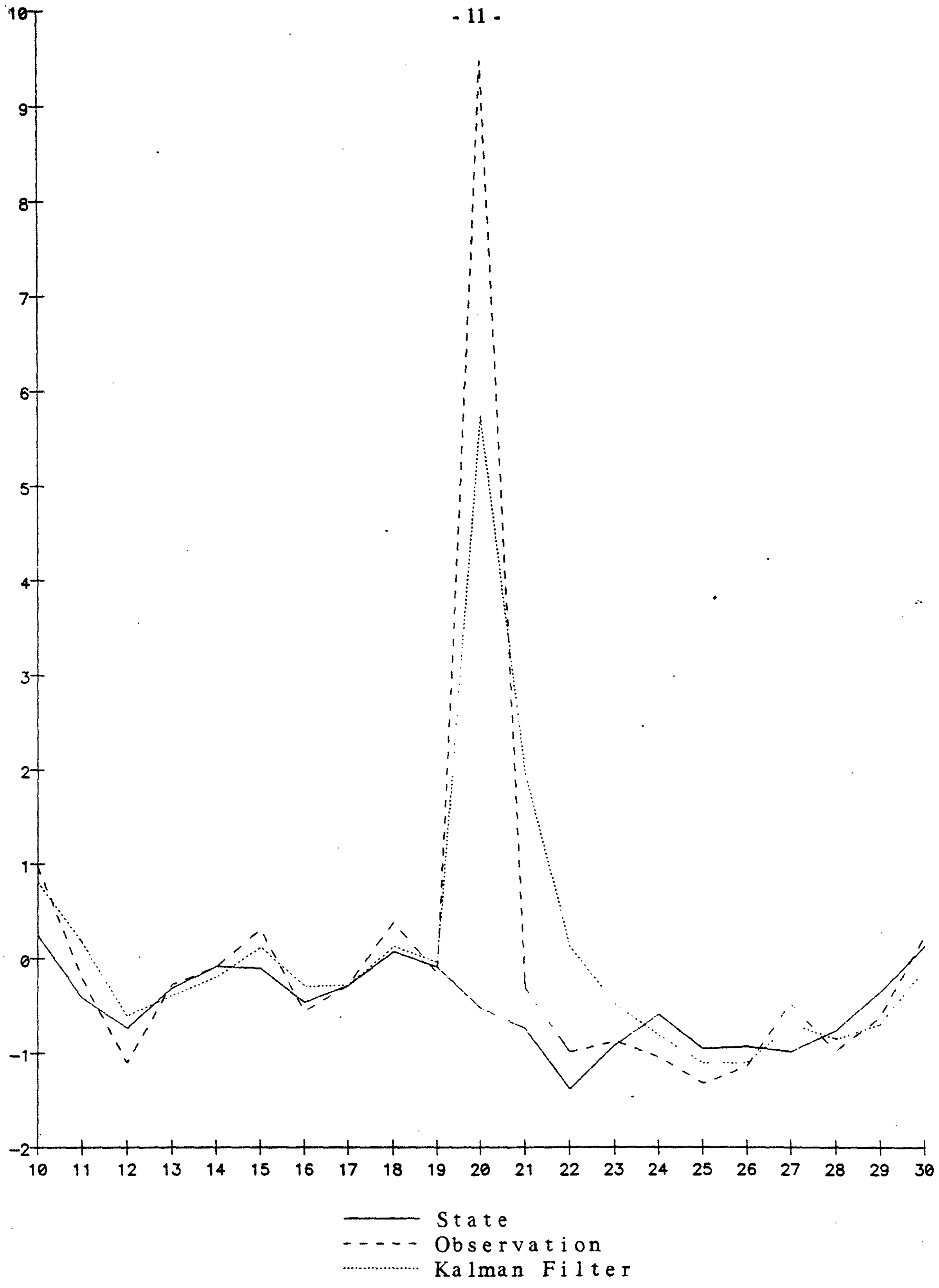

Figure 1.1 Kalman Filter at an observation outlier

Section 6.2.

The same justification as above also applies to the assumption 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 faults (except by observation of the subsequent behavior of the model, i.e. non-causally). Moreover, large enough process noise outliers can be utilized to determine the impulse response of the system, which makes them rather less interesting than observation outliers. Nevertheless, dealing with process noise outliers in real time is a problem for which satisfactory solutions remain unavailable, and is also briefly discussed in Section 6.2.

1.2 Survey of the Literature

Conscious of the deleterious effects of spurious observations on the Kalman Filter, engineers have long had recourse to *ad hoc* methods aimed at downweighting the influence of outliers on estimates. The simplest way employed is simply to discard observations for which the residual is "too large" (e.g. Meyr and Spies, 1984). Thus, the *a priori* estimate $F_n \hat{\underline{\theta}}_n$ of the state $\underline{\theta}_{n+1}$ would not be updated by $\underline{z}_{n+1}$ if, for example,

$$| [\underline{\gamma}_{n+1}]_i | > \alpha \sqrt{[\Gamma_{n+1}]_{ii}} \tag{1.10}$$

or

$$\underline{\gamma}^T_{n+1} \Gamma^{-1}_{n+1} \underline{\gamma}_{n+1} > \beta \tag{1.11}$$

for some thresholds α and β. This is equivalent to rewriting the Kalman Filter in Equation (1.3) as

$$\hat{\underline{\theta}}_{n+1} = F_n \hat{\underline{\theta}}_n + K_{n+1} \underline{\psi}_{n+1} (\underline{\gamma}_{n+1}), \tag{1.12}$$

where $\underline{\psi}_{n+1}$ 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 are sometimes used, but more for historical reasons than due to statistical considerations.) Moreover, no use whatsoever 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 statistics 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 redecending influence-bounding functions of this type give rise to non-robust covariances, since small changes in the values of the observations in the neighborhood of the thresholds may result in large variations in the value of the estimate (Huber, 1981, p.103).

Somewhat more sophisticated approaches have also been advanced to preprocess the data prior to its use in updating the Kalman Filter. In general, these techniques consist in replacing non-robust statistics by their robust counterparts. Thus, for instance, Kirlin and Moghaddamjoo (1986) use the median instead of the sample mean, while Hewer, Martin, and Zeh (1987) use Huber's M-estimator. Both papers report on applications to real data (target tracking in the case of the former, glint noise in that of the latter), where outliers were found to adversely affect 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:

(i) *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.

(ii) *Non-parametric 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.

(iii) *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 refered 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), 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, i.e. that observations are occasionally pure noise and contain no information at all. It would appear that this approach can conceptually be reduced to a simple hypothesis test to decide whether or not to update the estimate at the time of each observation. It differs considerably from the model assumed here, where the state is always observed, although the noise may occasionally contain outliers. Moreover, McGarty's method is non-recursive, 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 Section 4.1. However, while the number of components in the approximating sum grows exponentially with time in these papers, the formulation adopted in the present thesis (which exploits the exponential asymptotic stability of the Kalman Filter, as well as the fact that only one component in the mixture is of $O(1)$) 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.

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

$$R = (1-\varepsilon) R_{nom} + \varepsilon R_{out} \tag{1.13}$$

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 also 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) propose a more refined version of (1.13): they assume a distributional model for the observation noise, and compute a posterior observation noise covariance by using the posterior probability that an outlier has occurred, conditioned on the measurement. Similar approaches are discussed by Harrison and Stevens (1971, 1976). 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 this approach also tends to overestimate the covariance under nominal conditions, even though it does perform much better that (1.13). When the respective domains of the bulk of the probability masses for the underlying and outlier distributions are not sufficiently disjoint, the probability that an observation is an outlier does not decrease fast enough in the neighborhood of the mean, yielding inflated covariances and poor performance at the nominal model. Consider for instance the scalar case, with the fraction of outliers $\varepsilon = 0.1$ (as discussed in the paper), and the nominal and outlier models respectively given by N(0, 1) and N(0, 3) (as often assumed in the literature). Suppose that the innovation is $\gamma = 0$ -- i.e. the case where an outlier is least likely. Using Bayes' rule,

$$p(\, nominal \mid \gamma = 0\,) = \frac{p(\,\gamma = 0 \mid nominal\,)\, p(\, nominal\,)}{p(\,\gamma = 0\,)} \tag{1.14}$$

$$= \frac{(1-\varepsilon)\,\mathrm{N}(\,0;\,0,\,1\,)}{(1-\varepsilon)\,\mathrm{N}(\,0;\,0,\,1\,) + \varepsilon\,\mathrm{N}(\,0;\,0,\,3\,)} \tag{1.15}$$

$$= 0.94. \tag{1.16}$$

Thus, the effective covariance is given by

$$R = (0.94)\,(1) + (1-0.94)\,(3) \tag{1.17}$$

$$= 1.12, \tag{1.18}$$

implying that even in the best of all cases, the covariance is overestimated by 12%. This results in loss of efficiency at the nominal model -- as illustrated in Section 5, where the performance of this estimator is calculated for different values of ε and R_{out}. This problem carries over to virtually any model where the overlap between the nominal and outlier distributions is not negligible, such as any symmetric unimodal distribution.

A related method is that proposed by Ershov and Lipster (1978) and Ershov (1978a), where the framework is very similar to that of Guttman and Peña, but a *hard decision* is made 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. Indeed, simulations performed in the early stages of this thesis showed this filter to have excellent performance when the true noise distribution matches the modeled one. Furthermore, although the published derivation is for the scalar case, the multivariate extension is straight-forward. 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 non-robust 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 and Peña and the Ershov and Lipster filters can also be formulated in the form of Equation (1.12) -- the first with a sigmoidal 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 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 select the best among a set of competing models by means of the Akaike Information Criterion (AIC). The main difficulty with his approach, aside from the considerable computational burden it entails, lies with the mechanical and indiscriminate use of a criterion derived for another, very particular application, and not even universally accepted for that one. The well-known problems of AIC relative to order over-estimation and inapplicability to non-nested models are not addressed; as with an earlier paper by the same author on outlier detection, AIC is taken as an article of faith.

Another attempt at representing a distribution by simpler functions is that of Tsai and Kurz (1983), where a piecewise polynomial approximation is used to adaptively derive 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, very large samples are likely to be required for even moderate levels of confidence, particularly in the tails where accuracy matters most. 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 non-parametric, 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 (such as outliers) but perfect tracking of abrupt changes (such as 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. Yet, estimators are often used not merely to smooth, filter, or predict, but also for inference (e.g. model parameter estimation, jump detection, etc.), in which case knowledge of statistical properties is crucial. Finally, some non-parametric estimators may actually be special cases of more general formulations (such as the median as a limiting case of Huber's M-estimator), and should perhaps be studied in a more general framework.

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 worse 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 since 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. This is the approach taken in the present thesis, and details of the history of minimax robust estimation are provided throughout the text. Thus, only papers that specifically concern recursive state estimators are discussed here.

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 contradictory, making their complicated methods less accessible still. Moreover, their non-recursive nature makes them unsuitable for the present problem.

The literature most pertinent to this thesis (Masreliez, 1974, 1975; Masreliez and Martin, 1974, 1977; Tollet, 1976; Stanković and Kovacević, 1979; West, 1981; Stepiński, 1982) uses stochastic approximation of the Robbins-Monro type to get a recursive approximate conditional mean (minimum variance) estimator having the form of (1.12), with the influence-bounding function $\underline{\psi}_{n+1}$ given by the score of the conditional distribution of the observation, i.e.

$$\underline{\psi}_{n+1}(\underline{z}_{n+1}) = - \frac{\underline{\nabla}_{\underline{z}_{n+1}} p(\underline{z}_{n+1} \mid \underline{z}_0, \cdots . \underline{z}_n)}{p(\underline{z}_{n+1} \mid \underline{z}_0, \cdots, \underline{z}_n)} . \qquad (1.19)$$

This estimator has been found to perform well in simulation studies, 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. The present thesis extends these results and provides rigorous statistical derivations that will enable the use of this estimator for inference.

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 non-linear, 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; astonishingly, they too make the 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 (1.12) 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. Their algorithm is not strictly recursive, however, since it is based on processing all the elements of a moving window at each time step. Furthermore, while their scalar estimator is minimax, its multivariate generalization is *ad hoc* and does not obviously share this property.

Lastly, Boncelet and Dickinson (1983) describe a minimax filter obtained by applying a known robust regression technique 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.

2. Minimax Robust Estimation of a Location Parameter

It has long been recognized that spurious observations can totally offset even the soundest statistical practices, and early attempts at dealing with this problem were recorded at least as far back as the early nineteenth century (Huber, 1972; Hampel, 1973; Stigler, 1973). Nevertheless, the various methods for mitigating the effects of outliers in statistical analysis remained disparate and for the most part heuristic until Huber's landmark paper (1964). There, he proposed a new approach to robust estimation justified by *minimax theory*. This section attempts to reconstruct his argument, and is based mainly on Huber's own writings (1964, 1969, 1972, 1977, 1981).

The traditional approach to estimation is predicated upon a precise knowledge of the *form* of the probability distribution governing the random process under investigation -- if not the values of its parameters. Thus, commonly used estimators maximize or minimize some functional which derives from the distribution, as is the case with *maximum likelihood* or *maximum a posteriori probability* estimation. Alternately, a functional may be chosen for its simplicity, as is the case with *least squares* or *minimum modulus* estimation, but here again acceptance of the methodology depends on its justification through probabilistic arguments. In the case of the former, the normal (Gaussian) density, and in the case of the latter, the Laplacian (double-exponential) density, provide that justification. Indeed, Gauss formulated his density as having the form $e^{-\alpha x^2}$ precisely to justify his choice of quadratic functional (Gauss, 1821, p.98).

Robust estimation answers the need raised by the common situation where the distribution function is in fact *not* precisely known. In this case, a reasonable approach would be to assume that the density is a member of some set, or some family of parametric families, and to choose the *best* estimate for the *least favorable* member of that set -- in a sense to be discussed. While such an approach is bound to be overly pessimistic, since the true distribution may well not be the least favorable, it at least has the advantage of providing an *optimum lower bound* on performance. Consisting of Bayes solutions with respect to least favorable *a priori* distributions, minimax theory had been used earlier as a conservative approach to hypothesis testing and decision problems in the presence of statistical indeterminacy (see for instance Wald, 1950, pp.18, 89-99; Lehmann, 1959, pp.326-341; Blackwell and Girshick, 1954, pp.27, 195-199, 290-291), but Huber was apparently the first to formulate a minimax theory of robust estimation.

As a suitable performance measure for the robust estimator, Huber suggests its *asymptotic variance*. There are a number of *pros* and *cons* about this choice, including the following:

(i) The reason for having recourse to robust procedures is the lack of precise information about the distribution of the random variable(s); if the best one can do is think in asymptotic terms, why not estimate the distribution? The answer is that since outliers -- by definition -- are rare occurrences, it would take an enormous number of observations to obtain such estimates with any degree of confidence. (Nevertheless, some researchers have in fact opted for this approach, as briefly discussed in Section 1.2.) Thus, the minimax approach can be useful when the sample size is large enough to indicate deviations from the assumed model, yet not large enough to establish the precise nature of these deviations.

(ii) Since, by their very nature, outliers are infrequent, asymptotic results may not be applicable to small samples. This is a very valid criticism. At the same time, however, the probability that an outlier is present in a very small sample is remote, so that it is only in moderately large samples that outliers are likely to truly become problematical. Monte Carlo studies conducted so far suggest that asymptotic results become applicable quite rapidly.

(iii) 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. This has the added benefit of allowing its use in hypothesis testing and in the computation of confidence intervals. Seeking the distributional properties of robust estimators for small samples seems quite hopeless, in view of their complex and inevitably non-linear forms.

(iv) 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. Moreover, especially if the estimator is asymptotically normal, the "central part" of the distribution (which is of greatest importance) can better be approximated in terms of the asymptotic variance than the actual variance, yielding more accurate intervals for moderate levels of confidence.

For these and other reasons, all discussions here are based on using the asymptotic variance as measure of performance.

Huber's argument for a theory of robust estimation based upon minimax principles is reconstructed in Section 2.1, where conditions for the existence of a minimax robust estimator of a location parameter are derived. The robust estimator of location itself is rederived in Section 2.2: this result is subsequently generalized to the case where the location parameter is not constant.

2.1 Existence of the Estimator

Choosing the asymptotic variance as performance measure, it is necessary to obtain the least favorable distribution in the set, i.e. the distribution for which the minimum attainable asymptotic variance is maximum over the set. The estimator attaining that minimum asymptotic variance will then be the best robust estimator for this set of distributions. It is shown that under certain conditions, the least favorable distribution is that for which the Fisher information is minimized. While this is quite intuitive, in view of the Cramér-Rao lower bound. a more formal treatment is presented below.

Let $(\mathbf{X}, \mathbf{B})$ be a measurable space, and $\mathbf{P} := \{ P_\theta : \theta \in \Theta \}$ a family of probability measures on $(\mathbf{X}, \mathbf{B})$ such that for some σ-finite measure μ on $(\mathbf{X}, \mathbf{B})$, P_θ absolutely continuous with respect to μ for all $\theta \in \Theta$, $dP_\theta(x) \,/\, d\mu(x) := f_\theta(x)$ a.s. (μ) is a probability density in accordance with the Radon-Nikodym theorem (Halmos, 1964, pp.128-130; Loève, 1963, p.132). Suppose furthermore that $\partial f_\theta(x) \,/\, \partial\theta := f_\theta'(x)$ exists a.e. (μ) for all $\theta \in \Theta$.

Definition 2.1 The *Fisher information* of the density $f_\theta(x)$ at θ, $\theta \in \Theta$, is defined as

$$I(f_\theta) \ := \ E_{f_\theta}\left[\left[\frac{\partial}{\partial\theta} \log f_\theta(x) \right]^2 \right] \tag{2.1}$$

$$= \int \left[\frac{\partial}{\partial\theta} \log f_\theta(x) \right]^2 f_\theta(x)\, d\mu(x) \tag{2.2}$$

$$= \int \left[\frac{f_\theta'(x)}{f_\theta(x)} \right]^2 f_\theta(x)\, d\mu(x) \tag{2.3}$$

provided these expressions exist. (Kendall and Stuart, 1979, vol.2, p.10.)

The Fisher information is related to the asymptotic variance of an estimator by the following well-known relation, which is stated and proved for completeness:

Lemma 2.1 (*The Cramér-Rao lower bound*) Let $T : \mathbf{X} \to \Theta$ be an estimator of the parameter θ for the family of distributions $\mathbf{P} = \{ P_\theta : \theta \in \Theta \}$. Assume the distributions admit densities f_θ such that $f_\theta'(x)$ exists and is finite for all $\theta \in \Theta$ and all $x \in \mathbf{X}$. Let the bias of the estimator T be given by $b(\theta)$, i.e. $E_{f_\theta}[\, T \,] = \theta + b(\theta)$. Then the variance of $T(x)$ obeys

$$\mathrm{var}_{f_\theta}[\, T \,] \ \geq \ \frac{(\, 1 + b'(\theta) \,)^2}{I(f_\theta)}. \tag{2.4}$$

Proof Note first that

$$\int f_\theta(x)\, d\mu(x) \ = \ 1 \quad , \tag{2.5}$$

so that

$$\frac{\partial}{\partial\theta} \int f_\theta(x)\, d\mu(x) \ = \ 0. \tag{2.6}$$

Now, since f_θ is assumed to be differentiable with respect to θ, and $f_\theta' < \infty$ by hypothesis, there is for each θ a $\delta(\theta) > 0$ such that

$$\frac{1}{h} \left[f_{\theta+h} - f_\theta \right] \ < \ \infty \tag{2.7}$$

for $0 \leq |\, h \,| < \delta(\theta)$. Hence, taking the limit as $h \to 0$,

$$\frac{\partial}{\partial\theta} \int f_\theta(x)\, d\mu(x) \ = \ \int \frac{\partial}{\partial\theta} f_\theta(x)\, d\mu(x) \tag{2.8}$$

$$= \ 0 \tag{2.9}$$

where (2.8) follows from the Lebesgue dominated convergence theorem (Loève, 1963, pp.125-126), and (2.9) from (2.6).

By definition,

$$E_{f_\theta}[\,T\,] = \theta + b(\theta) \tag{2.10}$$

$$= \int T(x)\, f_\theta(x)\, d\mu(x) \tag{2.11}$$

Differentiating with respect to θ and once again using the dominated convergence theorem, it follows that

$$1 + b'(\theta) = \int T(x)\, \frac{\partial}{\partial\theta} f_\theta(x)\, d\mu(x) \tag{2.12}$$

$$= \int (\, T(x) - \theta - b(\theta)\,)\, \frac{\partial}{\partial\theta} f_\theta(x)\, d\mu(x) \tag{2.13}$$

$$= \int (\, T(x) - \theta - b(\theta)\,) \left[\frac{f_\theta'(x)}{f_\theta(x)} \right] f_\theta(x)\, d\mu(x) \tag{2.14}$$

where (2.13) follows from (2.9). Squaring, the Cauchy-Schwarz inequality yields

$$(\, 1 + b'(\theta)\,)^2 \le \int (\, T(x) - \theta - b(\theta)\,)^2 f_\theta(x)\, d\mu(x) \int \left[\frac{f_\theta'(x)}{f_\theta(x)} \right]^2 f_\theta(x)\, d\mu(x). \tag{2.15}$$

Recognizing that the first integral on the right hand side of (2.15) is the variance of T and the second defines the Fisher information, and dividing both sides of the inequality by the latter, proves the lemma. (See, for instance, Cox and Hinkley, 1974, p.254.) ■

The least favorable distribution is that for which the best estimator (i.e. the one minimizing the asymptotic variance) has the worst (highest) asymptotic variance over the entire class of distributions. Since the Cramér-Rao inequality (2.4) provides a lower bound on the asymptotic variance of an estimator, a reasonable approach might be to seek the distribution for which this lower bound is *highest* -- especially if assurances can be given that an estimator achieving that bound always exists. Unfortunately, dealing with bias in the minimax framework presents some difficulty, because there is no single expression for it that is valid over the entire set of probability distributions. This makes the use of (2.4) far from straightforward. If, however, it can be assumed that the estimator is *unbiased,* then $b'(\theta) = 0$ and the bound reduces to

$$\mathrm{var}_{f_\theta}[\,T\,] \ge \frac{1}{I(f_\theta)}. \tag{2.16}$$

Then, the least favorable distribution is simply the one minimizing the Fisher information, and the problem is considerably simplified. In the present application, robust estimators are sought for *location parameters.* In that special case, provided some restrictions are made on the class of probability distributions, it can be shown that Huber's robust estimator is unbiased so that (2.16) is indeed valid. This is assumed in the sequel.

As will become clear in the discussion of robust recursive estimators for linear dynamic systems (see Section 4), the principal case of interest here concerns estimators of a location parameter. Thus, it is assumed henceforth that **X** is the real line **R**, **B** the Borel σ-algebra, and μ the Lebesgue measure, and that $f_\theta(x) := f(\,x - \theta\,)$ a.s.

To determine the least favorable distribution in the manner discussed above, it is first necessary to prove the existence and uniqueness of a distribution minimizing the Fisher information. To this end, an alternative definition is proposed by Huber to incorporate situations where Definition 2.1 is inappropriate, in which case $I(f_\theta) := \infty$ is chosen.

Definition 2.2 Let **C** be the set of all continuously differentiable functions with compact support, such that for all $\psi \in$ **C**, $\int \psi^2(x)\, dP(x) > 0$. Then, the Fisher information for location of the distribution P on **R** is given by

$$I^*(P) := \sup_{\psi \in C} \frac{(\int \psi'(x)\, dP(x))^2}{\int \psi^2(x)\, dP(x)}. \tag{2.17}$$

It is shown in Theorem 2.1 that these definitions are equivalent when the expressions in equations (2.1-2.3) are well-defined. As becomes clear later, Definition 2.2 has certain features that are useful in proofs of existence and uniqueness. The following theorem is due to Huber.

Theorem 2.1 Let $\{ P_\theta : \theta \in \Theta \}$ be a location family. Then, the following two statements are equivalent:

(i) $I^*(P_\theta) < \infty$

(ii) P_θ has an absolutely continuous density f_θ, and $I(f_\theta) < \infty$.

In either case, $I(f_\theta) = I^*(P_\theta)$, and the asterisk is dropped in the sequel.

Proof Assume first that P_θ has an absolutely continuous density $f_\theta(x) = f(x - \theta)$, and that $I(f_\theta) < \infty$. Then, integrating by parts, noting that $\psi(x) = 0$ at $x = \pm\infty$, and that

$$\frac{\partial}{\partial x} f(x - \theta) = -\frac{\partial}{\partial \theta} f(x - \theta), \tag{2.18}$$

it follows that

$$\left[\int \psi'(x) f_\theta(x)\, dx \right]^2 = \left[-\int \psi(x) f_\theta'(x)\, dx \right]^2 \tag{2.19}$$

$$= \left[\int \psi(x) \frac{f_\theta'(x)}{f_\theta(x)} f_\theta(x)\, dx \right]^2 \tag{2.20}$$

$$\leq \int \psi^2(x) f_\theta(x)\, dx \int \left[\frac{f_\theta'(x)}{f_\theta(x)} \right]^2 f_\theta(x)\, dx \tag{2.21}$$

where (2.21) holds by the Cauchy-Schwarz inequality. Dividing by the first term on the right hand side of (2.21) (which is positive by definition), it then follows from (2.17) that

$$I^*(P_\theta) \leq I(f_\theta) \tag{2.22}$$

$$< \infty \tag{2.23}$$

by assumption. This proves (ii) → (i).

Conversely, assume that $I^*(P_\theta) < \infty$, and define by

$$A(\psi) = -\int \psi'(x)\, dP_\theta(x) \tag{2.24}$$

the linear functional $A : \mathbf{C} \rightarrow \mathbf{R}$ on the dense subset $\mathbf{C}$ of the Hilbert space $L_2(P_\theta)$ of square integrable functions with respect to P_θ. Noting that $\| A \|^2 = I^*(P_\theta)$ (from (2.17) and the definition of the norm), it follows that A is bounded, and can therefore be extended by continuity to the entire Hilbert space $L_2(P_\theta)$. By the Riesz representation theorem (Conway, 1985, pp.12-13; Bachman and Narici, 1966, p.15), there is a $g_\theta \in L_2(P_\theta)$ such that for all $\psi \in L_2(P_\theta)$,

$$A(\psi) = \int \psi(x)\, g_\theta(x)\, dP_\theta(x). \tag{2.25}$$

Define the function $f_\theta(x)$ as

$$f_\theta(x) = \int_{y \subseteq x} g_\theta(y)\, dP_\theta(y) \tag{2.26}$$

a.e., and proceed to prove that this yields the density associated with P_θ. Squaring (2.26) and using the Cauchy-Schwarz inequality,

$$f_\theta^2(x) \leq \int_{y \subseteq x} dP_\theta(y) \int_{y \subseteq x} g_\theta^2(y)\, dP_\theta(y) \tag{2.27}$$

$$= P_\theta(x) \int_{y \subseteq x} g_\theta^2(y)\, dP_\theta(y) \tag{2.28}$$

a.s., whence it follows that $f_\theta(x)$ is bounded a.s., and $f_\theta(x) \rightarrow 0$ for $x \rightarrow -\infty$. Furthermore, since from (2.24) and (2.25), $\int g_\theta(x)\, dP_\theta(x) = A(1) = 0$, equation (2.26) implies that $f_\theta(x) \rightarrow 0$ for $x \rightarrow +\infty$. Thus,

$$-\int \psi'(x) f_\theta(x)\, dx = -\int \psi'(x) \int_{y \subseteq x} g_\theta(y)\, dP_\theta(y)\; dx \tag{2.29}$$

$$= \int \psi(y)\, g_\theta(y)\, dP_\theta(y) \tag{2.30}$$

$$= A(\psi) \tag{2.31}$$

$$= -\int \psi'(x)\, dP_\theta(x) \tag{2.32}$$

where the order of integration is interchanged by virtue of Fubini's theorem (Halmos. 1964, p.148), (2.31) follows from (2.25), and (2.32) from (2.24). Thus, $f_\theta(x)\, dx$ and $dP_\theta(x)$ define the same linear functional on the set $\{ \psi' : \psi \in \mathbf{C} \}$ which is dense in $L_2(P_\theta)$; they therefore define the same measure, proving that f_θ is the density associated with the probability measure P_θ, and (differentiating 2.26) $g_\theta(x) = f_\theta'(x) \,/\, f_\theta(x)$ a.s..

From the Cauchy-Schwarz inequality,

$$\left[\int \psi(x)\, g_\theta(x)\, dP_\theta(x) \right]^2 \leq \int \psi^2(x)\, dP_\theta(x) \int g_\theta^2(x)\, dP_\theta(x) \tag{2.33}$$

with equality only if $\psi(x) = \alpha\, g_\theta(x)$ a.e. for some real-valued scalar α. It follows therefore that

$$I^*(P_\theta) = \sup_{\psi \in C} \frac{(\int \psi(x)\, g_\theta(x)\, dP_\theta(x))^2}{\int \psi^2(x)\, dP_\theta(x)} \tag{2.34}$$

$$= \int g_\theta^2(x)\, dP_\theta(x) \tag{2.35}$$

$$= \int \left[\frac{f_\theta'(x)}{f_\theta(x)} \right]^2 f_\theta(x)\, dx \tag{2.36}$$

$$= I(f_\theta). \tag{2.37}$$

which is finite by hypothesis, proving the theorem. (This proof follows those inspired by T. Liggett in Huber (1969, pp.78-81; 1977, p.30; 1981, pp.77-79); Huber (1964) provides a somewhat more cumbersome proof of the same theorem.) ■

The existence of a least favorable distribution for minimax problems 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. Wald (1950, pp.96-97) formulated necessary conditions for a least favorable distribution to exist, which included the restriction that the parameter set be compact. Lehmann (1952) provided some conditions under which this requirement could be relaxed, for tests involving a finite number of decisions. In general, however, proofs of existence involve some topological restrictions which 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 these difficulties, 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. This implies that some measures may be of mass less than unity, i.e. they may place nonzero mass at $\pm\infty$. According to Huber, such *substochastic* measures may in general be viewed as providing for "infinitely bad outliers", and the fact that they may have mass less than unity formalizes the practice of routinely discarding such grossly invalid data. In the present context, however, only location families are considered; since they always have the 0 measure as a limit, **P** can be assumed not to contain substochastic measures. In this framework, existence and uniqueness are proved in Theorem 2.2 using the following lemma, due to Huber.

Lemma 2.2 The Fisher information for location $I(P)$ is a convex function of P.

Proof Noting that, by linearity,

$$\frac{\partial^2}{\partial P^2} \int \psi'(x)\, dP(x) = 0 \tag{2.38}$$

and

$$\frac{\partial^2}{\partial P^2} \int \psi^2(x)\, dP(x) = 0, \tag{2.39}$$

it is easy to show that

$$\frac{\partial^2}{\partial P^2} \frac{(\int \psi'(x)\, dP(x))^2}{\int \psi^2(x)\, dP(x)} = 2 \int \psi^2(x)\, dP(x) \left[\frac{\partial}{\partial P} \frac{\int \psi'(x)\, dP(x)}{\int \psi^2(x)\, dP(x)} \right]^2 \tag{2.40}$$

$$\geq 0. \tag{2.41}$$

Thus, the quotient on the left hand side of (2.40) is a convex function of P, so that by (2.17) $I(P)$ is the supremum of a set of convex functions of P, and is therefore itself a convex function of P. (See Huber, 1981, pp.79-80.) ■

The following theorem is due to Huber.

Theorem 2.2 If $\mathbf{P}$ 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 Since $\mathbf{P}$ is vaguely compact, then from (2.17), the Fisher information $I(P)$ is the pointwise supremum of a set of vaguely continuous functions; consequently, it is lower semi-continuous as a function of P, and attains an infimum on $\mathbf{P}$. This proves the existence portion of the theorem.

Assume now that P_0 and P_1 both minimize $I(P)$. Then, by convexity (Lemma 2.2), $I(P_\lambda)$ must be constant over the subset of $\mathbf{P}$ defined by $0 \leq \lambda \leq 1$, so that

$$\frac{\partial^2}{\partial \lambda^2} I(P_\lambda) = 0. \tag{2.42}$$

Since $I(P_\lambda) < \infty$ by hypothesis, it follows from Theorem 2.1 that P_λ has an absolutely continuous density f_λ. Then, it is easy to show by straightforward differentiation that (for $f_\lambda \neq 0$)

$$\frac{\partial^2}{\partial \lambda^2} \frac{(f_\lambda')^2}{f_\lambda} = 2 \frac{(f_1' f_0 - f_0' f_1)^2}{f_\lambda^3} \tag{2.43}$$

$$\geq 0 \tag{2.44}$$

for $0 \leq \lambda \leq 1$ (where $f_\lambda' := \partial f_\lambda / \partial \theta$, and θ is the location parameter). It follows that $(f_\lambda')^2 / f_\lambda$ is a convex function of λ, and

$$\frac{1}{h} \left[\left[\frac{(f_\lambda')^2}{f_\lambda} \right]_{\lambda+h} - \frac{(f_\lambda')^2}{f_\lambda} \right]$$

is monotone in h. Thus, from (2.3), the limit as $h \to 0$ is taken and

$$\frac{\partial}{\partial \lambda} I(P_\lambda) = \frac{\partial}{\partial \lambda} \int \frac{(f_\lambda')^2}{f_\lambda}\, dx \tag{2.45}$$

$$= \int \frac{\partial}{\partial \lambda} \frac{(f_\lambda')^2}{f_\lambda}\, dx \tag{2.46}$$

by the monotone convergence theorem (Loève, 1963, p.124). Furthermore, (2.44) and monotonicity imply that

$$\frac{1}{h}\left[\left[\frac{\partial}{\partial\lambda}\frac{(f_\lambda')^2}{f_\lambda}\right]_{\lambda+h} - \frac{\partial}{\partial\lambda}\frac{(f_\lambda')^2}{f_\lambda}\right] \geq 0 \tag{2.47}$$

so that taking the limit as $h \rightarrow 0$,

$$\frac{\partial^2}{\partial\lambda^2} I(P_\lambda) = \frac{\partial}{\partial\lambda}\int \frac{\partial}{\partial\lambda}\frac{(f_\lambda')^2}{f_\lambda}\,dx \tag{2.48}$$

$$\geq \int \frac{\partial^2}{\partial\lambda^2}\frac{(f_\lambda')^2}{f_\lambda}\,dx \tag{2.49}$$

$$= \int 2\,\frac{(f_1' f_0 - f_0' f_1)^2}{f_\lambda^3}\,dx \tag{2.50}$$

$$\geq 0 \tag{2.51}$$

where (2.49) follows from Fatou's lemma (Halmos, p.113-114), and (2.50) from (2.43). Thus, from (2.42),

$$\int 2\,\frac{(f_1' f_0 - f_0' f_1)^2}{f_\lambda^3}\,dx = 0 \tag{2.52}$$

It follows that

$$\frac{f_1'}{f_1} = \frac{f_0'}{f_0} \tag{2.53}$$

a.e.

Integrating (2.53), since the support of f_0 is convex and therefore connected, it follows after exponentiation that

$$f_1 = \alpha f_0 \tag{2.54}$$

for some constant α. But from (2.3),

$$I(P_1) = \int \left[\frac{f_1'(x)}{f_1(x)}\right]^2 f_1(x)\,dx \tag{2.55}$$

$$= \int \left[\frac{\alpha f_0'(x)}{\alpha f_0(x)}\right]^2 \alpha f_0(x)\,dx \tag{2.56}$$

$$= \alpha I(P_0) \tag{2.57}$$

whence, since $I(P_0)$ and $I(P_1)$ are both minima and hence equal by hypothesis, $\alpha = 1$, and uniqueness is proved. (See Huber, 1964, pp.86-90; 1969, pp.81-85; 1981, pp.79-81.) ■

Theorem 2.2 proves that under suitable conditions, there is a distribution in **P** that minimizes the Fisher information. The question remains as to whether or not an estimator can be found that *achieves* the Cramér-Rao lower bound, i.e. an estimator whose asymptotic variance is the inverse of the Fisher information. It is well known that under suitable conditions, the *maximum likelihood estimator* achieves this bound. Specifically, if it is *consistent*, then it is *asymptotically efficient*, i.e. it is *asymptotically normal* with mean equal to the true parameter and variance equal to the inverse of the Fisher information. (See for instance Akahira and Takeuchi, 1981, p.58). Le Cam (1953) discusses the history of the maximum likelihood estimator as well as issues relating to its consistency and asymptotic efficiency (see also Le Cam, 1956). Wald (1949) provides a proof for consistency, which Huber (1967) modifies to hold under weaker conditions; furthermore, Huber also proves asymptotic normality under these weaker conditions. Many of the relevent regularity conditions (though not those of Huber) are further discussed in Le Cam (1970). Proofs of these results are involved and will be ommitted here.

This suggests that the maximum likelihood estimator based upon the least favorable distribution in a given set of distributions may yield the best robust estimator for that set. This is not immediate, however. It can be shown to hold -- at least for some cases -- by explicit verification of the *saddle point condition:* i.e. given a set **C** of allowable estimators and a set **P** of distributions, both defined on $(\mathbf{X}, \mathbf{B})$, and a gain function $J : \mathbf{C} \times \mathbf{P} \rightarrow \mathbf{R}$ to be maximized over **C** and minimized over **P**, the pair consisting of the estimator ψ_0 and distribution P_0 are such that

$$J(\psi, P_0) \leq J(\psi_0, P_0) \leq J(\psi_0, P) \qquad (2.58)$$

for all $\psi \in \mathbf{C}$ and all $P \in \mathbf{P}$. In other words, the pair (ψ_0, P_0) is a solution to the minimax problem. It must be noted that although such a saddle point solution yields the optimal robust estimator, the converse does not necessarily hold -- i.e. a least favorable distribution and the corresponding optimal estimator do not necessarily constitute a saddle point solution. Theorem 2.3, due to Verdú and Poor (1984), provides sufficient conditions such that every least favorable distribution forms a saddle point with its corresponding optimal estimator. The following definition and lemma are needed:

Definition 2.3 Given the minimax problem defined by the sets **C** and **P** and the function J, $(\psi_L, P_L) \in \mathbf{C} \times \mathbf{P}$ is a *regular pair* if and only if for every P such that $P_\lambda := (1-\lambda) P_L + \lambda P \in \mathbf{P}$, $0 \leq \lambda \leq 1$,

$$J(\psi^*(P_\lambda), P_\lambda) - J(\psi_L, P_\lambda) = o(\lambda) \qquad (2.59)$$

where $\psi^*(P) \in \mathbf{C}$ denotes the optimal estimator for the distribution $P \in \mathbf{P}$.

Lemma 2.3 Let $g : [0,1] \rightarrow \mathbf{R}$ be a convex function. Then, $g(0) \leq g(\lambda)$ for every $\lambda \in [0,1]$ if and only if

$$0 \leq \lim_{\lambda \downarrow 0} \frac{1}{\lambda} [\, g(\lambda) - g(0) \,]. \qquad (2.60)$$

Proof From the definition of convexity,

$$g(\,\delta x + (1-\delta) y\,) \;\le\; \delta\, g(x) \;+\; (1-\delta)\, g(y) \tag{2.61}$$

where $0 \le \delta \le 1$. Let $\{\lambda_n\}$ be a strictly decreasing sequence, $0 \le \lambda_n \le 1$ for all n. Substituting $x = \lambda_n$, $y = 0$, and $\delta = \lambda_{n+1} / \lambda_n$ in (2.61), it follows that

$$\frac{g(\lambda_{n+1}) - g(0)}{\lambda_{n+1}} \;\le\; \frac{g(\lambda_n) - g(0)}{\lambda_n} \tag{2.62}$$

for all n. In particular, this implies that

$$\frac{g(\lambda) - g(0)}{\lambda} \tag{2.63}$$

is an increasing function of λ, and its limit as $\lambda \downarrow 0$ exists. Suppose $g(0) \le g(\lambda)$ for every $\lambda \in [0,1]$. Then, (2.63) is non-negative, and therefore so is the limit. Conversely, suppose that (2.60) holds. Since (2.63) decreases with decreasing λ, it follows that for every $\lambda \in [0,1]$,

$$\frac{1}{\lambda}\;[\,g(\lambda) - g(0)\,] \;\ge\; \lim_{\lambda' \downarrow 0}\; \frac{1}{\lambda'}\;[\,g(\lambda') - g(0)\,] \tag{2.64}$$

$$\ge\; 0 \tag{2.65}$$

by hypothesis, so that $g(0) \le g(\lambda)$ for every $\lambda \in [0,1]$, and the lemma is proved. ■

Theorem 2.3 Consider the minimax problem defined by the sets **C** and **P** and the function J. If **P** is convex, and if $J(\,\psi, P\,)$ is convex on **P** for every $\psi \in$ **C**, then the following two statements are equivalent:

(i) P_0 is a least favorable distribution

(ii) The regular pair $(\,\psi_0, P_0\,)$ is a saddle point solution.

Proof Let P_0 be a least favorable distribution, and let $(\,\psi_0, P_0\,)$ be a regular pair. Choose $P = P_0$ in the statement of Definition 2.3, so that $P_\lambda = P_0$ for all λ. Then, letting $\lambda = 0$ in (2.59), regularity implies that

$$J(\,\psi^*(P_0), P_0\,) \;=\; J(\,\psi_0, P_0\,) \tag{2.66}$$

i.e. ψ_0 is the optimal estimator for the distribution P_0 and satisfies the left hand inequality in (2.58). This proves the first half of (i) → (ii); it remains to show that the right hand inequality in (2.58) is also satisfied.

For $P_0, P_1 \in$ **P** and $0 \le \lambda \le 1$, let $P_\lambda := (1-\lambda)\, P_0 + \lambda\, P_1$. Then by definition,

$$J(\,\psi^*(P_\lambda), P_\lambda\,) \;=\; \sup_{\psi \in \mathbf{C}}\; J(\,\psi, P_\lambda\,) \tag{2.67}$$

$$\le\; \sup_{\psi \in \mathbf{C}}\; [\,(1-\lambda)\, J(\,\psi, P_0\,) \;+\; \lambda\, J(\,\psi, P_1\,)\,] \tag{2.68}$$

$$\le\; (1-\lambda)\; \sup_{\psi \in \mathbf{C}}\; J(\,\psi, P_0\,) \;+\; \lambda\; \sup_{\psi \in \mathbf{C}}\; J(\,\psi, P_1\,) \tag{2.69}$$

$$= (1-\lambda)\, J(\,\psi^*(P_0), P_0\,) \;+\; \lambda\, J(\,\psi^*(P_1), P_1\,) \tag{2.70}$$

where (2.68) follows from the convexity (by hypothesis) of $J(\,\psi, P\,)$ on **P** given any $\psi \in$ **C**. This proves that $J(\,\psi^*(P), P\,)$ is also convex on **P**.

By definition, P_0 is a least favorable distribution if and only if

$$J(\,\psi^*(P_0), P_0\,) \;\le\; J(\,\psi^*(P), P\,) \tag{2.71}$$

for all $P \in \dot{\mathbf{P}}$. In particular, setting $P = P_\lambda$ and using Lemma 2.3, it follows that P_0 is a least favorable distribution if and only if

$$0 \le \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[\, J(\,\psi^*(P_\lambda), P_\lambda\,) \;-\; J(\,\psi^*(P_0), P_0\,)\,\right]. \tag{2.72}$$

Similarly, setting $P = P_\lambda$ and using Lemma 2.3, the right hand inequality in (2.58) holds if and only if

$$0 \le \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[\, J(\,\psi_0, P_\lambda\,) \;-\; J(\,\psi_0, P_0\,)\,\right]. \tag{2.73}$$

Note that

$$J(\,\psi^*(P_\lambda), P_\lambda\,) - J(\,\psi^*(P_0), P_0\,) = J(\,\psi^*(P_\lambda), P_\lambda\,) - J(\,\psi_0, P_0\,) \tag{2.74}$$

$$= J(\,\psi^*(P_\lambda), P_\lambda\,) - J(\,\psi_0, P_\lambda\,)$$

$$+\, J(\,\psi_0, P_\lambda\,) - J(\,\psi_0, P_0\,) \tag{2.75}$$

where (2.74) follows from (2.66). Dividing (2.75) by λ and taking the limit as $\lambda \downarrow 0$, and noting that

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \left[\, J(\,\psi^*(P_\lambda), P_\lambda\,) \;-\; J(\,\psi_0, P_\lambda\,)\,\right] = 0 \tag{2.76}$$

from the regularity assumption (2.59), a comparison with (2.72) and (2.73) shows that P_0 is a least favorable distribution if and only if the right hand inequality in (2.58) holds. This establishes that (i) $\longleftrightarrow$ (ii), and completes the proof. (See Verdú and Poor, 1984.) ■

Remark In the present context, the minimax problem consists in finding the estimator minimizing the asymptotic variance for the least favorable distribution. Instead of using the asymptotic variance as cost function, however, it is more convenient to utilize its inverse, the Fisher information, as gain function. This is because, as shown earlier (Lemma 2.2), $I(P)$ is convex on **P**. Thus, if **P** itself is also convex, then the conditions of Theorem 2.3 are satisfied, subject to the regularity condition, eliminating the need to verify that any given pair consisting of a least favorable distribution and the corresponding optimal estimator is a saddle point solution to the minimax problem. It has thus been established that the maximum likelihood estimate corresponding to the distribution minimizing the Fisher information is the optimal robust estimator in the minimax sense, provided that it is contained in **C** and that **P** satisfies certain conditions.

2.2 Derivation of the Estimator

Consider the measure space ($\mathbf{X}, \mathbf{B}, \mu$) defined earlier, and let $\{ x_1, \cdots, x_n \}$ be a sample of independent random variates taking values in $\mathbf{X}$, with a common distribution function P. Let $\mathbf{P} := \{ P_\theta : \theta \in \Theta \}$ be a family of probability measures on ($\mathbf{X}, \mathbf{B}$) such that for all $\theta \in \Theta$, P_θ is absolutely continuous with respect to μ and admits the density f_θ in accordance with the Radon-Nikodym theorem.

Let $\mathbf{X}^n$ be the product of n copies of $\mathbf{X}$, and let $T_n : \mathbf{X}^n \rightarrow \Theta$ be an estimator for the parameter θ. A broad class of such estimators are solutions to maximization problems of the form

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \rho_\theta(x_i). \tag{2.77}$$

For instance, if $\rho_\theta(x) := \log f_\theta(x)$, then the solution of (2.77) is the maximum likelihood estimate; if $\rho_\theta(x) := - \| x - \theta \|^2$, it is the least squares estimate; if $\rho_\theta(x) := - | x - \theta |$, it is the minimum modulus estimate, i.e. the median. Huber calls these *M-estimators*. Since the optimal robust estimator described in Section 2.1 is a maximum likelihood estimator (for a certain appropriately chosen family of distributions), it is also of the form (2.77).

An alternative way of stating (2.77), provided that ρ_θ is differentiable and that Θ is an open set, is

$$\sum_{i=1}^{n} \psi_\theta(x_i) = 0 \tag{2.78}$$

where $\psi_\theta(x) := \alpha \partial \rho_\theta(x) / \partial \theta$ a.e., and α is an arbitrary constant. For the minimax robust estimator, $\alpha = -1$ is chosen for aesthetic reasons (as will become clear below), and

$$\psi_\theta(x) = -\frac{\partial}{\partial \theta} \log f_\theta(x) \tag{2.79}$$

$$= -\frac{f_\theta'(x)}{f_\theta(x)} \tag{2.80}$$

a.s., where f_θ is the density corresponding to the least favorable distribution as described in Section 2.1.

Since the minimax robust estimator is a maximum likelihood estimator, it has the properties known to hold for such estimators in general. Specifically, under rather mild conditions, it is consistent as well as asymptotically efficient. (In this context, it is useful to recall that Huber (1967) proves the consistency and asymptotic normality of the maximum likelihood estimator for the case where the true distribution f^* underlying the observations does not necessarily belong to the parametric family $\{ f_\theta : \theta \in \Theta \}$ defining the estimator. In that case, convergence is to $\theta_0 \in \Theta$ satisfying $E_{f^*}[\log f_\theta(x)] < E_{f^*}[\log f_{\theta_0}(x)]$ for all $\theta \in \Theta$, $\theta \neq \theta_0$.) As discussed later, however, the minimax estimator is not always the most appropriate for any given application, so that Huber's more general results concerning a class of M-estimators are reviewed below.

As before, it is assumed here that $f_\theta(x) := f(x - \theta)$; the same follows, of course, for ψ and ρ.

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. Let the estimator of location $T_n(x_1, \cdots, x_n)$ be the solution of

$$\sum_{i=1}^{n} \psi(x_i - T) = 0 \tag{2.81}$$

(as with equation (2.78)), and let

$$\xi(T) := \int \psi(x-T)\, dP(x) \tag{2.82}$$

denote the expectation of ψ with shift T. It is clear, from (2.81), that $T_n(x_1+c, \cdots, x_n+c) = T_n(x_1, \cdots, x_n) + c$, i.e. T_n is translation invariant. This fact is used in the sequel.

The following lemma, due to Huber, establishes the existence of the expectation in (2.82), and the fact that it crosses zero.

Lemma 2.4 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 Since ρ is convex, ψ is a monotone increasing function of its argument. (This is easy to demonstrate and is analogous to Lemma 2.3 -- see Royden, 1968, pp.108-109.) Thus, $\psi(x-T)$ is monotone *decreasing* in T, so that for $T^* < T$,

$$\psi(x-T^*) - \psi(x-T) \geq 0 \tag{2.83}$$

a.s., and consequently

$$\int \left[\psi(x-T^*) - \psi(x-T) \right] dP(x) \tag{2.84}$$

exists (though it is not necessarily finite). But by hypothesis,

$$\xi(T^*) = \int \psi(x-T^*)\, dP(x) \tag{2.85}$$

also exists and is finite; taken together, (2.84) and (2.85) imply that $\xi(T)$ exists for $T^* < T$ (though it is not necessarily finite). A symmetric argument for $T < T^*$ extends the result to all T. Moreover, since $\psi(x-T)$ is monotone decreasing in T, so is $\xi(T)$.

Decompose ψ into its positive and negative parts, i.e. let $\psi = \psi^+ - \psi^-$, where

$$\psi^+(x) = \max (\psi(x), 0) \tag{2.86}$$

and

$$\psi^-(x) = -\min (\psi(x), 0) \tag{2.87}$$

a.s. For any given x_0, $\psi(x_0-T)$ is monotone decreasing in T and takes both positive and negative values by hypothesis. It follows that for large enough T, $\psi^+(x_0-T) = 0$ and $\psi^-(x_0-T) > 0$, so that

$$\lim_{T\to\infty} \xi(T) = \lim_{T\to\infty} \left[\int \psi^+(x-T)\, dP(x) - \int \psi^-(x-T)\, dP(x) \right] \tag{2.88}$$

$$< 0, \tag{2.89}$$

where both integrals in (2.88) exist since $\xi(T)$ is defined for all T, and the limit exists since both ψ^+ and ψ^- are monotone in T, and therefore so are the integrals. A symmetric argument for $T \to -\infty$ completes the proof. (The proof of existence is suggested in Huber, 1981, p.48; the remainder follows Huber, 1964; 1969, pp.64-65.) ■

The following theorem is due to Huber.

Theorem 2.4 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

$$\int |\psi(x-T)| \, dP(x) < \infty, \tag{2.90}$$

then $T_n(x_1, \cdots, x_n) \to T^*$ 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

$$0 < \int \psi^2(x-T) \, dP(x) < \infty \tag{2.91}$$

is continuous in a neighborhood of T^*, then

$$\mathbf{L}(\sqrt{n} \ (T_n - T^*)) \to \mathbf{N}\left[0, \frac{\int \psi^2(x-T^*) \, dP(x)}{(\xi'(T^*))^2} \right] \tag{2.92}$$

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

Proof Let $\delta > 0$. If (2.90) holds, then the Kolmogorov strong law of large numbers (Loève, 1963, p.239) implies that as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \psi(x_i - (T^* - \delta)) \to \xi(T^* - \delta) < 0 \tag{2.93}$$

a.s. and i.p., where the inequality holds by hypothesis. Similarly, as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \psi(x_i - (T^* + \delta)) \to \xi(T^* + \delta) > 0 \tag{2.94}$$

a.s. and i.p. Since $\xi(T)$ is monotone in T, it follows that for each $\delta > 0$, there is an $n(\delta)$ such that for all $n > n(\delta)$,

$$T^* - \delta < T_n(x_1, \cdots, x_n) < T^* + \delta, \tag{2.95}$$

(recall that $T_n(x_1, \cdots, x_n)$ solves (2.81)), and similarly, as $n \to \infty$,

$$\text{prob}[\, T^* - \delta < T_n(x_1, \cdots, x_n) < T^* + \delta \,] \to 1. \tag{2.96}$$

Letting $\delta \to 0$ proves the first assertion.

Because of translation invariance, it can be assumed with no loss of generality that $T^* = 0$. Since $\psi(x-T)$ is monotone decreasing in T and $T_n(x_1, \cdots, x_n)$ solves (2.81), it follows that $T_n(x_1, \cdots, x_n) < T$ if and only if

$$\sum_{i=1}^{n} \psi(x_i - T) \leq 0. \tag{2.97}$$

a.s., for any given T. Rewriting (2.97) to center the sum on its expectation and bound its variance,

$$n^{-1/2} \sum_{i=1}^{n} \left[\psi(x_i - T) - \xi(T) \right] \leq -n^{1/2}\, \xi(T) \tag{2.98}$$

a.s. (from (2.82)), so that

$$\begin{aligned} &\text{prob}[\sqrt{n}\ T_n(x_1, \cdots, x_n) < T\ ] \\ &\qquad = \text{prob}\left[n^{-1/2} \sum_{i=1}^{n} \left[\psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T) \right] \leq -n^{1/2}\, \xi(n^{-1/2}T) \right]. \end{aligned} \tag{2.99}$$

Thus, showing that the right hand side of (2.99) tends towards a normal distribution would establish the asymptotic normality of $\sqrt{n}\ T_n(x_1, \cdots, x_n)$.

Note first that since $\{ x_i \}$ are independent and identically distributed,

$$\text{var}\left[\psi(x_i - n^{-1/2}T) \,\middle|\, \sum_{k=1}^{i-1} \psi(x_k - n^{-1/2}T) \right] = \text{var}[\ \psi(x_i - n^{-1/2}T)\ ] \tag{2.100}$$

for all i, so that

$$\sum_{i=1}^{n} E\left[\left| \text{var}\left[\psi(x_i - n^{-1/2}T) \,\middle|\, \sum_{k=1}^{i-1} \psi(x_k - n^{-1/2}T) \right] - \text{var}[\ \psi(x_i - n^{-1/2}T)\ ] \right| \right] = 0 \tag{2.101}$$

identically. Moreover, by independence,

$$\text{var}[\ \psi(x_i - n^{-1/2}T)\ ] = \int \psi^2(x - n^{-1/2}T)\, dP(x) - \xi^2(n^{-1/2}T) \tag{2.102}$$

for all i, so that

$$\sum_{i=1}^{n} \text{var}[\ n^{-1/2}\, (\ \psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T)\ )\ ] = \int \psi^2(x - n^{-1/2}T)\, dP(x) - \xi^2(n^{-1/2}T) \tag{2.103}$$

$$< \infty, \tag{2.104}$$

at least for large n (i.e. for $n^{-1/2}T$ near 0), where (2.104) follows from (2.91), the continuity of $\xi(T)$ in a neighborhood of 0, and $\xi(0) = 0$ (by hypothesis). Finally, define

$$A(\,n, \delta, T\,) := \left\{ x :\ n^{-1/2} \left| \psi(x - n^{-1/2}T) - \xi(n^{-1/2}T) \right| \geq \delta \right\} \tag{2.105}$$

for some given $\delta > 0$. Then, by independence,

$$\begin{aligned} &\sum_{i=1}^{n} \int_{A(n,\delta,T)} \left[n^{-1/2} \left[\psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T) \right] \right]^2 dP(x_i) \\ &\qquad = \int_{A(n,\delta,T)} \left[\psi(x - n^{-1/2}T) - \xi(n^{-1/2}T) \right]^2 dP(x) \end{aligned} \tag{2.106}$$

$$\to 0 \tag{2.107}$$

as $n \to \infty$, since, as with (2.103)-(2.104) and from (2.90),

$$\left| \psi(x - n^{-1/2}T) - \xi(n^{-1/2}T) \right| < \infty \tag{2.108}$$

a.s., so that

$$\lim_{n\to\infty} A(n, \delta, T) = \varnothing \tag{2.109}$$

(or possibly a set of measure 0). Thus, Lindeberg's conditions (Loève, 1963, pp.377-378) for asymptotic normality are satisfied, and the right hand side of (2.99) tends towards a normal distribution.

Since $\xi(0) = 0$ and $\xi(T)$ is differentiable in a neighborhood of 0 by hypothesis,

$$\xi(T) = T\,\xi'(0) + O(T^2) \tag{2.110}$$

so that

$$-\,n^{1/2}\,\xi(n^{-1/2}T) = -\,n^{1/2}\,(\,n^{-1/2}T\xi'(0) + O(n^{-1})\,) \tag{2.111}$$

$$\rightarrow\ -T\xi'(0) \tag{2.112}$$

as $n \rightarrow \infty$. Thus, recalling that $\xi(T)$ is strictly monotone (decreasing) in a neighborhood of 0 by hypothesis, so that $\xi'(0) < 0$, the limit of (2.98) can be written as

$$\lim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \frac{\psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T)}{|\xi'(0)|} \le T \tag{2.113}$$

so that, comparing with (2.99),

$$n^{-1/2} \sum_{i=1}^{n} \frac{\psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T)}{|\xi'(0)|} \overset{d}{=} \sqrt{n}\; T_n(\,x_1, \cdots, x_n\,) \tag{2.114}$$

asymptotically. This establishes the asymptotic normality of $\sqrt{n}\ T_n(\,x_1, \cdots, x_n\,)$, and it only remains to derive its limiting variance. Once again by independence,

$$\mathrm{var}\left[\lim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \frac{\psi(x_i - n^{-1/2}T) - \xi(n^{-1/2}T)}{|\xi'(0)|}\right]$$

$$= \mathrm{var}\left[\lim_{n\to\infty} \frac{\psi(x - n^{-1/2}T) - \xi(n^{-1/2}T)}{|\xi'(0)|}\right] \tag{2.115}$$

$$= \mathrm{var}\left[\frac{\psi(x)}{|\xi'(0)|}\right] \tag{2.116}$$

$$= \frac{\int \psi^2(x)\,dP(x)}{(\,\xi'(0)\,)^2} \tag{2.117}$$

where use is made of the fact that $E[\,\psi(x)\,] = \xi(0) = 0$, concluding the proof. (See Huber, 1964; 1969, pp.66-72; also 1981, pp.45-50.) ■

Corollary 2.1 For a given family of symmetric distributions with location parameter θ, let the least favorable density f_θ be such that $I(f_\theta) < \infty$ for all θ, let the corresponding influence-bounding function ψ_θ (given by equation (2.80)) satisfy the conditions of Theorem 2.4, and let $T_n(\,x_1, \cdots, x_n\,)$ be the minimax robust estimator of θ, i.e. the solution of (2.81). If the true underlying distribution is $f_{\theta*}$, then

$$\mathbf{L}(\sqrt{n}\ (T_n - \theta^*)) \rightarrow \mathbf{N}\left[0, \frac{1}{I(f_{\theta^*})}\right] \tag{2.118}$$

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

Proof Because of translation invariance, it can be assumed with no loss of generality that $\theta^* = 0$. Note first that since f_0 is symmetric with respect to 0, it follows that

$$f_0(x) = f_0(-x) \tag{2.119}$$

and similarly,

$$\psi_0(x) = -\frac{f_0'(x)}{f_0(x)} \tag{2.120}$$

$$= \frac{f_0'(-x)}{f_0(-x)} \tag{2.121}$$

$$= -\psi_0(-x). \tag{2.122}$$

Thus,

$$\xi(0) = \int \psi_0(x) f_0(x)\, dx \tag{2.123}$$

$$= \int_{-\infty}^{0} \psi_0(x) f_0(x)\, dx + \int_{0}^{\infty} \psi_0(x) f_0(x)\, dx \tag{2.124}$$

$$= -\int_{\infty}^{0} \psi_0(-x) f_0(-x)\, dx + \int_{0}^{\infty} \psi_0(x) f_0(x)\, dx \tag{2.125}$$

$$= -\int_{0}^{\infty} \psi_0(x) f_0(x)\, dx + \int_{0}^{\infty} \psi_0(x) f_0(x)\, dx \tag{2.126}$$

$$= 0, \tag{2.127}$$

where (2.123) holds by definition, (2.125) follows from a change of variable (replacing x by $-x$), and (2.126) follows from (2.119) and (2.122). It is furthermore easy to show that $\xi(T)$ has a unique root, at 0: suppose $\xi(T_1) = 0$ also, for $T_1 > 0$, and define $T_\lambda = \lambda T_1$ for $\lambda \in [0,1]$. Then,

$$\xi'(0) = \lim_{\lambda \downarrow 0} \frac{\xi(T_\lambda) - \xi(0)}{T_\lambda} \tag{2.128}$$

$$\geq \lim_{\lambda \downarrow 0} \frac{\xi(T_1) - \xi(0)}{T_\lambda} \tag{2.129}$$

$$= 0, \tag{2.130}$$

where (2.129) follows from the fact that $\xi(T)$ is monotone decreasing and $T_1 \geq T_\lambda$ by definition, and (2.130) holds identically since T_1 and 0 are both roots of $\xi(T)$. But this is a contradiction, since $\xi'(0) < 0$ by hypothesis (see Theorem 2.4). Thus, there can be no root $T_1 > 0$. A similar argument for

$T_1 < 0$ establishes unicity. Thus, by Theorem 2.4, $\sqrt{n}\; T_n(x_1, \cdots, x_n)$ is normally distributed with mean 0.

Note next that

$$\xi(T) = \int \psi_0(x-T)\, f_0(x)\, dx \tag{2.131}$$

$$= \int \psi_0(x)\, f_0(x+T)\, dx \tag{2.132}$$

by a change of variable, so that

$$\xi'(T) = \int \psi_0(x)\, f_0{}'(x+T)\, dx. \tag{2.133}$$

Equation (2.133) is justified (*via* the Lebesgue dominated convergence theorem) by the boundedness and differentiability of f_0, implicit in the assumption that $I(f_0) < \infty$; the change of variable in (2.132) thus allows the proof to proceed without making any further assumptions as to the differentiability of ψ_0. Substituting for ψ_0 yields

$$\xi'(T) = -\int \frac{f_0{}'(x)}{f_0(x)}\, f_0{}'(x+T)\, dx \tag{2.134}$$

and thus,

$$\xi'(0) = -I(f_0). \tag{2.135}$$

But

$$\int \psi_0^2(x)\, f_0(x)\, dx = \int \left[-\frac{f_0{}'(x)}{f_0(x)} \right]^2 f_0(x)\, dx \tag{2.136}$$

$$= I(f_0). \tag{2.137}$$

Comparing (2.135) and (2.137) with the asymptotic variance in (2.92) proves the assertion. (See also Huber, 1969, pp.72-73.) ■

Remark The condition in Theorem 2.4 that $\xi(T)$ be differentiable in some neighborhood of T^* is restrictive. It is often not met, in which case weaker statements can be made -- concerning the asymptotic normality of $\sqrt{n}\; \xi(T_n)$, but not that of $\sqrt{n}\; T_n$. Corollary 2.1 shows that the minimax robust estimator is asymptotically efficient under certain conditions -- specifically, when the true underlying distribution is in fact the least favorable one, and has finite Fisher information. On the other hand, small sample theory on the distributional properties of M-estimators is unfortunately very limited; their non-linearity and the rather uncooperative forms of least favorable densities make such results very hard to obtain.

A specific case is now treated in some detail. This case has been investigated in the literature, and forms the basis of a considerable part of what follows.

Definition 2.4 A convenient model of indeterminacy, proposed by Huber (1964), is the *ε-contaminated normal neighborhood*

$$P_\varepsilon := \{ (1-\varepsilon)\Phi + \varepsilon H : H \in \mathbf{S} \}, \tag{2.138}$$

where Φ is the standard normal distribution, **S** is the set of all probability distributions symmetric with respect to the origin (i.e. such that $P(-x) = 1 - P(x)$), and $0 \leq \varepsilon < 1$ is the known fraction of "contamination." The location family (of neighborhoods) generated by P_ε is then defined as

$$\mathbf{P}_\varepsilon := \{ P(x-\theta) : P \in P_\varepsilon, \theta \in \Theta \}. \tag{2.139}$$

The presence of outliers in a nominally normal sample can be modeled here by a distribution H with tails heavier than normal. Note that symmetry ensures the unbiasedness of the maximum likelihood estimator, making the expression for its asymptotic variance considerably simpler as discussed earlier. Although this restriction obviously precludes cases where outliers are grouped on one side of the mean of the nominal ("underlying") distribution, the model is general enough to represent many realistic situations. (The assymetric case has been studied by Jaeckel (1971) and Collins (1976).) Note also that allowing H to be substochastic would ensure vague compactness.

Lemma 2.5 P_ε is a convex set.

Proof Let $P_0, P_1 \in P_\varepsilon$ be two distributions respectively corresponding to H_0 and $H_1 \in$ S. Then, for $\lambda \in [0,1]$,

$$P_\lambda := (1-\lambda)\, P_0 + \lambda\, P_1 \tag{2.140}$$

$$= (1-\lambda)\, [\, (1-\varepsilon)\, \Phi + \varepsilon\, H_0 \,] + \lambda\, [\, (1-\varepsilon)\, \Phi + \varepsilon\, H_1 \,] \tag{2.141}$$

$$= (1-\varepsilon)\, \Phi + \varepsilon\, [\, (1-\lambda)\, H_0 + \lambda\, H_1 \,] \tag{2.142}$$

$$:= (1-\varepsilon)\, \Phi + \varepsilon\, H_\lambda \tag{2.143}$$

$$\in P_\varepsilon \tag{2.144}$$

since, being a weighted sum of two symmetric distributions, with weights summing to unity, $H_\lambda \in$ S also. ∎

From Lemma 2.2, $I(P)$ is a convex function of P, and from Lemma 2.5, P_ε is convex. It follows by Lemma 2.3 that P_0 minimizes $I(P)$ if and only if

$$0 \leq \lim_{\lambda \downarrow 0} \frac{1}{\lambda} \, [\, I(P_\lambda) - I(P_0) \,] \tag{2.145}$$

for all $P_1 \in P_\varepsilon$, where P_λ is defined by (2.140). Equation (2.46) yields

$$\left[\frac{\partial}{\partial \lambda} I(P_\lambda) \right]_{\lambda=0} = \int \left[\frac{\partial}{\partial \lambda} \frac{(f_\lambda')^2}{f_\lambda} \right]_{\lambda=0} dx \tag{2.146}$$

$$= \int \left[2 \left[\frac{f_0'}{f_0} \right] (f_1' - f_0') - \left[\frac{f_0'}{f_0} \right]^2 (f_1 - f_0) \right] dx \tag{2.147}$$

$$\geq 0, \tag{2.148}$$

where f_λ is the Radon-Nikodym derivative of P_λ, $\lambda \in [0,1]$, and $f_\lambda' := \partial P(x-\theta) / \partial\theta$. Moreover, (2.148) follows from (2.145) and must hold for all $P_1 \in P_\varepsilon$. Integrating by parts, and assuming that f_0'' exists at all but a countable number of points,

$$\int \left[\frac{f_0'}{f_0} \right] (f_1' - f_0') \, dx = - \int \frac{f_0 f_0'' - (f_0')^2}{f_0^2} (f_1 - f_0) \, dx \tag{2.149}$$

so that (2.147) may be rewritten as

$$\int \left[2 \left[\frac{f_0'}{f_0} \right] (f_1' - f_0') - \left[\frac{f_0'}{f_0} \right]^2 (f_1 - f_0) \right] dx$$

$$= \int \left[-2 \frac{f_0 f_0'' - (f_0')^2}{f_0^2} - \left[\frac{f_0'}{f_0} \right]^2 \right] (f_1 - f_0) \, dx \tag{2.150}$$

$$= \int \frac{(f_0')^2 - 2 f_0 f_0''}{f_0^2} (f_1 - f_0) \, dx \tag{2.151}$$

$$= -4 \int \frac{(\sqrt{f_0})''}{\sqrt{f_0}} (f_1 - f_0) \, dx \tag{2.152}$$

$$\geq 0 \tag{2.153}$$

where (2.152) can easily be shown to reduce to (2.151), and (2.153) follows from (2.148). Note, furthermore, that the minimizing distribution f_0 can be assumed not to be substochastic, so that

$$\int (f_1 - f_0) \, dx \leq 0 \tag{2.154}$$

and (2.153) holds if

$$\int \left[\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \alpha^2 \right] (f_1 - f_0) \, dx \leq 0 \tag{2.155}$$

for some real-valued constant α.

Huber does not provide details as to how this problem can be solved in the general case, i.e. how to find a P_0 such that (2.153) holds for all P_1 given *any* family of distributions. Rather, in the case of the ε-contaminated normal neighborhood, he draws upon heuristic arguments -- as well as some analogies to the Schrödinger equation for an electron moving in a given potential -- to propose a solution, and proceeds to show that it satisfies (2.153) (1969, pp.82-89; 1981, pp.82-86). That approach is taken below.

The problem essentially consists in finding f_0 minimizing $I(f)$ subject to the constraints

$$f_0(x) \geq (1-\varepsilon) \, \phi(x) \tag{2.156}$$

a.e., and

$$\int f_0(x)\,dx \;=\; 1, \tag{2.157}$$

where ϕ is the Radon-Nikodym derivative of Φ, (2.156) follows from Definition 2.4, and (2.157) once again assumes f_0 is not substochastic (otherwise, equality must be replaced by $\leq$). Given this formulation, it is more than likely that there is *some* region where the inequality constraint (2.156) is active, i.e.

$$f_0(x) \;=\; (1-\varepsilon)\;\phi(x) \tag{2.158}$$

for x in some $X \subset \mathbf{X}$. In that region, $f_1 \geq f_0$ for all $P_1 \in P_\varepsilon$ (compare (2.138) and (2.158)), so that

$$\int_X \left[\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \alpha^2 \right] (\,f_1 - f_0\,)\,dx \;\leq\; 0 \tag{2.159}$$

only if

$$\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \alpha^2 \;\leq\; 0 \tag{2.160}$$

for $x \in X$. It is not hard to verify, by substituting (2.158), that (2.160) holds in some neighborhood of 0, i.e. in the "center" of the distribution. If, on the other hand,

$$f_0(x) \;>\; (1-\varepsilon)\,\phi(x) \tag{2.161}$$

for $x \in \mathbf{X}\backslash X$, then $(\,f_1 - f_0\,)$ may be either positive or negative, depending on f_1. In that case, to ensure that

$$\int_{\mathbf{X}\backslash X} \left[\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \alpha^2 \right] (\,f_1 - f_0\,)\,dx \;\leq\; 0 \tag{2.162}$$

for all $P_1 \in P_\varepsilon$, one may require

$$\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \alpha^2 \;=\; 0. \tag{2.163}$$

for $x \in \mathbf{X}\backslash X$. This, in turn, implies that for $x \in \mathbf{X}\backslash X$ -- the regions away from the center, i.e. the "tails" -- f_0 has exponential form. All these arguments lead to the following least favorable distribution, due to Huber.

Theorem 2.5 For the set P_ε of ε-contaminated normal distributions, the least favorable distribution (i.e. the distribution minimizing the Fisher information $I(f\,)$) is given by

$$f^*(x) \;:=\; \begin{cases} (\,1-\varepsilon\,)\;\phi(k)\;e^{kx+k^2} & x < -k \\ (\,1-\varepsilon\,)\;\phi(x) & -k \leq x \leq k \\ (\,1-\varepsilon\,)\;\phi(k)\;e^{-kx+k^2} & k < x \end{cases} \tag{2.164}$$

where k is related to the fraction of contamination ε by

$$2\left[\frac{\phi(k)}{k} \;-\; \Phi(-k) \right] \;=\; \frac{\varepsilon}{1-\varepsilon}. \tag{2.165}$$

Proof It is first shown that $f^* \in P_\varepsilon$, i.e. there is an $h^* \in \mathbf{S}$ such that

$$f^*(x) \;=\; (1-\varepsilon)\,\phi(x) + \varepsilon\, h^*(x). \tag{2.166}$$

It follows from (2.164) and (2.166) that

$$h^*(x) \;=\; \begin{cases} \dfrac{1-\varepsilon}{\varepsilon}\left[\phi(k)\,e^{kx+k^2} - \phi(x)\right] & x < -k \\ 0 & -k \le x \le k \\ \dfrac{1-\varepsilon}{\varepsilon}\left[\phi(k)\,e^{-kx+k^2} - \phi(x)\right] & k < x \end{cases} \tag{2.167}$$

which is clearly symmetric with respect to the origin. Furthermore, substituting for ϕ, it follows that for $x < -k$,

$$\frac{1-\varepsilon}{\sqrt{2\pi}\,\varepsilon}\left[e^{-k^2/2}\,e^{kx+k^2} - e^{-x^2/2}\right] \;\ge\; 0 \tag{2.168}$$

if and only if

$$e^{-k^2/2}\,e^{kx+k^2} \;\ge\; e^{-x^2/2}, \tag{2.169}$$

or equivalently (taking logarithms and rearranging)

$$\frac{k^2}{2} + kx \;\ge\; -\frac{x^2}{2}, \tag{2.170}$$

or

$$(k + x)^2 \;\ge\; 0, \tag{2.171}$$

which holds for all $x \in \mathbf{X}$, and for $x < -k$ in particular. Since h^* is symmetric with respect to 0, it follows that $0 \le h^*(x)$ for all x. Finally, again by symmetry,

$$\int h^*(x)\,dx \;=\; 2\,\frac{1-\varepsilon}{\varepsilon}\int_{-\infty}^{-k}\left[\phi(k)\,e^{kx+k^2} - \phi(x)\right]dx \tag{2.172}$$

$$=\; 2\,\frac{1-\varepsilon}{\varepsilon}\left[\frac{\phi(k)}{k} - \Phi(-k)\right] \tag{2.173}$$

$$=\; 1 \tag{2.174}$$

from (2.165). Thus, $h^* \in \mathbf{S}$, so that $f^* \in P_\varepsilon$.

Next, it is necessary to show that f^* minimizes $I(f)$, i.e. that it satisfies (2.153). Note first that for $k < |x|$,

$$\frac{(f^{*\prime})^2 - 2\,f^*\,f^{*\prime\prime}}{f^{*2}} \;=\; -\,k^2, \tag{2.175}$$

while for $|x| < k$,

$$\frac{(f^{*\prime})^2 - 2\,f^*\,f^{*\prime\prime}}{f^{*2}} \;=\; 1 - x^2. \tag{2.176}$$

Thus, (2.151) may be rewritten as

$$\int \frac{(f^{*\prime})^2 - 2\, f^* f^{*\prime\prime}}{f^{*2}} (f_1 - f^*)\, dx$$

$$= -k^2 \int_{-\infty}^{-k} (f_1 - f^*)\, dx \;+\; \int_{-k}^{k} (1 - x^2)(f_1 - f^*)\, dx$$

$$- \; k^2 \int_{k}^{\infty} (f_1 - f^*)\, dx \tag{2.177}$$

$$= \int_{-k}^{k} (k^2 + 1 - x^2)(f_1 - f^*)\, dx \;-\; k^2 \int (f_1 - f^*)\, dx \tag{2.178}$$

$$\geq 0 \tag{2.179}$$

for all $f_1 \in P_\varepsilon$. Here, (2.179) follows from the fact that for $|x| < k$, $0 \leq (k^2 + 1 - x^2)$ and $0 \leq (f_1 - f^*)$, and furthermore

$$\int (f_1 - f^*)\, dx \;\leq\; 0 \tag{2.180}$$

since f_1 may be substochastic, but f^* is not. This proves that f^* minimizes $I(f)$.

Finally, it must be shown that f^* is unique. Note first that for $k < |x|$,

$$\left[\frac{f^{*\prime}}{f^*} \right]^2 = k^2, \tag{2.181}$$

while for $|x| < k$,

$$\left[\frac{f^{*\prime}}{f^*} \right]^2 = x^2. \tag{2.182}$$

It follows that

$$I(f^*) = \int \left[\frac{f^{*\prime}}{f^*} \right]^2 f^*\, dx \tag{2.183}$$

$$= \int_{-\infty}^{-k} k^2 f^*\, dx \;+\; \int_{-k}^{k} x^2 f^*\, dx \;+\; \int_{k}^{\infty} k^2 f^*\, dx \tag{2.184}$$

$$= (1 - \varepsilon)\, [\, 2\, \Phi(k) - 1\,] \tag{2.185}$$

whence $0 < I(f^*) < \infty$ for $0 < k$, which is consistent with (2.165). Moreover, the support of f^* is **R**, which is convex. Thus, the conditions of Theorem 2.2 are met, and f^* is unique. This concludes the proof. (Outlines of this proof can be found in Huber, 1969, pp.87-89; 1981, pp.84-85.) ■

Remark It is, in retrospect, somewhat surprising that the least favorable distribution has tails that do not descend *slower* than exponentially. One explanation is provided by the following qualitative argument: if the tails were very heavy, then it would be "too easy" to discriminate outlying observations from those due to the underlying (normal) distribution. Thus, the *least informative* situation occurs when tails are "just heavy enough" for outliers to be most difficult to discriminate.

It follows from (2.80) and (2.164) that

$$\psi_\varepsilon(x) = \begin{cases} -k & x < -k \\ x & -k \le x \le k \\ k & k < x \end{cases} \tag{2.186}$$

a.s., where k is related to ε through (2.165). 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 explains the choice of multiplicative constant discussed earlier. Plots for $f^*(x)$ and $\psi_\varepsilon(x)$ appear in Figure 2.1 (a-b). The function $\psi_\varepsilon(x)$ illustrates well the concept of *bounded influence* estimation. Since wild observations are truncated, no single data point can totally dominate the others; this is in stark contrast to the sample mean, for instance, where any data point may have arbitrarily large influence on the estimate of the parameter. Note also that the function $\psi_\varepsilon(x)$ is closely related to the practice of *Winsorization* (see for example Tukey and Laughlin, 1963), where the j smallest and k largest observations in a sample of size n are replaced by the values of the $j+1$st smallest and $n-k$th largest observations, respectively. While Winsorization does not result in a bounded-influence estimator, its relationship to (2.186) is clear. The main difference between the two approaches is that in the former, truncation does not occur at preset values but is a function of the sample.

Since it is assumed that ρ is differentiable, and therefore continuous, integrating (2.186) yields (within an additive constant)

$$\rho_\varepsilon(x) = \begin{cases} -kx - \dfrac{k^2}{2} & x < -k \\ \dfrac{1}{2}x^2 & -k \le x \le k \\ kx - \dfrac{k^2}{2} & k < x \end{cases} \tag{2.187}$$

a.s. In other words, it is quadratic in the center and linear in the tails. It follows that the estimator defined by (2.81) with $\psi_\varepsilon(x)$ given by (2.186) (or equivalently by (2.77) with ρ_ε given by (2.187)) represents in some sense a continuum between the sample mean and the sample median. As $\varepsilon \to 0$, (2.165) 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).

Assume that the true distribution P_{θ^*} belongs to the location family generated by the ε-contaminated normal neighborhood, i.e. $P_{\theta^*} \in \mathbf{P}_\varepsilon$; because of translation invariance, it can be assumed with no loss of generality that $\theta^* = 0$. It is clear that since ψ_ε is odd and P_0 is symmetric, $\xi(0) = 0$ (see equations (2.123)-(2.127)), so that $\xi(T)$ exists for all T by Lemma 2.4. Moreover, from (2.82),

$$\xi(T) = \int_{-\infty}^{0} \psi_\varepsilon(x-T)\, dP_0(x) + \int_{0}^{\infty} \psi_\varepsilon(x-T)\, dP_0(x) \tag{2.188}$$

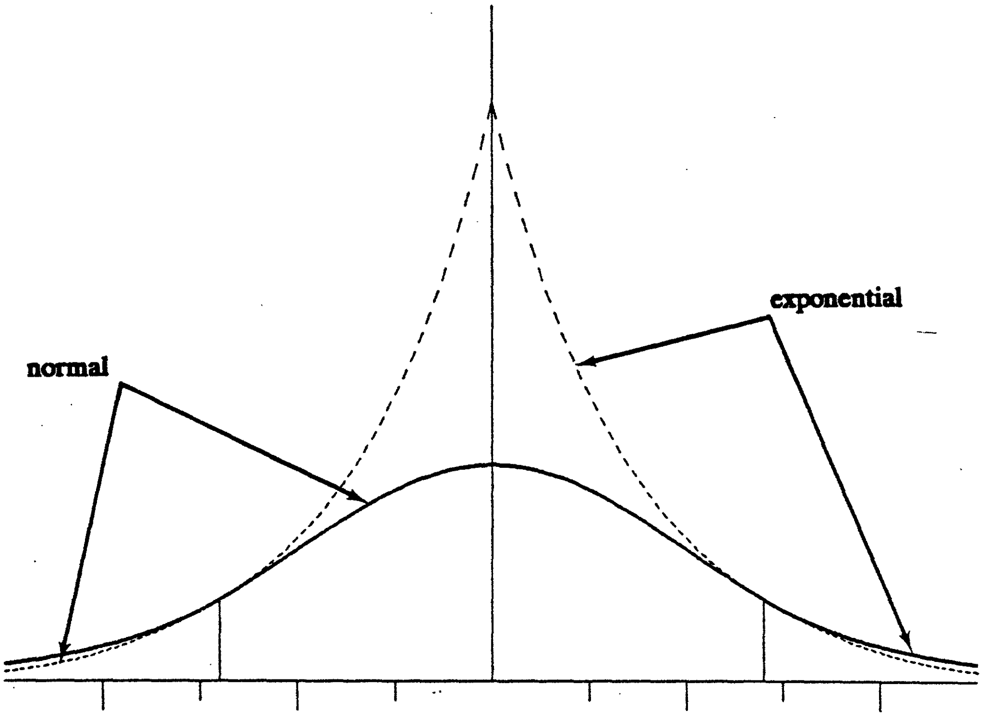

Figure 2.1 (a) Least favorable distribution for ϵ-contaminated normal family.

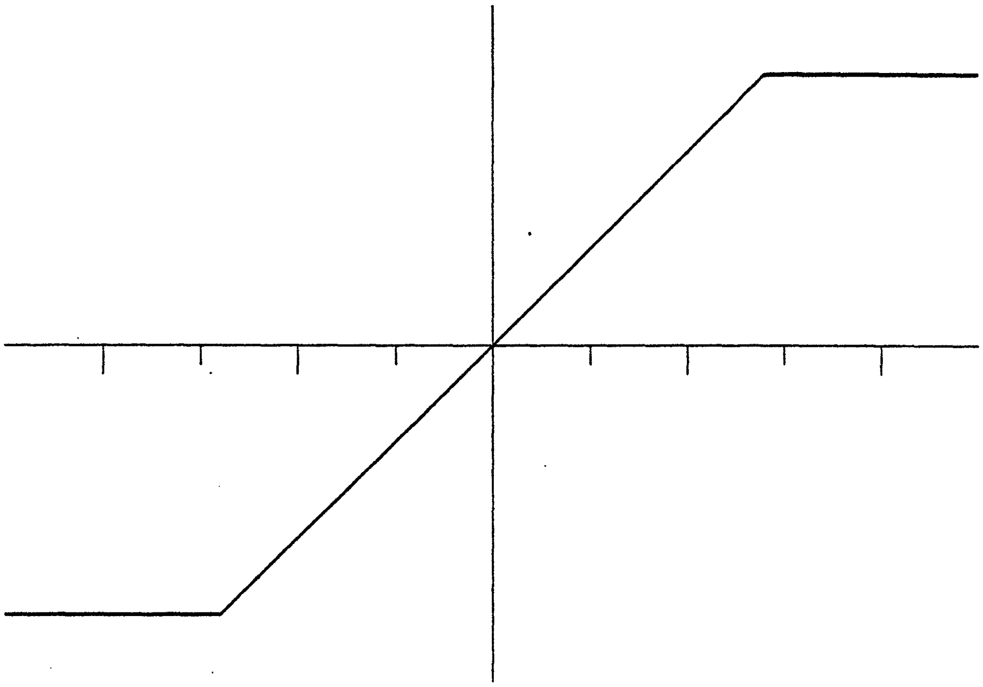

Figure 2.1 (b) ψ-function for ϵ-contaminated normal family.

$$= \int_0^\infty \psi_\varepsilon(-x-T)\, dP_0(-x) \;+\; \int_0^\infty \psi_\varepsilon(x-T)\, dP_0(x) \tag{2.189}$$

$$= \int_0^\infty \left[\psi_\varepsilon(x-T) - \psi_\varepsilon(x+T) \right] dP_0(x) \tag{2.190}$$

where (2.189) follows from a change of variable (replacing x by $-x$), and (2.190) follows from (2.122) and the symmetry of P_0. Thus, $\xi(T) > 0$ for any P_0 if and only if

$$\psi_\varepsilon(\, x - T \,) \;>\; \psi_\varepsilon(\, x + T \,) \tag{2.191}$$

a.s. for $x \in [0,\infty)$. In turn, this holds (provided P_0 has nonzero mass on $[-k,+k]$, which is always true for $P_0 \in \mathbf{P}_\varepsilon$) if and only if

$$x - T \;>\; x + T \tag{2.192}$$

a.s., or

$$T \;<\; 0, \tag{2.193}$$

since $\psi_\varepsilon(x)$ is strictly monotone for $x \in [-k,k]$, from (2.186). A similar argument demonstrates that $\xi(T) < 0$ if and only if $T > 0$. Finally, since $-k \le \psi_\varepsilon \le k$ a.s. from (2.186), it follows that

$$\int \,|\,\psi_\varepsilon(x-T)\,|\; dP_0(x) \;\le\; \int k\; dP_0(x) \tag{2.194}$$

$$\le\; k \tag{2.195}$$

(since P_0 may be substochastic), which is finite for $\varepsilon > 0$. Thus, the first set of conditions in Theorem 2.4 are satisfied, so that the estimator $T_n(\, x_1, \;\cdots, x_n \,)$ solving (2.81) with $\psi = \psi_\varepsilon$ a.s. is consistent.

As stated earlier (in the proof of Corollary 2.1), $\xi(T)$ is continuous and differentiable if $I(P_0) < \infty$; moreover, if P_0 has nonzero mass on $[-k,+k]$, then $\xi(T)$ is strictly monotone in a neighborhood of 0; finally, as before,

$$\int \psi_\varepsilon^2(x-T)\, dP_0(x) \;\le\; k^2, \tag{2.196}$$

which is finite for $\varepsilon > 0$. Thus, the second set of conditions in Theorem 2.4 are also satisfied, and $T_n(\, x_1, \;\cdots, x_n \,)$ is asymptotically normal.

Of course the ε-contaminated normal neighborhood of distributions is only one possible model of indeterminacy. Another proposed model is the ε-*normal* neighborhood, containing distributions whose Kolmogorov distance to the normal distribution is at most ε. More formally, this neighborhood is given by

$$P_\varepsilon{}' \;:=\; \{\; P : \sup_{x \in \mathbf{X}} \;|\; P(x) - \Phi(x) \;|\; \le \varepsilon,\;\; P \in \mathbf{S} \;\}, \tag{2.197}$$

and was investigated by Huber (1964; 1969, pp.89-90; 1981, pp.86-90) as well as Sacks and Ylvisaker (1972), who also analyzed the neighborhood

$$P_{A,p} \;:=\; \{\; P : \int_{-A}^{+A} dP(x) \ge p,\;\; P \in \mathbf{S} \;\}. \tag{2.198}$$

While neither (2.197) nor (2.198) is of particular interest to the present application, which deals primarily with robustness in the presence of heavy-tailed noise, these examples do point to the fact that the choice of a distributional family is rather *ad hoc*. Such arbitrariness, however, seems unavoidable in view of the fact that incomplete or inaccurate information lies at the very core of the robust estimation problem.

There is a certain correspondence between families of distributions and their least favorable members, and, by corollary, between families of distributions and their minimax ψ-functions. (See Poljak and Tsypkin, 1978, 1980.) One can therefore speak of a certain duality between the choice of a distribution family and the selection, on the basis of experience and judgement, of a ψ-function. In other words, it may be of interest to investigate the properties of robust estimators designed with specific influence-bounding functions ψ in mind. For instance, a continuously differentiable (smooth) approximation for the general form of $\psi_\varepsilon(x)$ is the function corresponding to the logistic distribution

$$P(x) = \frac{1}{1+e^{-x}}, \tag{2.199}$$

given by

$$\psi(x) = \frac{1-e^{-x}}{1+e^{-x}}. \tag{2.200}$$

a.s. While this function is not necessarily optimal in the minimax sense, it has the advantage of not containing "corners," which may cause numerical difficulties for some iterative techniques. Thus, the relationship between ψ, ρ, and f is worth exploring in greater detail. Integrating (2.80) yields

$$f(x) \propto e^{-\int \psi(x)\, dx} \tag{2.201}$$

$$= e^{-\rho(x)}, \tag{2.202}$$

a.s., where the proportionality constant is chosen so as to give f unit mass. Several researchers have investigated choices of ψ, and a large number of curves are pictured in Andrews *et al.* (1972, pp.96-101); some are critically reviewed by Rey (1983, pp.100-116). Clearly, for f to be a proper distribution function, ψ must satisfy certain conditions -- e.g. $\rho(x) \to \infty$ as $x \to \pm\infty$. Yet, there are instances where intuition suggests such properties *should* be violated. For instance, if it is known with certainty (say, because of a physical impossibility) that very large observations contain no information whatsoever, then it might be more reasonable to entirely discard rather than merely truncate them. This would call for *redescending* ψ-functions, and loosely corresponds to *trimming* -- where, however, the censoring fraction is not preset but depends on the sample (see for example Tukey and Laughlin, 1963; also Prescott, 1978). Note that non-monotone ψ-functions (i.e. non-convex ρ-functions) do not satisfy the conditions of Lemma 2.4 and Theorem 2.4, and the theory is much less developed for estimators based on them.

Numerous redescending ψ-functions have been proposed. These include Hampel's piecewise linear function (Hampel, 1974; Andrews *et al.*, 1972, p.14), Andrews' sine wave (Andrews, 1974; Andrews *et al.*, 1972, p.15), and Tukey's biweight (Mosteller and Tukey, 1977, p.353; Gross, 1977). The fact that these ψ-functions do not correspond to the least favorable member of any given family of distributions -- indeed that substitution in (2.201) does not even yield a proper density -- diminishes the theoretical justification for this methodology. Nevertheless, Agee, Turner and Gomez (1979) have somewhat formalized this approach by terming the expression in (2.202) a *pseudo-density* and deriving *a posteriori* "densities" based on it. For his part, Huber (1981, p.100-102; see also Collins, 1976) retains the minimax approach by solving the original problem subject to the additional constraint that $\psi(x) = 0$ for $c < |x|$, where the cutoff parameter c is arbitrary. He also observes that an important issue to consider in designing influence-bounding functions is that they must not redescend too steeply: otherwise, the estimate would be very sensitive to small changes in those observations lying in the interval where the function redescends, violating a fundamental tenet of robustness. This suggests that the simple (and frequently used) practice of discarding observations that are "too large," i.e. using

$$\psi(x) = \begin{cases} x & |x| \le c \\ 0 & c < |x| \end{cases} \qquad (2.203)$$

is unwise from the standpoint of robustness, as discussed in Section 1.2.

3. Robust Recursive Estimation of a Deterministic Parameter

As before, let $\{ x_1, \cdots, x_n \}$ be a sample of independent random variates with a common distribution function P. Define

$$\beta_n(x_1, \cdots, x_n; T) := \sum_{i=1}^{n} \psi(x_i - T) \tag{3.1}$$

and recall that the estimator $T_n(x_1, \cdots, x_n)$ is defined as the root of (3.1), i.e.

$$\beta_n(x_1, \cdots, x_n; T_n(x_1, \cdots, x_n)) = 0. \tag{3.2}$$

Since the estimator is consistent (provided certain conditions are satisfied), successive solutions of (3.2) for increasing n tend towards the true value of the location parameter almost surely and in probability, as shown in Section 2.2. Since (3.2) is nonlinear, however, its solution for any given n and $\{ x_1, \cdots, x_n \}$ necessitates some kind of iterative procedure (Huber, 1972). For instance, the Newton-Raphson method is of the form

$$T_n^{(k+1)} = T_n^{(k)} - \frac{\beta_n(x_1, \cdots, x_n; T_n^{(k)})}{\beta_n'(x_1, \cdots, x_n; T_n^{(k)})} \tag{3.3}$$

for $k = 0, 1, \cdots$ and some arbitrary $T_n^{(0)}$ -- an intelligent choice might be the median of $\{ x_1, \cdots, x_n \}$. It is assumed in (3.3) that β_n is differentiable, and furthermore that its slope only vanishes at the root, if at all; since such is not the case, for example, for $\psi = \psi_\varepsilon$ a.s., some safeguards would have to be provided to deal with corners as well as flat extremities. This difficulty aside, however, it is well known that recursions of the form (3.3) converge quadratically near the root (e.g. see Dahlquist and Björk, 1974, pp.222-224). For given $\{ x_1, \cdots, x_n \}$, the process is entirely deterministic, and so long as β_n is relatively well-behaved and $T_n^{(0)}$ is reasonable, the correct solution is virtually assured.

Nevertheless, there are some disadvantages to this kind of "batch" processing -- i.e. to solving (3.2) over and over again each time a new observation x_{n+1} becomes available. On the one hand, this procedure involves the solution of increasingly complex nonlinear equations: recall that ψ is generally nonlinear, so that the sum β_n of n variously shifted ψ-functions gets more and more complicated to handle. On the other, it requires the availability of *all* past observations at all times, a potentially serious memory problem for even moderately high sampling rates. Thus, despite recent advances in computer technology, it appears highly desirable to formulate a sequence of estimators recursively updated by a function of only the most recent observation. This can be achieved with the Robbins-Monro *stochastic approximation* procedure (for general reviews, see for instance Wasan, 1969, pp.8-35; Nevel'son and Has'minskii, 1973, pp.79-83, 88-94; Kushner and Clark, 1978, pp.19-47), first proposed in the context of robust estimation by Martin (1972), Martin and Masreliez (1975), Nevel'son (1975), and Price and Vandelinde (1979). See also Englund, Holst, and Ruppert (1988), who investigate the colored noise case.

3.1 The Method of Stochastic Approximation

Suppose $\xi(T^*) = 0$, where $\xi(T) = E_P[\ \psi(\ x - T\ )\ ]$ as before, and consider the recursion

$$T^R_{n+1} \ = \ T^R_n \ + \ a_n \ \psi(\ x_n - T^R_n\ ), \tag{3.4}$$

where $n = 1, 2, \cdots$, $\{a_n\}$ is a given real-valued sequence, and T^R_1 is an arbitrary (possibly random) starting point. The problem, first posed in a more general setting by Robbins and Monro (1951), is to determine conditions under which $T^R_n \to T^*$ as $n \to \infty$. Note that while the correction term in (3.3) approaches zero (under suitable regularity conditions) as $T^{(k)}_n \to T_n(\ x_1, \ \cdots, x_n\ )$, an analogous statement does not necessarily hold for (3.4): since the value of $\psi(\ x_n - T^R_n\ )$ is random, it is necessary for $\{a_n\}$ to obey certain conditions in order to ensure convergence. Specifically, a_n must tend towards zero at a rate sufficient for the error variance to vanish asymptotically; yet, it must not reach zero for $n < \infty$, since it must be able to compensate for any and all random perturbations due to the $\{x_n\}$ -- 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, 1984, p.34).

The results presented below draw upon a considerable body of literature, where increasingly general conclusions are obtained under weaker and weaker conditions (see e.g. Derman, 1956; Schmetterer, 1961). In their landmark paper, Robbins and Monro prove the mean-square convergence (and hence, the convergence in probability) of recursions of the form (3.4) by assuming that the observation is bounded in probability -- i.e. (in the present case) that there exists an $\alpha < \infty$ such that

$$\text{prob}\Big[\ |\ \psi(\ x\ - T\ )\ | \le \alpha\ \Big] \ = \ 1. \tag{3.5}$$

Kallianpur (1954) also assumes (3.5) to derive estimates for the order of magnitude of the error variance $E\ [\ (\ T^R_n - T^*\ )^2\ ]$. Although this condition is satisfied in the case of bounded-influence estimators (e.g. using ψ_ε, for which $\alpha = k$), it is in general too restrictive; in particular, it is violated by ψ-functions which reduce, but do not necessarily bound, the influence of large x Wolfowitz (1952) proves mean-square convergence by assuming that there exists an $\alpha < \infty$ such that $|\xi(T)| \le \alpha$ for all T, and a $\sigma^2 < \infty$ such that

$$E_P\Big[\ (\ \psi(x-T) - \xi(T)\ )^2\ \Big] \ \le \ \sigma^2 \tag{3.6}$$

for all T. While bounded variance is also assumed in deriving the asymptotic distribution of the M-estimator (see Theorem 2.4), the bound on $\xi(T)$ is once again violated by certain robust or near-robust estimators. A further weakened condition is provided by Blum (1954a), who assumes -- besides (3.6) -- that there exist suitable $0 \le \alpha_1 < \infty$ and $0 \le \alpha_2 < \infty$ such that

$$|\ \xi(T)\ | \ \le \ \alpha_1 \ + \ \alpha_2\ |\ T - T^*\ | \tag{3.7}$$

for all T. Moreover, he is able to prove convergence with probability one. Dvoretzky (1956) proves mean-square convergence as well as convergence with probability one for vastly more general situations, but his setup also requires (3.7) in the special case of Robbins-Monro; indeed, he argues that this condition is necessary to prevent estimates from diverging. Wolfowitz (1956) and Derman and Sacks (1959) provide alternative proofs for Dvoretzky's results, and other researchers also assume conditions at least as strong as (3.7). It is worth noting, however, that this condition is not restrictive in the present application: since the objective is to mitigate the influence of large observations, it is hard to conceive of situations where $|\psi(x-T)|$ grows faster than linearly with large values of its argument; thus, $\xi(T)$ may be assumed to obey (3.7) without realistic loss of generality. Nevertheless, this condition is relaxed in Theorem 3.2, where an alternative proof is employed.

Another class of results obtained for recursions of the form (3.4) concerns the behavior of the moments of T_n^R, as well as its asymptotic distribution. The first such results are due to Chung (1954), who not only provides bounds on the former but also shows that they tend towards the moments of a normal distribution. Unfortunately, he assumes that $\xi(T)$ is bounded by straight lines with nonvanishing slopes from both above *and* below, a condition clearly violated in the case of bounded-influence estimators (such as that obtained with ψ_ε). Hodges and Lehmann (1956) are able to weaken that assumption to (3.7), although at the expense of information on the asymptotic moments. Burkholder (1956) defines a broader class of stochastic approximation algorithms of which the Robbins-Monro process is a subclass, and proves asymptotic normality, as well as obtaining asymptotic confidence intervals free of unknowns, under this weakened condition. Sacks (1958) proves asymptotic normality in both cases by utilizing a central limit theorem rather than Chung's method of moments, and Fabian (1968) does so by obtaining the asymptotic characteristic function.

Some asymptotic results are now stated and proved. Note that generality is not sought beyond that required by the present application. The following lemmas, due to Burkholder and to Chung, are used in the proof of Theorem 3.1.

Lemma 3.1 Let $\{b_n\}$ be a real sequence such that, for some n_0, $b_n \geq 0$ and

$$b_{n+1} \leq b_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} \tag{3.8}$$

for all $n \geq n_0$, where $\{c_n\}$ is a real sequence with

$$\liminf_{n\to\infty} \quad c_n = c > p, \tag{3.9}$$

$\{d_n\}$ is a real sequence with

$$\limsup_{n\to\infty} \quad d_n = d \geq 0, \tag{3.10}$$

and $p > 0$. Then,

$$\limsup_{n\to\infty} \quad n^p\, b_n \leq \frac{d}{c-p}. \tag{3.11}$$

Proof For any δ such that $c-p > \delta > 0$, there is by (3.9) and (3.10) a large enough $n(\delta) \geq n_0$ such that $c_n \geq c-\delta$ and $d_n \leq d+\delta$ for all $n \geq n(\delta)$. Thus, (3.8) may be rewritten as

$$b_{n+1} \leq b_n \left[1 - \frac{c-\delta}{n} \right] + \frac{d+\delta}{n^{p+1}} \tag{3.12}$$

for all $n \geq n(\delta)$, where use is made of the non-negativity of b_n for $n \geq n_0$.

Note now that for any $p > 0$,

$$\frac{1}{(n+1)^p} - \left[1 - \frac{c-\delta}{n}\right]\frac{1}{n^p} = \frac{c-\delta}{n^{p+1}} - \left[\frac{1}{n^p} - \frac{1}{(n+1)^p}\right] \tag{3.13}$$

$$= \frac{c-\delta}{n^{p+1}} - \frac{1}{n^p}\left[1 - \left[1 + \frac{1}{n}\right]^{-p}\right] \tag{3.14}$$

$$= \frac{c-\delta}{n^{p+1}} - \frac{1}{n^p}\left[1 - \left[1 - \frac{p}{n} + O(n^{-2})\right]\right] \tag{3.15}$$

$$= \frac{c-\delta-p}{n^{p+1}} + O(n^{-(p+2)}), \tag{3.16}$$

where the leading term in $O(n^{-(p+2)})$ is positive. Thus, multiplying through by $(d+\delta)/(c-\delta-p)$, one can always find a $c_1 > 0$ such that

$$\frac{d+\delta}{n^{p+1}} \le \frac{d+\delta}{c-\delta-p}\left[\frac{1}{(n+1)^p} - \left[1 - \frac{c-\delta}{n}\right]\frac{1}{n^p}\right] + \frac{c_1}{n^{p+2}}. \tag{3.17}$$

Moreover, choose some p' such that $p < p' < p+1$ and $p' < c-\delta$ (which is possible since $c-\delta > p$ by hypothesis). Then, since $c-\delta-p' > 0$, equation (3.16) (with p replaced by p') implies that there is a large enough $n(p')$ such that

$$\frac{1}{(n+1)^{p'}} - \left[1 - \frac{c-\delta}{n}\right]\frac{1}{n^{p'}} > 0 \tag{3.18}$$

for all $n \ge n(p')$. Thus, one can always find a $c_2 > 0$ such that

$$\frac{c_1}{n^{p+2}} \le c_2\left[\frac{1}{(n+1)^{p'}} - \left[1 - \frac{c-\delta}{n}\right]\frac{1}{n^{p'}}\right] \tag{3.19}$$

for all $n \ge n(p')$, since $p+2 > p'+1$ by hypothesis. Substituting (3.17) and (3.19) into (3.12) and rearranging, it follows that

$$b_{n+1} - \frac{d+\delta}{(c-\delta-p)(n+1)^p} - \frac{c_2}{(n+1)^{p'}} \le \left[1 - \frac{c-\delta}{n}\right]\left[b_n - \frac{d+\delta}{(c-\delta-p)\,n^p} - \frac{c_2}{n^{p'}}\right] \tag{3.20}$$

for all $n \ge n_1 = \max(\, n(\delta), n(p')\,)$. If, for some $n_2 > \max(\, c-\delta, n_1\,)$,

$$b_{n_2} - \frac{d+\delta}{(c-\delta-p)\,n_2^p} - \frac{c_2}{n_2^{p'}} \le 0, \tag{3.21}$$

then (3.21) holds for all $n \ge n_2$, so that

$$b_n \le \frac{d+\delta}{(c-\delta-p)\,n^p} + \frac{c_2}{n^{p'}} \tag{3.22}$$

for all $n \ge n_2$. Otherwise, given some $n_2 > \max(\, c-\delta, n_1\,)$, for any $n \ge n_2$,

$$b_n - \frac{d+\delta}{(c-\delta-p)\,n^p} - \frac{c_2}{n^{p'}} \le \left[b_{n_2} - \frac{d+\delta}{(c-\delta-p)\,n_2^p} - \frac{c_2}{n_2^{p'}}\right]\prod_{j=n_2}^{n-1}\left[1 - \frac{c-\delta}{j}\right] \tag{3.23}$$

$$\leq \left[b_{n_2} - \frac{d+\delta}{(c-\delta-p)\, n_2^p} - \frac{c_2}{n_2^{p'}} \right] \prod_{j=n_2}^{n-1} e^{-(c-\delta)/j} \tag{3.24}$$

$$= \left[b_{n_2} - \frac{d+\delta}{(c-\delta-p)\, n_2^p} - \frac{c_2}{n_2^{p'}} \right] e^{-(c-\delta) \sum_{j=n_2}^{n-1} \frac{1}{j}} \tag{3.25}$$

$$= O(e^{-(c-\delta)\log n}) \tag{3.26}$$

$$= O(n^{-(c-\delta)}), \tag{3.27}$$

where (3.24) follows from inequality 4.2.30, and (3.26) from equation 4.1.32, of Abramowitz and Stegun (n.d.). Combining (3.22) and (3.27), it follows that in either case,

$$b_n \leq \frac{d+\delta}{(c-\delta-p)\, n^p} + O(n^{-p'} + n^{-(c-\delta)}), \tag{3.28}$$

or, since $p' > p$ and $c - \delta > p$ by hypothesis,

$$\limsup_{n\to\infty} \; n^p\, b_n \leq \frac{d+\delta}{c-\delta-p}, \tag{3.29}$$

and letting $\delta \downarrow 0$ proves the assertion. (This is a lemma due to Burkholder and inspired by Chung. The proof follows Wasan, 1969, pp.175-178, and Chung, 1954.) ■

Lemma 3.2 Let $\{b_n\}$ be a real sequence such that, for some n_0, $b_n \geq 0$ and

$$b_{n+1} \geq b_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} \tag{3.30}$$

for all $n \geq n_0$, where $\{c_n\}$ is a real sequence with

$$\limsup_{n\to\infty} \; c_n = c > p, \tag{3.31}$$

$\{d_n\}$ is a real sequence with

$$\liminf_{n\to\infty} \; d_n = d \geq 0, \tag{3.32}$$

and $p > 0$. Then,

$$\liminf_{n\to\infty} \; n^p\, b_n \geq \frac{d}{c-p}. \tag{3.33}$$

Proof The proof is virtually identical to that of Lemma 3.1, and is ommitted. (See Chung, 1954; Wasan, 1969, pp.178-179.) ■

Lemma 3.3 Let $\{b_n\}$ be a real sequence such that, for some n_0, $b_n \geq 0$ and

$$b_{n+1} = b_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} \tag{3.34}$$

for all $n \geq n_0$, where $\{c_n\}$ is a real sequence with

$$\lim_{n \to \infty} c_n = c > p, \tag{3.35}$$

$\{d_n\}$ is a real sequence with

$$\lim_{n \to \infty} d_n = d \geq 0, \tag{3.36}$$

and $p > 0$. Then,

$$\lim_{n \to \infty} n^p \, b_n = \frac{d}{c - p}. \tag{3.37}$$

Proof The proof consists of successive applications of Lemmas 3.1 and 3.2, and is ommitted. (The result is suggested by Burkholder; for a proof, see Wasan, 1969, pp.179-180.) ■

The following theorem is based on the results of Blum and of Burkholder.

Theorem 3.1 Let $\xi(T)$ exist for all T, and let there be a T^* such that $0 < \xi(T)$ for $T < T^*$ and $\xi(T) < 0$ for $T^* < T$. Let (3.6) and (3.7) be satisfied, and let $\{a_n\}$ be a sequence such that $a_n > 0$ for all n,

$$\sum_{n=1}^{\infty} a_n = \infty, \tag{3.38}$$

and

$$\sum_{n=1}^{\infty} a_n^2 < \infty. \tag{3.39}$$

Then, given any $T_1^R < \infty$, $T_n^R \to T^*$ as $n \to \infty$ w.p.1 (i.e. T_n^R is *consistent*).

If, moreover, $\xi(T^*) = 0$, $\xi(T)$ is continuous, differentiable and strictly monotone in a neighborhood of T^* with $|\xi'(T^*)| < \infty$, if

$$\int (\psi(x-T) - \xi(T))^2 \, dP(x) > 0 \tag{3.40}$$

and is continuous and bounded in a neighborhood of T^*,

$$\int | \psi(x-T) - \xi(T) |^r \, dP(x) < \infty \tag{3.41}$$

in a neighborhood of T^* for all natural numbers r, and finally if

$$\lim_{n \to \infty} n \, a_n = a > -\frac{1}{2 \, \xi'(T^*)}, \tag{3.42}$$

then

$$\mathbf{L}(\sqrt{n} \, (T_n^R - T^*)) \to \mathbf{N} \left[0, -\frac{a^2 \int \psi^2(x-T^*) \, dP(x)}{2 \, a \, \xi'(T^*) + 1} \right] \tag{3.43}$$

(i.e. T_n^R is *asymptotically normal*).

Proof Note first that

$$T_{n+1}^R = T_1^R + \sum_{j=1}^{n} a_j \, \psi(x_j - T_j^R) \tag{3.44}$$

from (3.4), so that

$$T_{n+1}^R - \sum_{j=1}^{n} a_j \, \xi(T_j^R)$$

$$= T_{n+1}^R - \sum_{j=1}^{n} a_j \, \psi(x_j - T_j^R) + \sum_{j=1}^{n} a_j \left[\psi(x_j - T_j^R) - \xi(T_j^R) \right] \tag{3.45}$$

$$= T_1^R + \sum_{j=1}^{n} a_j \left[\psi(x_j - T_j^R) - \xi(T_j^R) \right] \tag{3.46}$$

from (3.44). Now:

$$E\left[\psi(x_j - T_j^R) \,\middle|\, \psi(x_1 - T_1^R), \cdots, \psi(x_{j-1} - T_{j-1}^R) \right] = E_P \left[\psi(x_j - T_j^R) \mid T_j^R \right] \tag{3.47}$$

$$= \xi(T_j^R) \tag{3.48}$$

w.p.1, where (3.47) follows from the independence of $\{x_n\}$ and from (3.44), and (3.48) holds by definition. Furthermore,

$$E\left[(\psi(x_j - T_j^R) - \xi(T_j^R))^2 \right] = E\left[E_P [(\psi(x_j - T_j^R) - \xi(T_j^R))^2 \mid T_j^R] \right] \tag{3.49}$$

$$\leq \sigma^2 \tag{3.50}$$

w.p.1, from (3.6). It follows, therefore, that

$$\sum_{j=1}^{\infty} E\left[(a_j \, (\psi(x_j - T_j^R) - \xi(T_j^R)))^2 \right] \leq \sigma^2 \sum_{j=1}^{\infty} a_j^2 \tag{3.51}$$

$$< \infty \tag{3.52}$$

by hypothesis (from (3.39)). Finally, since $\xi(T_j^R)$ is a deterministic function of the random variable T_j^R,

$$\sum_{j=1}^{\infty} E\left[a_j \, (\psi(x_j - T_j^R) - \xi(T_j^R)) \,\middle|\, a_1 \, (\psi(x_1 - T_1^R) - \xi(T_1^R)), \cdots, \right.$$

$$\left. a_{j-1} \, (\psi(x_{j-1} - T_{j-1}^R) - \xi(T_{j-1}^R)) \right]$$

$$= \sum_{j=1}^{\infty} E\left[a_j \, (\psi(x_j - T_j^R) - \xi(T_j^R)) \mid T_j^R \right] \tag{3.53}$$

$$= 0 \tag{3.54}$$

w.p.1, where (3.53) follows from (3.47), and (3.54) holds identically from (3.48). Thus, by a convergence theorem due to Loève (1963, p.387; also 1951), the sum in (3.46) converges w.p.1, and therefore so does the left hand side of (3.45).

Suppose now that

$$\lim_{n\to\infty} T_n^R = \infty \tag{3.55}$$

for some sequence $\{T_n^R\}$. It follows that there is a large enough N such that $T_n^R > T^*$ for all $n \geq N$, and thus

$$a_n\, \xi(T_n^R) < 0 \tag{3.56}$$

for all $n \geq N$, by hypothesis. Then, it must hold that

$$\sum_{j=1}^{\infty} a_j\, \xi(T_j^R) < \infty \tag{3.57}$$

so that

$$\lim_{n\to\infty} \left[T_{n+1}^R - \sum_{j=1}^{n} a_j\, \xi(T_j^R) \right] = \infty \tag{3.58}$$

from (3.55) and (3.57). But since this expression was shown to converge w.p.1, (3.58) is an event of probability zero, whence it follows that

$$\text{prob}\left[\lim_{n\to\infty} T_n^R = \infty \right] = 0. \tag{3.59}$$

A similar argument for $-\infty$ proves that $\{T_n^R\}$ converges to a finite limit, if at all.

Suppose the sequence $\{T_n^R\}$ does not have a limit, i.e.

$$\liminf_{n\to\infty} T_n^R \neq \limsup_{n\to\infty} T_n^R. \tag{3.60}$$

Assume first that

$$\limsup_{n\to\infty} T_n^R > T^*, \tag{3.61}$$

and choose the numbers a and b such that

$$T^* < a < b \tag{3.62}$$

and

$$[a.b] \subset (\liminf_{n\to\infty} T_n^R, \limsup_{n\to\infty} T_n^R). \tag{3.63}$$

Since the left hand side of (3.45) converges to a finite number, it follows that for $m > n$,

$$\lim_{n\to\infty} \left| T_m^R - \sum_{j=1}^{m-1} a_j\, \xi(T_j^R) - T_n^R + \sum_{j=1}^{n-1} a_j\, \xi(T_j^R) \right| = 0. \tag{3.64}$$

Thus, for any $\delta_1 > 0$, there is a large enough $N(\delta_1)$ such that

$$\left| T_m^R - T_n^R - \sum_{j=n}^{m-1} a_j\, \xi(T_j^R) \right| \leq \delta_1 \tag{3.65}$$

provided $N(\delta_1) \leq n < m$. Similarly, it follows from (3.39) that

$$\lim_{n\to\infty} a_n = 0, \tag{3.66}$$

and for any $\delta_2 > 0$, there is a large enough $N(\delta_2)$ such that

$$a_n \le \delta_2 \tag{3.67}$$

provided $N(\delta_2) \le n$. In particular, let

$$\delta_1 = \frac{b-a}{3} \tag{3.68}$$

and

$$\delta_2 = \min\left\{ \frac{1}{3\,\alpha_2}, \frac{b-a}{3\,\alpha_1} \right\}. \tag{3.69}$$

(It is assumed here that α_1 and α_2 are nonzero; this causes no loss of generality, however, since a positive number can always be substituted for zero without affecting the validity of the bound in (3.7).) Choose n and m such that $\max\,(N(\delta_1), N(\delta_2)) \le n < m$, with

$$T_n^R < a, \tag{3.70}$$

$$b < T_m^R, \tag{3.71}$$

and, if $m \ne n+1$, then

$$T_j^R \in [\,a, b\,] \tag{3.72}$$

for all $n < j < m$; this is possible by (3.60). Then, from (3.65) and (3.68),

$$T_m^R - T_n^R \le \frac{b-a}{3} + \sum_{j=n}^{m-1} a_j\, \xi(T_j^R) \tag{3.73}$$

$$\le \frac{b-a}{3} + a_n\, \xi(T_n^R), \tag{3.74}$$

since $\xi(T_j^R) < 0$ for $n < j$ by hypothesis, in view of (3.62) and (3.72). If $T^* \le T_n^R$, then $\xi(T_n^R) \le 0$ also, so that (3.74) yields

$$T_m^R - T_n^R \le \frac{b-a}{3}, \tag{3.75}$$

which contradicts (3.70)-(3.71). If on the other hand $T_n^R < T^*$, then (3.74) yields

$$T_m^R - T_n^R \le \frac{b-a}{3} + a_n\,(\,\alpha_1 + \alpha_2\,|\,T_n^R - T^*\,|\,) \tag{3.76}$$

$$\le \frac{b-a}{3} + a_n\,(\,\alpha_1 + \alpha_2\,(\,T_m^R - T_n^R\,)\,) \tag{3.77}$$

from (3.7) and the fact that $T_n^R < T^* < a < b < T_m^R$ by hypothesis. Equation (3.77) can be rewritten as

$$(\,1 - a_n\,\alpha_2\,)\,(\,T_m^R - T_n^R\,) \le \frac{b-a}{3} + a_n\,\alpha_1, \tag{3.78}$$

or, from (3.67) and (3.69),

$$T_m^R - T_n^R \le b - a \tag{3.79}$$

which once again contradicts (3.70)-(3.71). Thus, (3.61) cannot be true. A similar argument for

$$\limsup_{n\to\infty} \; T_n^R \; \le \; T^* \tag{3.80}$$

also results in a contradiction, proving that

$$\liminf_{n\to\infty} \; T_n^R \; = \; \limsup_{n\to\infty} \; T_n^R, \tag{3.81}$$

i.e. the sequence $\{ T_n^R \}$ converges to a finite number. It remains to show that the limit is T^*.

Assume the contrary, i.e.

$$\lim_{n\to\infty} \; T_n^R \; = \; T_0 \; \ne \; T^* \tag{3.82}$$

for some sequence $\{ T_n^R \}$. Suppose first that $T^* < T_0$. Then, for every $T^* < \delta < T_0$, there is a large enough $n(\delta)$ such that $T_n^R > \delta$ for all $n \ge n(\delta)$. Since $\xi(T)$ is monotone decreasing by Lemma 2.4, it follows that

$$\xi(T_n^R) \; < \; \xi(\delta) \tag{3.83}$$

for all $n \ge n(\delta)$, where use is made of the hypothesis that $\xi(T)$ exists for all T. Thus,

$$\sum_{n=1}^{\infty} a_n \; \xi(T_n^R) \; = \; \sum_{n=1}^{n(\delta)-1} a_n \; \xi(T_n^R) \; + \; \sum_{n=n(\delta)}^{\infty} a_n \; \xi(T_n^R) \tag{3.84}$$

$$\le \; \sum_{n=1}^{n(\delta)-1} a_n \; \xi(T_n^R) \; + \; \xi(\delta) \sum_{n=n(\delta)}^{\infty} a_n \tag{3.85}$$

$$= \; -\infty \tag{3.86}$$

from (3.38), since $\xi(\delta) < 0$ for $\delta > T^*$ by hypothesis. But this is a contradiction: equation (3.82) and the convergence w.p.1 of the left hand side of (3.45) imply that the left hand side of (3.84) converges w.p.1, i.e. that (3.86) is an event of probability zero. A similar argument for $T_0 < T^*$ completes the proof, establishing that $T_n^R \to T^*$ as $n \to \infty$ w.p.1.

The proof of asymptotic normality proceeds in two steps: the result is first proved for a "truncated" version of the recursion (3.4), where $\xi(T)$ is bounded from both above and below by straight lines with finite, nonvanishing slopes; it is subsequently extended to the original recursion subject to the consistency property proved above.

Define

$$S(T) \; := \; \begin{cases} -\dfrac{\xi(T)}{T - T^*} & T \ne T^* \\ -\xi'(T^*) & T = T^* \end{cases} \tag{3.87}$$

which is possible, since $\xi'(T^*)$ exists by hypothesis. Since $\xi(T)$ is strictly monotone descending in a neighborhood of T^*, one can find numbers s_1 and s_2 such that

$$0 \; < \; \frac{1}{2a} \; < \; s_1 \; < \; -\xi'(T^*) \; < \; s_2 \tag{3.88}$$

(where the first inequalities follow from (3.42)) and by the continuity of $\xi(T)$ in a neighborhood of T^*, there exists a $\delta_1 > 0$ such that

$$s_1 \leq S(T) \leq s_2 \tag{3.89}$$

provided $T \in [\, T^* - \delta_1, T^* + \delta_1 \,]$. Similarly, let

$$\sigma^2(T) := \int (\, \psi(x-T) - \xi(T)\,)^2 \, dP(x) \tag{3.90}$$

denote the variance of ψ with shift T. Equation (3.40) and boundedness (by hypothesis) in a neighborhood of T^* imply that one can find numbers σ_1^2 and σ_2^2 such that

$$0 < \sigma_1^2 < \sigma^2(T^*) < \sigma_2^2 < \infty, \tag{3.91}$$

and by the continuity of $\sigma^2(T)$ in a neighborhood of T^*, there exists a $\delta_2 > 0$ such that

$$\sigma_1^2 \leq \sigma^2(T) \leq \sigma_2^2 \tag{3.92}$$

provided $T \in [\, T^* - \delta_2, T^* + \delta_2 \,]$.

By the convergence of T_n^R to T^* w.p.1, there exists for $\delta_3 := \min(\, \delta_1, \delta_2\,)$ and for any $\delta_4 > 0$ a large enough $n(\delta_3, \delta_4)$ such that

$$\text{prob}\,[\; |T_n^R - T^*| < \delta_3 \text{ for all } n \geq n(\delta_3, \delta_4)\;] \;\geq\; 1 - \delta_4. \tag{3.93}$$

In other words, the probability that T_n^R lies in an arbitrary neighborhood of T^* can, in view of consistency, be made arbitrarily large by choosing a large enough n.

Given some $n_1 \geq n(\delta_3, \delta_4)$, define the "truncated" recursion

$$T_{n+1}^{\circ} = T_n^{\circ} + a_n \, \psi_{\circ}(\, x_n - T_n^{\circ}\,) \tag{3.94}$$

for $n \geq n_1$, where

$$T_{n_1}^{\circ} := \begin{cases} T_{n_1}^R & T_{n_1}^R \in [\, T^* - \delta_3, T^* + \delta_3 \,] \\ 0 & \text{otherwise} \end{cases} \tag{3.95}$$

and

$$\psi_{\circ}(x_n - T_n^{\circ}) := \begin{cases} \psi(x_n - (T^* - \delta_3)) - \xi(T^* - \delta_3) + \xi'(T^*)\,(T_n^{\circ} - T^*) & T_n^{\circ} < T^* - \delta_3 \\ \psi(x_n - T_n^{\circ}) & T_n^{\circ} \in [\, T^* - \delta_3, T^* + \delta_3 \,] \\ \psi(x_n - (T^* + \delta_3)) - \xi(T^* + \delta_3) + \xi'(T^*)\,(T_n^{\circ} - T^*) & T_n^{\circ} > T^* + \delta_3 \end{cases} \tag{3.96}$$

a.e. Defining $\xi_{\circ}(T)$, $\sigma_{\circ}^2(T)$, and $S_{\circ}(T)$ analogously to (2.82), (3.90), and (3.87), respectively, it is easy to verify that

$$\xi_{\circ}(T) = \begin{cases} \xi(T) & T \in [\, T^* - \delta_3, T^* + \delta_3 \,] \\ \xi'(T^*)\,(\, T - T^* \,) & \text{otherwise} \end{cases} \tag{3.97}$$

so that

$$s_1 \leq S_{\circ}(T) \leq s_2 \tag{3.98}$$

for all T. Similarly,

$$\sigma_{\circ}^2(T) = \begin{cases} \sigma^2(T^* - \delta_3) & T < T^* - \delta_3 \\ \sigma^2(T) & T \in [\, T^* - \delta_3, T^* + \delta_3 \,] \\ \sigma^2(T^* + \delta_3) & T > T^* + \delta_3 \end{cases} \tag{3.99}$$

so that

$$\sigma_1^2 \leq \sigma_o^2(T) \leq \sigma_2^2 \tag{3.100}$$

for all T. Asymptotic normality is now proved for this bounded case, i.e. for the recursion (3.94)-(3.96).

For economy of notation, define for all non-negative integers r

$$b_n^{(r)} := E[\,(T_n^o - T^*)^r\,] \tag{3.101}$$

and

$$\beta_n^{(r)} := E[\,|\,T_n^o - T^*\,|^r\,], \tag{3.102}$$

and note that their finite existence is guaranteed by (3.41). From (3.94) and (3.101),

$$b_{n+1}^{(r)} = E[\,(T_{n+1}^o - T_n^o + T_n^o - T^*)^r\,] \tag{3.103}$$

$$= E\left[\sum_{k=0}^{r} \begin{bmatrix} r \\ k \end{bmatrix} (T_n^o - T^*)^{r-k} \left[a_n\ \psi_o(x_n - T_n^o)\right]^k\right] \tag{3.104}$$

$$= b_n^{(r)} + \sum_{k=1}^{r} \begin{bmatrix} r \\ k \end{bmatrix} a_n^k\ H_k(r, n) \tag{3.105}$$

where

$$H_k(r, n) := E\left[(T_n^o - T^*)^{r-k}\ \psi_o^k(x_n - T_n^o)\right] \tag{3.106}$$

for all $k \leq r$. Moreover,

$$|\,H_k(r, n)\,| \leq E\left[\,|T_n^o - T^*|^{r-k}\ |\psi_o(x_n - T_n^o)|^k\,\right] \tag{3.107}$$

$$= E\left[\,|T_n^o - T^*|^{r-k}\ E[\,|\psi_o(x_n - T_n^o)|^k \mid T_n^o\,]\,\right] \tag{3.108}$$

$$= E\left[\,|T_n^o - T_{\cdot}^*|^{r-k}\ E[\,|\psi_o(x_n - T_n^o) - \xi_o(T_n^o) + \xi_o(T_n^o)|^k \mid T_n^o\,]\,\right] \tag{3.109}$$

$$\leq 2^{k-1} E\Big[\,|T_n^o - T^*|^{r-k}$$

$$\Big[E[\,|\psi_o(x_n - T_n^o) - \xi_o(T_n^o)|^k \mid T_n^o\,] + |\xi_o(T_n^o)|^k\Big]\,\Big] \tag{3.110}$$

$$= 2^{k-1}\Big(E\left[\,|T_n^o - T^*|^{r-k}\ E[\,|\psi_o(x_n - T_n^o) - \xi_o(T_n^o)|^k \mid T_n^o\,]\,\right]$$

$$+ E\left[\,|T_n^o - T^*|^r\ |S_o(T_n^o)|^k\,\right]\Big), \tag{3.111}$$

w.p.1, where (3.110) follows from the c_r-inequality (Loève, 1963, p.155) and (3.111) holds by definition. It then follows, using (3.41) and (3.98), that

$$|\,H_k(r, n)\,| = O(1)\,\beta_n^{(r-k)} + O(1)\,\beta_n^{(r)} \tag{3.112}$$

for all $k \leq r$.

This result is now used to prove by induction that for each positive integer r there exists a $B_r > 0$ such that

$$\limsup_{n\to\infty} \quad n^{r/2}\,\beta_n^{(r)} \;\le\; B_r. \tag{3.113}$$

Note that since $(\beta_n^{(r)})^{1/r}$ is nondecreasing in r for $r > 0$ (as a consequence of Hölder's inequality; see Loève, 1963, p.156), it suffices to establish (3.113) for all even r. In that case, $\beta_n^{(r)} = b_n^{(r)}$.

Note first that

$$H_1(\,r\,,n\,) \;=\; E\Big[\,E[\,(T_n^\circ - T^*\,)^{r-1}\;\psi_\circ(x_n - T_n^\circ) \mid T_n^\circ\,]\,\Big] \tag{3.114}$$

$$=\; E\Big[\,(T_n^\circ - T^*\,)^{r-1}\,\xi_\circ(T_n^\circ)\,\Big] \tag{3.115}$$

$$=\; E\Big[\,(T_n^\circ - T^*\,)^{r}\,(\,-\,S_\circ(T_n^\circ)\,)\,\Big] \tag{3.116}$$

$$\le\; -\,s_1\,\beta_n^{(r)} \tag{3.117}$$

w.p.1 (provided r is even), where (3.115) and (3.116) hold by definition and (3.117) follows from (3.98). Similarly,

$$H_2(\,r\,,n\,) \;=\; E\Big[\,E[\,(T_n^\circ - T^*\,)^{r-2}\;\psi_\circ^2(x_n - T_n^\circ) \mid T_n^\circ\,]\,\Big] \tag{3.118}$$

$$=\; E\Big[\,(T_n^\circ - T^*\,)^{r-2}\,\Big[\,\sigma_\circ^2(T_n^\circ) + \xi_\circ^2(T_n^\circ)\,\Big]\,\Big] \tag{3.119}$$

$$=\; E\Big[\,(T_n^\circ - T^*\,)^{r-2}\,\sigma_\circ^2(T_n^\circ)\,\Big] + E\Big[\,(T_n^\circ - T^*\,)^{r}\,S_\circ^2(T_n^\circ)\,\Big] \tag{3.120}$$

$$\le\; \sigma_2^2\,\beta_n^{(r-2)} \;+\; s_2^2\,\beta_n^{(r)} \tag{3.121}$$

w.p.1 (provided r is even), where (3.119) and (3.120) hold by definition and (3.121) follows from (3.98) and (3.100). Substituting (3.117) and (3.121) into (3.105) with $r = 2$, it follows that

$$\beta_{n+1}^{(2)} \;\le\; \beta_n^{(2)} \;-\; 2\,a_n\,s_1\,\beta_n^{(2)} \;+\; a_n^2\,(\,\sigma_2^2 + s_2^2\,\beta_n^{(2)}\,) \tag{3.122}$$

$$=\; \beta_n^{(2)}\left[\,1 - \frac{n\,a_n\,(\,2\,s_1 - a_n\,s_2^2\,)}{n}\,\right] \;+\; \frac{n^2\,a_n^2\,\sigma_2^2.}{n^2}. \tag{3.123}$$

Since

$$\liminf_{n\to\infty} \quad n\,a_n\,(\,2\,s_1 - a_n\,s_2^2\,) \;=\; 2\,a\,s_1 \;>\; 1 \tag{3.124}$$

(from (3.42) and (3.88)), and

$$\limsup_{n\to\infty} \quad n^2\,a_n^2\,\sigma_2^2 \;=\; a^2\,\sigma_2^2 \;>\; 0 \tag{3.125}$$

(from (3.42) and (3.91)), it follows by Lemma 3.1 (with $p = 1$) that

$$\limsup_{n\to\infty} \quad n\,\beta_n^{(2)} \;\le\; \frac{a^2\,\sigma_2^2}{2\,a\,s_1 - 1} \;:=\; B_2, \tag{3.126}$$

establishing (3.113) for the case $r = 2$.

Assume now that (3.113) holds for all $k \le r-2$ for some even r, i.e. that

$$\beta_n^{(k)} \;=\; O(n^{-k/2}) \tag{3.127}$$

for all $k \leq r-2$. Since, moreover,

$$a_n = O(n^{-1}) \tag{3.128}$$

from (3.42), it follows that

$$\sum_{k=3}^{r} \begin{bmatrix} r \\ k \end{bmatrix} a_n^k \, H_k(r, n) = \sum_{k=3}^{r} \begin{bmatrix} r \\ k \end{bmatrix} O(n^{-k}) \left[O(1)\, \beta_n^{(r-k)} + O(1)\, \beta_n^{(r)} \right] \tag{3.129}$$

$$= \sum_{k=3}^{r} \begin{bmatrix} r \\ k \end{bmatrix} O(n^{-k})\, O(n^{-(r-k)/2})$$

$$+ \beta_n^{(r)} \sum_{k=3}^{r} \begin{bmatrix} r \\ k \end{bmatrix} O(n^{-k}) \tag{3.130}$$

$$= O(n^{-(r+3)/2}) + O(n^{-3})\, \beta_n^{(r)}, \tag{3.131}$$

where (3.129) follows from (3.112), and (3.130) from (3.127). Substituting (3.117), (3.121), and (3.131) into (3.105) yields

$$\beta_{n+1}^{(r)} \leq \beta_n^{(r)} - r\, a_n\, s_1 \beta_n^{(r)} + \frac{r(r-1)}{2}\, a_n^2 \left[\sigma_2^2\, \beta_n^{(r-2)} + s_2^2\, \beta_n^{(r)} \right]$$

$$+ O(n^{-(r+3)/2}) + O(n^{-3})\, \beta_n^{(r)} \tag{3.132}$$

$$= \beta_n^{(r)} \left[1 - \frac{n\, a_n\, (r\, s_1 - r\,(r-1)\, a_n\, s_2^2 \,/\, 2) + O(n^{-2})}{n} \right]$$

$$+ \frac{r\,(r-1)\,(n^2 a_n^2)\, \sigma_2^2\, (n^{(r-2)/2}\, \beta_n^{(r-2)})\,/\,2 + O(n^{-1/2})}{n^{(r/2)+1}}. \tag{3.133}$$

Since

$$\liminf_{n\to\infty} \; n\, a_n\, (r\, s_1 - r\,(r-1)\, a_n\, s_2^2 \,/\, 2) + O(n^{-2}) = r\, a\, s_1 > \frac{r}{2} \tag{3.134}$$

(from (3.42) and (3.88)), and

$$0 \leq \limsup_{n\to\infty} \; r\,(r-1)\,(n^2 a_n^2)\, \sigma_2^2\, (n^{(r-2)/2} \beta_n^{(r-2)})\,/\,2 + O(n^{-1/2}) \tag{3.135}$$

$$\leq \frac{r(r-1)}{2}\, a^2\, \sigma_2^2\, B_{r-2} \tag{3.136}$$

(where (3.135) follows from (3.91) and the non-negativity of $\beta_n^{(r-2)}$, and (3.136) from (3.42) and (3.113) under the induction hypothesis), it follows by Lemma 3.1 (with $p = r/2$) that

$$\limsup_{n\to\infty} \; n^{r/2}\, \beta_n^{(r)} \leq \frac{(r-1)\, a^2 \sigma_2^2\, B_{r-2}}{2a\, s_1 - 1} := B_r, \tag{3.137}$$

completing the proof of (3.113) for all even r, and hence for all r.

This result is now used to prove by induction that for each positive integer r,

$$\lim_{n\to\infty} \; n^{r/2}\, b_n^{(r)} = \begin{cases} 0 & r \text{ odd} \\ (r-1)(r-3)\cdots 1 \left[- \dfrac{a^2 \sigma^2(T^*)}{2a\, \xi'(T^*) + 1} \right]^{r/2} & r \text{ even} \end{cases} \tag{3.138}$$

Note first that for any $\delta > 0$,

$$\beta_n^{(r)} = E\left[\, |T_n^\circ - T^*|^r \;\middle|\; |T_n^\circ - T^*| \geq \delta \,\right] \text{ prob } \left[\, |T_n^\circ - T^*| \geq \delta \,\right]$$

$$+\; E\left[\, |T_n^\circ - T^*|^r \;\middle|\; |T_n^\circ - T^*| < \delta \,\right] \text{ prob } \left[\, |T_n^\circ - T^*| < \delta \,\right] \tag{3.139}$$

$$\geq E\left[\, |T_n^\circ - T^*|^r \;\middle|\; |T_n^\circ - T^*| \geq \delta \,\right] \text{ prob } \left[\, |T_n^\circ - T^*| \geq \delta \,\right] \tag{3.140}$$

since all terms in (3.139) are non-negative. Choose any $q > 0$. Then,

$$E\left[\, |T_n^\circ - T^*|^r \;\middle|\; |T_n^\circ - T^*| \geq \delta \,\right] \text{ prob } \left[\, |T_n^\circ - T^*| \geq \delta \,\right]$$

$$\leq E\left[\, |T_n^\circ - T^*|^r \left[\frac{|T_n^\circ - T^*|}{\delta} \right]^q \;\middle|\; |T_n^\circ - T^*| \geq \delta \,\right] \text{prob } \left[\, |T_n^\circ - T^*| \geq \delta \,\right] \tag{3.141}$$

$$\leq \delta^{-q}\, \beta_n^{(r+q)} \tag{3.142}$$

$$= O(n^{-(r+q)/2}) \tag{3.143}$$

$$= o(n^{-r/2}), \tag{3.144}$$

where (3.142) follows from (3.140), (3.143) from (3.127), and (3.144) from the fact that $q > 0$ by hypothesis. Alternatively, if δ is replaced by a sequence $\{\delta_n\}$ such that $\delta_n > 0$ for all n and

$$\delta_n^{-1} = o(n^{1/2}), \tag{3.145}$$

then

$$\delta_n^{-q}\, \beta_n^{(r+q)} = o(n^{q/2})\, O(n^{-(r+q)/2}) \tag{3.146}$$

$$= o(n^{-r/2}), \tag{3.147}$$

as before.

Since $\xi(T)$ is continuous in a neighborhood of T^* (by hypothesis), so is $S_o(T)$. It follows that there is a $\delta > 0$ and a $K(\delta) > 0$ such that

$$|\, S_o(T) - S_o(T^*)\, | \;<\; K(\delta) \tag{3.148}$$

if and only if

$$|\, T - T^*\, | \;<\; \delta. \tag{3.149}$$

Choose sequences $\{\delta_n\}$ and $\{K(\delta_n)\}$ such that $0 < \delta_n \leq \delta$ and $K(\delta_n) > 0$ for all n, (3.145) and (3.148)-(3.149) are satisfied, and

$$K(\delta_n) = o(1). \tag{3.150}$$

(This last condition is possible by virtue of the continuity of $S_o(T)$ in a neighborhood of T^*, and by (3.145) which implies that $\delta_n = o(1)$.) Moreover, from (3.87) and (3.97),

$$\begin{aligned} \mid S_o(T) - S_o(T^*) \mid &= \mid S_o(T) + \xi'(T^*) \mid && (3.151) \\ &\leq \mid S_o(T) \mid + \mid \xi'(T^*) \mid && (3.152) \\ &\leq s_2 - \xi'(T^*) && (3.153) \end{aligned}$$

for all T, from (3.98) and the negativity of $\xi'(T^*)$ by hypothesis. It follows that

$$\begin{aligned} &\left| E\left[(T_n^\circ - T^*)^r \left[S_o(T_n^\circ) + \xi'(T^*) \right] \right] \right| \\ &\quad \leq E\left[\mid T_n^\circ - T^* \mid^r \mid S_o(T_n^\circ) + \xi'(T^*) \mid \right] && (3.154) \\ &\quad = E\left[\mid T_n^\circ - T^* \mid^r \mid S_o(T_n^\circ) + \xi'(T^*) \mid \;\Big|\; \mid T_n^\circ - T^* \mid < \delta_n \right] \text{prob} \left[\mid T_n^\circ - T^* \mid < \delta_n \right] \\ &\qquad\quad + E\left[\mid T_n^\circ - T^* \mid^r \mid S_o(T_n^\circ) + \xi'(T^*) \mid \;\Big|\; \mid T_n^\circ - T^* \mid \geq \delta_n \right] \\ &\qquad\quad \text{prob} \left[\mid T_n^\circ - T^* \mid \geq \delta_n \right] && (3.155) \\ &\quad \leq K(\delta_n) \, E\left[\mid T_n^\circ - T^* \mid^r \;\Big|\; \mid T_n^\circ - T^* \mid < \delta_n \right] \text{prob} \left[\mid T_n^\circ - T^* \mid < \delta_n \right] \\ &\qquad\quad + (s_2 - \xi'(T^*)) \, E\left[\mid T_n^\circ - T^* \mid^r \;\Big|\; \mid T_n^\circ - T^* \mid \geq \delta_n \right] \\ &\qquad\quad \text{prob} \left[\mid T_n^\circ - T^* \mid \geq \delta_n \right] && (3.156) \\ &\quad \leq K(\delta_n) \, \beta_n^{(r)} + o(n^{-r/2}) && (3.157) \\ &\quad = o(n^{-r/2}) && (3.158) \end{aligned}$$

where (3.156) follows from (3.148) and (3.153), (3.157) from (3.147), and (3.158) from (3.127) and (3.150). Thus, from (3.116),

$$\begin{aligned} H_1(r, n) &= E\left[(T_n^\circ - T^*)^r (- S_o(T_n^\circ)) \right] && (3.159) \\ &= E\left[(T_n^\circ - T^*)^r \, \xi'(T^*) \right] - E\left[(T_n^\circ - T^*)^r (\xi'(T^*) + S_o(T_n^\circ)) \right] && (3.160) \\ &= \xi'(T^*) \, b_n^{(r)} + o(n^{-r/2}) && (3.161) \end{aligned}$$

from (3.158). Setting $r = 1$ and substituting into (3.105) yields

$$\begin{aligned} b_{n+1}^{(1)} &= b_n^{(1)} + a_n \, H_1(1, n) && (3.162) \\ &= b_n^{(1)} \left[1 + \frac{n \, a_n \, \xi'(T^*)}{n} \right] + a_n \, o(n^{-1/2}), && (3.163) \end{aligned}$$

implying that

$$\mid b_{n+1}^{(1)} \mid \leq \mid b_n^{(1)} \mid \left| 1 + \frac{n \, a_n \, \xi'(T^*)}{n} \right| + \mid a_n \, o(n^{-1/2}) \mid . \tag{3.164}$$

(Note that this step is necessary because $b_n^{(1)}$ may be negative, violating a condition of Lemma 3.3.)

But by (3.42), there is, for any $\xi'(T^*)$ and $\{a_n\}$ satisfying the conditions of this theorem, a large enough N such that

$$1 + a_n \xi'(T^*) > 0 \tag{3.165}$$

for all $n \geq N$. Hence, (3.164) may be rewritten as

$$|b_{n+1}^{(1)}| \leq |b_n^{(1)}| \left[1 + \frac{n\, a_n\, \xi'(T^*)}{n} \right] + \frac{n\, a_n\, o(1)}{n^{3/2}} \tag{3.166}$$

for all $n \geq N$ (where the last term is implicitly positive). Since

$$\liminf_{n \to \infty} \; -n\, a_n\, \xi'(T^*) = -a\, \xi'(T^*) > 1/2 \tag{3.167}$$

(from (3.42)), and

$$\limsup_{n \to \infty} \; n\, a_n\, o(1) = 0 \tag{3.168}$$

(from (3.42)), it follows by Lemma 3.1 (with $p = 1/2$) that

$$\limsup_{n \to \infty} \; n^{1/2}\, |b_n^{(1)}| \leq 0, \tag{3.169}$$

implying that

$$\lim_{n \to \infty} \; n^{1/2} b_n^{(1)} = 0. \tag{3.170}$$

This establishes (3.138) for the case $r = 1$.

In analogous fashion, the continuity of $\sigma^2(T)$ and hence $\sigma_o^2(T)$ in a neighborhood of T^* (by hypothesis) yields that

$$E\left[(T_n^o - T^*)^{r-2}\, \sigma_o^2(T_n^o) \right] = \sigma^2(T^*)\, b_n^{(r-2)} + o(n^{-(r-2)/2}), \tag{3.171}$$

while it follows from an argument similar to (3.154)-(3.161) that

$$E\left[(T_n^o - T^*)^{r}\, S_o^2(T_n^o) \right] = (\xi'(T^*))^2\, b_n^{(r)} + o(n^{-r/2}). \tag{3.172}$$

Thus, (3.120) may be used to obtain

$$H_2(\, r, n\,) = (\xi'(T^*))^2\, b_n^{(r)} + \sigma^2(T^*)\, b_n^{(r-2)} + o(n^{-(r-2)/2}) \tag{3.173}$$

(neglecting the lower order term in (3.172)). Setting $r = 2$, substituting into (3.105) and rearranging yields

$$b_{n+1}^{(2)} = b_n^{(2)} \left[1 + \frac{n\, a_n\, (\, 2\,\xi'(T^*) + a_n\, (\xi'(T^*))^2\,)}{n} \right] + \frac{n\, a_n\, (\, 2\, o(1) + n\, a_n\, (\, \sigma^2(T^*) + o(1)\,)\,)}{n^2}. \tag{3.174}$$

Since

$$\lim_{n \to \infty} \; -n\, a_n\, (\, 2\,\xi'(T^*) + a_n\, (\xi'(T^*))^2\,) = -2\, a\, \xi'(T^*) > 1 \tag{3.175}$$

(from (3.42)), and

$$\lim_{n\to\infty} \; n\, a_n \,(2\,o(1) + n\, a_n \,(\sigma^2(T^*) + o(1))) \;=\; a^2 \sigma^2(T^*) \;>\; 0 \tag{3.176}$$

(from (3.42) and (3.40)), it follows by Lemma 3.3 (with $p = 1$) that

$$\lim_{n\to\infty} \; n\, b_n^{(2)} \;=\; -\,\frac{a^2\sigma^2(T^*)}{2a\,\xi'(T^*) + 1}. \tag{3.177}$$

This proves (3.138) for the case $r = 2$.

Assume now that (3.138) (with r replaced by k) holds for all $k < r$, given some $r > 2$, and note that

$$\sum_{k=3}^{r} \begin{bmatrix} r \\ k \end{bmatrix} \; a_n^k \; H_k(\, r, n \,) \;=\; O(n^{-(r+3)/2}) \tag{3.178}$$

by (3.127) and (3.131). Substituting (3.161), (3.173), and (3.178) into (3.105) yields

$$\begin{aligned} b_{n+1}^{(r)} \;=\; & b_n^{(r)} \;+\; a_n\, r \;\Big[\, \xi'(T^*)\, b_n^{(r)} + o(n^{-r/2}) \;\Big] \\ & + \frac{r\,(r-1)}{2}\; a_n^2 \;\Big[(\xi'(T^*))^2 b_n^{(r)} + \sigma^2(T^*)\, b_n^{(r-2)} + o(n^{-(r-2)/2}) \Big] \\ & + \; O(n^{-(r+3)/2}) \end{aligned} \tag{3.179}$$

$$\begin{aligned} =\; & b_n^{(r)} \left[\, 1 + \frac{n\, a_n\, r\, \xi'(T^*) \,(\, 1 + a_n \,(r-1)\, \xi'(T^*) \,/\, 2 \,)}{n} \,\right] \\ & + \frac{1}{n^{(r/2)+1}} \;\Bigg[\, r\,(n\, a_n)\, o(1) + \frac{r\,(r-1)}{2}\,(n^2 a_n^2) \\ & \qquad \Big[\, \sigma^2(T^*)\,(\, n^{(r-2)/2}\, \beta_n^{(r-2)} \,) + \; o(1) \,\Big] \;+\; O(n^{-1/2}) \;\Bigg]. \end{aligned} \tag{3.180}$$

Since

$$\lim_{n\to\infty} \; -n\, a_n\, r\, \xi'(T^*) \,(\, 1 + a_n \,(r-1)\, \xi'(T^*) \,/\, 2 \,) \;=\; -r\, a\, \xi'(T^*) \;>\; \frac{r}{2} \tag{3.181}$$

(from (3.42)), and

$$\begin{aligned} & \lim_{n\to\infty} \; r\,(n\, a_n)\, o(1) + r\,(r-1)\,(n^2 a_n^2)\,(\, \sigma^2(T^*)\,(\, n^{(r-2)/2}\, \beta_n^{(r-2)}\,) + o(1)\,)\,/\,2 \\ & \quad = \frac{r\,(r-1)}{2}\; a^2 \sigma^2(T^*) \; \lim_{n\to\infty} \;(\, n^{(r-2)/2}\, \beta_n^{(r-2)} \,) \end{aligned} \tag{3.182}$$

$$= \begin{cases} 0 & r \text{ odd} \\ \dfrac{r\,(r-1)}{2}\,(r-3)\cdots 1\; \dfrac{(\, a^2 \sigma^2(T^*)\,)^{\,r/2}}{(\,-2a\,\xi'(T^*) - 1\,)^{(r-2)/2}} & r \text{ even} \end{cases} \tag{3.183}$$

(from (3.42) and (3.138) under the induction hypothesis), where the last term is clearly positive in view of (3.42) and (3.40). Lemma 3.3 (with $p = r/2$) completes the proof of (3.138) for all r.

A comparison of (3.138) with the moments of a normal distribution (e.g. Kendall and Stuart, 1977, vol.1, p.62) reveals that

$$\mathbf{L}(\sqrt{n}\ (T_n^\circ - T^*)\ ) \ \rightarrow\ \mathbf{N}\left[0, -\frac{a^2\, \sigma^2(T^*)}{2\ a\ \xi'(T^*) + 1} \right], \tag{3.184}$$

i.e. that T_n° is asymptotically normal with the parameters given in equation (3.43). It remains to extend this result to the original recursion T_n^R.

From (3.93) and the definition of the truncated recursion (equations (3.94)-(3.96)), it follows that

$$\text{prob}\ [\ T_n^R \neq T_n^\circ\ ]\ \leq\ \delta_4 \tag{3.185}$$

for all $n\ \geq n_1$. Choose any $t \in \mathbf{R}$. Then,

$$[\ T_n^R > t\,, T_n^\circ \leq t\ ]\ \subseteq\ [\ T_n^R \neq T_n^\circ\ ]\,, \tag{3.186}$$

where brackets denote events. It follows that

$$\text{prob}\ [\ T_n^R > t\,, T_n^\circ \leq t\ ]\ \leq\ \text{prob}\ [\ T_n^R \neq T_n^\circ\ ]\,, \tag{3.187}$$

so that

$$\text{prob}\ [\ T_n^R \leq t\ ]\ \geq\ \text{prob}\ [\ T_n^R \leq t,\, T_n^\circ \leq t\ ] \tag{3.188}$$

$$=\ \text{prob}\ [\ T_n^\circ \leq t\ ]\ -\ \text{prob}\ [\ T_n^R > t\,, T_n^\circ \leq t\ ] \tag{3.189}$$

$$\geq\ \text{prob}\ [\ T_n^\circ \leq t\ ]\ -\ \delta_4 \tag{3.190}$$

from (3.185) and (3.186). Thus,

$$\text{prob}\ [\ T_n^\circ \leq t\ ] - \text{prob}\ [\ T_n^R \leq t\ ]\ \leq\ \delta_4. \tag{3.191}$$

By symmetry,

$$\text{prob}\ [\ T_n^R \leq t\ ] - \text{prob}\ [\ T_n^\circ \leq t\ ]\ \leq\ \delta_4 \tag{3.192}$$

also, so that

$$\left|\ \text{prob}\ [\ T_n^R \leq t\ ] - \text{prob}\ [\ T_n^\circ \leq t\ ]\ \right|\ \leq\ \delta_4. \tag{3.193}$$

But from (3.184), there is a large enough $n(\delta_4)$ such that

$$\left|\ \text{prob}\ [\ T_n^\circ \leq t\ ] - \Phi(t)\ \right|\ \leq\ \delta_4 \tag{3.194}$$

for all $n\ \geq n(\varepsilon)$, where Φ denotes the normal distribution in (3.184). Thus,

$$\left|\ \text{prob}\ [\ T_n^R \leq t\ ] - \Phi(t)\ \right|$$

$$=\ \left|\ \text{prob}\ [\ T_n^R \leq t\ ] - \text{prob}\ [\ T_n^\circ \leq t\ ] + \text{prob}\ [\ T_n^\circ \leq t\ ] - \Phi(t)\ \right| \tag{3.195}$$

$$\leq\ \left|\ \text{prob}\ [\ T_n^R \leq t\ ] - \text{prob}\ [\ T_n^\circ \leq t\ ]\ \right|\ +\ \left|\ \text{prob}\ [\ T_n^\circ \leq t\ ] - \Phi(t)\ \right| \tag{3.196}$$

$$\leq\ 2\,\delta_4 \tag{3.197}$$

for all $n \geq n(\delta_4)$, from (3.193) and (3.194). Letting $\delta_4 \downarrow 0$, using (3.90) with $T = T^*$, and noting that $\xi(T^*) = 0$ by hypothesis, completes the proof. (The proof of consistency follows, with some modifications, that of Blum, 1954a; the proof of asymptotic normality uses a truncation argument due to Hodges and Lehmann, and is a special case -- with modifications -- of that of Burkholder, 1956.) ■

Corollary 3.1 Theorem 3.1 holds also if T_1^R is a random variable, provided that

$$E[\,(T_1^R)^r\,] < \infty \tag{3.198}$$

for all natural numbers r, and T_1^R is independent of x_n, $n = 2, 3, \cdots$. If, moreover, T_1^R is a translation-invariant function of x_1, then T_n^R is translation invariant.

Proof The proof of the first part of the corollary follows that of Theorem 3.1 identically. In the proof of consistency, the condition of independence is required in order for equation (3.48) to hold; furthermore, T_1^R must have bounded variance in order for (3.46), and hence the left hand side of (3.45), to converge w.p.1. In the proof of asymptotic normality, (3.198) is required in addition to (3.41) in order to ensure that the expressions in (3.101) and (3.102) exist finitely.

The proof of translation invariance proceeds by induction. By hypothesis,

$$T_1^R(\,x_1+c\,) = T_1^R(\,x_1\,) + c. \tag{3.199}$$

Assume now that $T_n^R(\,x_1, \cdots, x_{n-1}\,)$ is translation invariant for some n; then, from (3.4),

$$T_{n+1}^R(\,x_1+c, \cdots, x_n+c\,) = T_n^R(\,x_1+c, \cdots, x_{n-1}+c\,)$$

$$+\ a_n\ \psi(\,x_n+c - T_n^R(\,x_1+c, \cdots, x_{n-1}+c\,)\,) \tag{3.200}$$

$$= T_n^R(\,x_1, \cdots, x_{n-1}\,) + c$$

$$+\ a_n\ \psi(\,x_n+c - T_n^R(\,x_1, \cdots, x_{n-1}\,) - c\,) \tag{3.201}$$

$$= T_{n+1}^R(\,x_1, \cdots, x_n\,) + c, \tag{3.202}$$

where (3.201) holds by the induction hypothesis, and (3.202) by (3.4), completing the proof. ■

Remark This corollary suggests that -- in the absence of additional information -- a convenient starting point for the recursion might be $T_1^R = x_1$ (see, for instance, Martin, 1972; Martin and Masreliez, 1975; Price and Vandelinde, 1979). This has the advantages of simplicity and translation invariance, in addition to the obvious fact that the observation x_1 will generally be a better estimator of location than an arbitrary constant. On the other hand, this choice implies that one observation can have arbitrarily large influence on T_n^R, $n \geq 1$, so that an outlying x_1 could severely degrade the small sample performance of the estimator. This is contrary to the philosophy of robust estimation, and is therefore not desirable. While a better initial value might be a bounded version of the observation, such as $\psi(x_1)$, this choice does not generally lead to translation invariance. Alternatively, the first few observations could be utilized to obtain a Huber minimax robust M-estimator, which could then be used as the initial value T_1^R of the recursive minimax robust estimator with the *remaining* observations. Since Huber's estimator is translation invariant, as shown earlier, this approach does yield a translation invariant *and* robust estimator T_n^R.

Corollary 3.2 Under the conditions of Theorems 2.4 and 3.1, the recursive minimax robust estimator T_n^R has asymptotic variance no smaller than that of Huber's minimax robust M-estimator. The asymptotic variance of T_n^R is minimized for the choice

$$a^* \; = \; - \frac{1}{\xi'(T^*)}, \tag{3.203}$$

for which the two estimators are asymptotically equivalent.

Proof From Theorem 3.1,

$$\lim_{n\to\infty} \operatorname{var}[\, T_n^R \,] \; = \; - \frac{a^2\, \sigma^2(T^*)}{2a\, \xi'(T^*) + 1} \tag{3.204}$$

so that

$$\frac{\partial}{\partial a} \lim_{n\to\infty} \operatorname{var}[\, T_n^R \,] \; = \; - \frac{2a\, \sigma^2(T^*)\,(\, 2a\, \xi'(T^*) + 1\,) \; - \; 2a^2\, \sigma^2(T^*)\, \xi'(T^*)}{(\, 2a\, \xi'(T^*) + 1\,)^2} \tag{3.205}$$

$$= \; 0 \tag{3.206}$$

at the optimal value $a = a^*$. Since $a^* \neq 0$ (by (3.42)) and $\sigma^2(T^*) > 0$ by hypothesis, (3.206) may be rewritten as

$$a^*\, \xi'(T^*) + 1 \; = \; 0, \tag{3.207}$$

which yields (3.203); note that this value is consistent with the inequality in (3.42), since $\xi'(T^*) < 0$ by hypothesis.

Moreover,

$$\frac{\partial^2}{\partial a^2} \lim_{n\to\infty} \operatorname{var}[\, T_n^R \,] \; = \; - \frac{2\, \sigma^2(T^*)}{(\, 2a\, \xi'(T^*) + 1\,)^4} \Big[(\, 2a\, \xi'(T^*) + 1\,)^3$$

$$- \; 4a\, \xi'(T^*)\,(\, a\, \xi'(T^*) + 1\,)\,(\, 2a\, \xi'(T^*) + 1\,) \Big] \tag{3.208}$$

$$= \; - \frac{2\, \sigma^2(T^*)}{(\, 2a\, \xi'(T^*) + 1\,)^3} \tag{3.209}$$

$$> \; 0 \tag{3.210}$$

for all a satisfying (3.42), confirming that (3.203) corresponds to a global minimum.

Finally, substituting (3.203) into (3.204) yields

$$\min_{a} \lim_{n\to\infty} \operatorname{var}[\, T_n^R \,] \; = \; \frac{\sigma^2(T^*)}{(\, \xi'(T^*)\,)^2} \tag{3.211}$$

which, using (3.90) with $T = T^*$, corresponds to the asymptotic variance of Huber's minimax robust M-estimator in equation (2.92). ■

Corollary 3.3 For a given family of symmetric distributions with location parameter θ, let the least favorable density f_θ be such that the corresponding influence-bounding function ψ_θ satisfies the conditions of Theorem 3.1. Let T_n^R be the recursive minimax robust estimator of θ defined by equation (3.4), with coefficients $\{a_n\}$ satisfying the conditions of Theorem 3.1 as well as (3.203). If the true underlying distribution is $f_{\theta*}$, then

$$\mathbf{L}(\sqrt{n}\ (T_n^R - \theta^*\ )\ ) \rightarrow \mathbf{N}\left[0, \frac{1}{I(f_{\theta*})}\right] \tag{3.212}$$

as $n \rightarrow \infty$ (i.e. T_n^R is *asymptotically efficient*). In that case,

$$a^* = \frac{1}{I(f_{\theta*})}. \tag{3.213}$$

Proof Note first that by equations (2.131)-(2.135) (with 0 replaced by θ),

$$\xi'(\theta) = -I(f_\theta) \tag{3.214}$$

provided that the true distribution is the least favorable one. Thus, the condition that $|\xi'(\theta)| < \infty$ in Theorem 3.1 implies that $I(f_\theta) < \infty$, and the corollary is a direct consequence of Corollaries 2.1 and 3.2. ■

Corollaries 3.2 and 3.3 give some hints as to the choice of coefficients $\{a_n\}$ that yields minimum asymptotic variance. Although these results are of little immediate practical value, since neither T^* nor the true distribution P (and therefore the function $\xi(T)$) are known *a priori*, they can potentially serve to help making clever choices of coefficients. For instance, minimax optimality can be ensured by chosing P to be the least favorable distribution. The latest estimate T_n^R of T^* can be substituted for T^* in the expression for a_n, making this latter a function of the data (i.e. *adaptive* gains). That the recursion would still converge under this scheme, and have the various properties derived earlier, remains to be demonstrated. .

The result in Corollary 3.3 is implicitly used, for instance, by Price and Vandelinde (1979) for the special case $a_n = an^{-1}$. While this form is used widely, and is in fact assumed by Sacks (1958) and by those who base their work on his results (such as Martin and Masreliez), it is not necessarily a good choice for the small-sample behavior of T_n^R. For example, Dvoretzky (1956) shows in the special case where $\xi(T)$ is bounded from above and below by straight lines with finite, nonvanishing slopes, and furthermore $(T_1^R - T^*)^2$ (or equivalently $E[\ (T_1^R - T^*)^2\ ]$ if T_1^R is a random variable) is bounded by a function of these slopes, that the choice

$$a_n = \frac{K_1}{K_2 + n} \tag{3.215}$$

(where K_1 and K_2 are constants satisfying certain conditions) is minimax in the sense of providing the minimum upper bound on the estimation error variance $E[\ (T_n^R - T^*)^2\ ]$ for all n. In other words, for any $\{a_n\}$ other than that given by (3.215), there exist T_1^R and $\psi(x_n - T_n^R)$ satisying the above condition on $\xi(T)$ as well as all the conditions of Theorem 3.1, for which Dvoretzky's upper bound is violated for some n. Moreover, under certain conditions, larger values of K_2 can result in tighter upper bounds on the error variance. On the other hand, in general, (3.215) leads to estimates with asymptotic variance greater than the minimum given by (3.211), a loss in asymptotic efficiency that is "the price paid for small-sample optimality" (Derman, 1956).

The choice of optimal coefficients $\{a_n\}$ remains a difficult problem to which no satisfactory solution presently exists. Dvoretzky's assumptions are limiting, especially in the case of robust estimation, as discussed earlier in the context of the work of Chung. Moreover, it is not clear that an upper bound is necessarily the performance measure of choice in selecting optimal a_n; on the other hand, since $E[\,(T_n^R-T^*\,)^2\,]=O(\,n^{-1}\,)$ as a consequence of Theorem 3.1, the sum

$$\sum_{n=1}^{\infty} E[\,(T_n^R-T^*\,)^2\,] \tag{3.216}$$

does not converge and can therefore not be used as an objective function either. One possible approach might be to choose a finite N and then find

$$\min_{\{a_n\}} \sum_{n=1}^{N} E[\,(T_n^R-T^*\,)^2\,] \tag{3.217}$$

subject to the constraints

$$E[\,T_{n+1}^R\,] \;=\; E[\,T_n^R\,] \;+\; a_n\,E[\,\psi(x_n-T_n^R)\,] \tag{3.218}$$

$$=\; E[\,T_n^R\,] \;+\; a_n\,E\Big[\,E[\,\psi(x_n-T_n^R)\mid T_n^R\,]\,\Big] \tag{3.219}$$

$$=\; E[\,T_n^R\,] \;+\; a_n\,E[\,\xi(T_n^R)\,] \tag{3.220}$$

w.p.1, and $n=1,\,\cdots,N-1$, where (3.218) follows from (3.4) (and is used instead of this latter in the usual way in order to ensure that the solution is not a function of the data $\{x_n\}$), and (3.220) follows from (2.82). What makes this problem especially difficult, however, is that $\xi(T)$ is generally nonlinear, and moreover $\xi(T_n^R)$, $n=1,\,\cdots,N-1$ are not independent nor identically distributed. Moreover, there is no guarantee that the results thus obtained for finite N would be consistent at the limit with the asymptotic properties discussed earlier. The problem remains intractable even for (realistic) special cases, so that only asymptotic results are sought in the sequel.

It is noted, finally, that the recursive robust estimator has been studied for the ε-contaminated normal neighborhood by Martin (1972); the "p-point" neighborhood

$$P'_{A,p} \;:=\; \{\,P:\int_{+A}^{\infty} dP(x)=p\,/2,\;\; P\in \mathbf{S},\;\; P \text{ continuous at } A\,\} \tag{3.221}$$

(which is a special case of $P_{A,p}$ in (2.198)) by Martin and Masreliez (1975); and the "ε-G" neighborhood

$$P_{\varepsilon,G} \;:=\; \{\,P:\sup_{x\in \mathbf{X}}\;|\,P(x)-G(x)\,|\le\varepsilon,\;\; P,G\in \mathbf{S}\,\} \tag{3.222}$$

(which is a generalization of P_ε' in (2.197)), and the "generalized moment-constrained" neighborhood

$$P_{\{p_n\}} \;:=\; \{\,P:\int x^n\,dP(x)\le p_n,\;\; P\in \mathbf{S},\;\; 0\le n\le N\,\} \tag{3.223}$$

by Price and Vandelinde (1979).

For the problem of robustness in the presence of outliers, those of primary interest here are the ε-contaminated normal and the p-point neighborhoods. Indeed, the latter seems particularly appealing: if A is chosen, based on physical considerations, to be at the limit of acceptable noise (so that anything beyond it can be viewed as an "outlier"), then $P'_{A,p}$ yields itself to an interpretation as the neighborhood of all symmetric noise distributions containing a certain fraction p of outliers. This is very general and does not assume underlying normality, as does the ε-contaminated neighborhood. Furthermore, the estimator has constant asymptotic variance over a class of noise distributions. However, there is a price to this generality: just like ψ_ε, the influence-bounding function derived by Martin and Masreliez for the p-point neighborhood is flat beyond $\pm A$; however, it is not linear in the "center," so that *all* observations *are* processed, to a greater or lesser extent. Even the slope *at* the point of symmetry does not generally equal unity. This suggests a loss of efficiency at the nominal (underlying) model -- especially when it is nearly normal, as is often the case. For this reason, only the ε-contaminated normal neighborhood is used in the present analysis.

3.2 The Multivariate Case

So far, the discussion has been limited to estimators of scalar location parameters from univariate observations corrupted by noise. In the present section, these results are extended to vector location parameters and multivariate observations.

Multivariate extensions of the Robbins-Monro stochastic approximation procedure have been proposed by Blum (1954b), Block (1956), Sacks (1958), Derman and Sacks (1959), Epling (1964), and Fabian (1968). The primary limitation of these results is that all but Fabian's are restricted to scalar sequences $\{a_n\}$, and thus do not provide a means of attaining minimum asymptotic variance; this latter, on the other hand, makes a last-minute assumption (as to the normality of the updates) that neither appears in the statement of his theorem, nor is consistent with his claim of total generality.

Consider, as before, the measure space $(\mathbf{X}, \mathbf{B}, \mu)$ where $\mathbf{X}$ is now $\mathbf{R}^p$, $\mathbf{B}$ the Borel σ-algebra, and μ the Lebesgue measure. Let $\{\underline{x}_1, \cdots, \underline{x}_n\}$ be a sample of independent random variates taking values in $\mathbf{X}$, with a common distribution function P; let $\mathbf{P} := \{P_{\underline{\theta}} : \underline{\theta} \in \Theta\}$, where $\Theta \subseteq \mathbf{R}^q$, be a family of probability measures on $(\mathbf{X}, \mathbf{B})$ such that for all $\underline{\theta} \in \Theta$, $P_{\underline{\theta}}$ is absolutely continuous with respect to μ and admits the density $f_{\underline{\theta}}$ in accordance with the Radon-Nikodym theorem.

Let $\mathbf{X}^n$ be the product of n copies of $\mathbf{X}$, and let $T_n : \mathbf{X}^n \to \Theta$ be Huber's minimax robust M-estimator for the parameter $\underline{\theta}$, i.e. the maximum-likelihood estimator given the least favorable distribution $f^*_{\underline{\theta}} \in \mathbf{P}$. It follows that T_n is the solution of maximization problem

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f^*_{\underline{\theta}}(\underline{x}_i), \tag{3.224}$$

or, alternatively (provided that Θ is an open set) of the system of equations

$$\sum_{i=1}^{n} \underline{\psi}_{\underline{\theta}}(\underline{x}_i) = 0 \tag{3.225}$$

where

$$\underline{\psi}_{\underline{\theta}}(\underline{x}) = -\underline{\nabla}_{\underline{\theta}} \log f^*_{\underline{\theta}}(\underline{x}) \tag{3.226}$$

$$= -\frac{1}{f^*_{\underline{\theta}}(\underline{x})} \; \underline{\nabla}_{\underline{\theta}} f^*_{\underline{\theta}}(\underline{x}) \tag{3.227}$$

a.s., within an arbitrary multiplicative constant. ($\underline{\nabla}_{\underline{\theta}}$ denotes the gradient with respect to the parameter vector $\underline{\theta}$; compare equations (2.77)-(2.80).) That this is minimax follows directly from the results of Section 2. Note also that for the case of location parameters, $p = q$; this is assumed in the sequel.

As before, consider the recursion

$$\underline{T}^R_{n+1} = \underline{T}^R_n + A_n \, \underline{\psi}(\underline{x}_n - \underline{T}^R_n), \tag{3.228}$$

where $n = 1, 2, \cdots$, $\{A_n\}$ is a given matrix sequence with $A_n \in \mathbf{R}^{q \times q}$, and $\underline{T}^R_1$ is an arbitrary (possibly random) starting point. Let

$$\underline{\xi}(\underline{T}) := E_P \, [\, \underline{\psi}(\underline{x} - \underline{T}) \,], \tag{3.229}$$

$$\Sigma(\underline{T}) := E_P \left[(\, \underline{\psi}(\underline{x}-\underline{T}) - \underline{\xi}(\underline{T}) \,) (\, \underline{\psi}(\underline{x}-\underline{T}) - \underline{\xi}(\underline{T}) \,)^{\mathrm{T}} \right], \tag{3.230}$$

and define

$$J(\underline{T}) := \left[\frac{\partial}{\partial t_j} \, \xi_i(\underline{t}) \right]_{\underline{t} = \underline{T}} \tag{3.231}$$

as the Jacobian of $\underline{\xi}(\underline{T})$, provided it exists. The following is a generalization of Lemma 3.3.

Lemma 3.4 Let $\{b_n\}$ be a real sequence such that, for some n_0,

$$b_{n+1} = b_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} \tag{3.232}$$

for all $n \geq n_0$, where $\{c_n\}$ is a real sequence with

$$\lim_{n \to \infty} c_n = c > p, \tag{3.233}$$

$\{d_n\}$ is a real sequence with

$$\lim_{n \to \infty} d_n = d, \tag{3.234}$$

and $p > 0$. Then,

$$\lim_{n \to \infty} n^p \, b_n = \frac{d}{c - p}. \tag{3.235}$$

Proof Assume first that $d \neq 0$. Then, there is a large enough n_1 such that

$$1 - \frac{c_n}{n} > 0 \tag{3.236}$$

and (from (3.234)) either $d_n > 0$ (if $d > 0$) or $d_n < 0$ (if $d < 0$) for all $n \geq n_1$. If, for any $n_2 \geq \max(n_0, n_1)$, b_{n_2} has the same sign as d, then clearly b_n has the same sign as d for all $n \geq n_2$, and Lemma 3.3 (after multiplying (3.232) through by -1 if $d < 0$) establishes (3.235).

Assume that b_{n_2} and d have opposite signs for some $n_2 \geq \max(n_0, n_1)$, and (with no loss of generality, since the other case can be reduced to this one by multiplying (3.232) through by -1) let $b_{n_2} < 0$ and $d > 0$. Rewriting (3.232) as

$$| b_{n+1} | \leq | b_n | \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} \tag{3.237}$$

for all $n \geq n_2$, Lemma 3.1 yields

$$\limsup_{n \to \infty} n^p | b_n | \leq \frac{d}{c - p}, \tag{3.238}$$

or

$$\liminf_{n \to \infty} n^p b_n \geq - \frac{d}{c - p}. \tag{3.239}$$

It follows that, for any $\delta > 0$, there is a large enough $n(\delta)$ such that

$$b_n \geq - \frac{1}{n^p} \left[\frac{d}{c - p} + \delta \right] \tag{3.240}$$

for all $n \geq n(\delta)$. Thus, defining

$$\tilde{b}_n := b_n + \frac{1}{n^p} \left[\frac{d}{c - p} + \delta \right], \tag{3.241}$$

and noting that $\tilde{b}_n \geq 0$ for all $n \geq n(\delta)$ from (3.240), equation (3.232) may be rewritten as

$$\tilde{b}_{n+1} = \tilde{b}_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}} + \left[\frac{1}{(n+1)^p} - \left[1 - \frac{c_n}{n} \right] \frac{1}{n^p} \right] \left[\frac{d}{c - p} + \delta \right] \tag{3.242}$$

for $n \geq \max(n_0, n_1, n(\delta))$, where, from (3.13)-(3.16),

$$\frac{1}{(n+1)^p} - \left[1 - \frac{c_n}{n} \right] \frac{1}{n^p} = \frac{c_n - p}{n^{p+1}} + O(n^{-(p+2)}). \tag{3.243}$$

Substituting (3.243) into (3.242), and noting that

$$\lim_{n \to \infty} d_n + (c_n - p) \left[\frac{d}{c - p} + \delta \right] + O(n^{-1}) = 2d + (c - p)\delta > 0 \tag{3.244}$$

from (3.233)-(3.234) and by hypothesis, Lemma 3.3 implies that

$$\lim_{n \to \infty} n^p \tilde{b}_n = \frac{2d + (c - p)\delta}{c - p}, \tag{3.245}$$

and (3.241) establishes (3.235).

Finally, if $d = 0$, choose any $\delta > 0$ and define

$$\tilde{b}_n := b_n + \frac{\delta}{n^p}. \tag{3.246}$$

Then, (3.232) may be rewritten as

$$\tilde{b}_{n+1} = \tilde{b}_n \left[1 - \frac{c_n}{n} \right] + \frac{d_n}{n^{p+1}}$$

$$+ \left[\frac{1}{(n+1)^p} - \left[1 - \frac{c_n}{n} \right] \frac{1}{n^p} \right] \delta \tag{3.247}$$

$$= \tilde{b}_n \left[1 - \frac{c_n}{n} \right] + \frac{1}{n^{p+1}} \left[d_n + (c_n - p)\,\delta + O(n^{-1}) \right] \tag{3.248}$$

for $n \geq n_0$, from (3.43). Since

$$\lim_{n \to \infty} d_n + (c_n - p)\,\delta + O(n^{-1}) = (c - p)\,\delta > 0 \tag{3.249}$$

from (3.233) and by hypothesis, the problem is reduced to one already solved, and

$$\lim_{n \to \infty} n^p \, \tilde{b}_n = \frac{(c - p)\,\delta}{c - p} \tag{3.250}$$

$$= \delta, \tag{3.251}$$

so that (3.246) establishes (3.235) with $d = 0$. ■

The following theorem, based on the results of Blum and of Fabian, is a multivariate generalization of Theorem 3.1.

Theorem 3.2 Let $\underline{\xi}(\underline{T})$ exist for all $\underline{T}$, and let there be a $\underline{T}^*$ such that for any $\delta > 0$ and all $q \times q$ matrices $M > 0$,

$$\sup_{\delta \leq \|\underline{T} - \underline{T}^*\|} (\underline{T} - \underline{T}^*)^{\mathrm{T}} M \, \underline{\xi}(\underline{T}) < 0. \tag{3.252}$$

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

$$E_P \left[\underline{\psi}(\underline{x} - \underline{T}) \, \underline{\psi}^{\mathrm{T}}(\underline{x} - \underline{T}) \right] \leq S_0 \tag{3.253}$$

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

$$\sum_{n=1}^{\infty} A_n = \infty, \tag{3.254}$$

and

$$\sum_{n=1}^{\infty} A_n^{\mathrm{T}} A_n < \infty. \tag{3.255}$$

Then, given any $\underline{T}_1^R < \infty$, $\underline{T}_n^R \to \underline{T}^*$ as $n \to \infty$ a.s. (i.e. $\underline{T}_n^R$ is *consistent*).

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

$$\lim_{n\to\infty} n\, A_n = A > -\frac{1}{2} J^{-1}(\underline{T}^*), \tag{3.256}$$

then

$$\mathbf{L}(\sqrt{n}\ (\underline{T}_n^R - \underline{T}^*)) \rightarrow \mathbf{N}\left[0, V\, Q\, V^{\mathrm{T}}\right], \tag{3.257}$$

where $Q = [\, q_{ij}\,]$,

$$q_{ij} = \frac{[\, V^{\mathrm{T}} A\ \Sigma(\underline{T}^*)\, A^{\mathrm{T}} V\,]_{ij}}{\lambda_i + \lambda_j - 1} \tag{3.258}$$

and V is an orthogonal matrix and $\Lambda = [\, \lambda_i\,]$ a diagonal matrix such that

$$V^{\mathrm{T}} A\, J(\underline{T}^*)\, V = -\Lambda \tag{3.259}$$

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

Proof Equation (3.228) may be rewritten as

$$\underline{T}_{n+1}^R - \underline{T}^* = \underline{T}_n^R - \underline{T}^* + A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,), \tag{3.260}$$

whence, squaring and taking expectations, it follows that

$$\begin{aligned} E&\left[(\, \underline{T}_{n+1}^R - \underline{T}^*\,)^{\mathrm{T}} (\, \underline{T}_{n+1}^R - \underline{T}^*\,) \right] \\ &= E\left[(\, \underline{T}_n^R - \underline{T}^*\,)^{\mathrm{T}} (\, \underline{T}_n^R - \underline{T}^*\,) \right] \\ &\qquad + 2\, E\left[(\, \underline{T}_n^R - \underline{T}^*\,)^{\mathrm{T}} A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,) \right] \\ &\qquad + E\left[\underline{\psi}^{\mathrm{T}}(\, \underline{x}_n - \underline{T}_n^R\,)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,) \right] && (3.261) \\ &= (\, \underline{T}_1^R - \underline{T}^*\,)^{\mathrm{T}} (\, \underline{T}_1^R - \underline{T}^*\,) \\ &\qquad + 2 \sum_{j=1}^{n} E\left[(\, \underline{T}_j^R - \underline{T}^*\,)^{\mathrm{T}} A_j\, \underline{\psi}(\, \underline{x}_j - \underline{T}_j^R\,) \right] \\ &\qquad + \sum_{j=1}^{n} E\left[\underline{\psi}^{\mathrm{T}}(\, \underline{x}_j - \underline{T}_j^R\,)\, A_j^{\mathrm{T}} A_j\, \underline{\psi}(\, \underline{x}_j - \underline{T}_j^R\,) \right]. && (3.262) \end{aligned}$$

Note first that since (3.262) is scalar,

$$\begin{aligned} \underline{\psi}^{\mathrm{T}}(\, \underline{x}_n - \underline{T}_n^R\,)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,) &= \mathrm{tr}\left[\underline{\psi}^{\mathrm{T}}(\, \underline{x}_n - \underline{T}_n^R\,)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,) \right] && (3.263) \\ &= \mathrm{tr}\left[\underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,)\, \underline{\psi}^{\mathrm{T}}(\, \underline{x}_n - \underline{T}_n^R\,)\, A_n^{\mathrm{T}} A_n \right], && (3.264) \end{aligned}$$

so that

$$E\left[\underline{\psi}^{\mathrm{T}}(\, \underline{x}_n - \underline{T}_n^R\,)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\, \underline{x}_n - \underline{T}_n^R\,) \right]$$

$$= \operatorname{tr}\left[E\left[E_P\left[\underline{\psi}(\underline{x}_n - \underline{T}_n^R)\, \underline{\psi}^{\mathrm{T}}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_n^R \right] \right] A_n^{\mathrm{T}} A_n \right] \tag{3.265}$$

$$\leq \operatorname{tr}\left[S_0 A_n^{\mathrm{T}} A_n \right] \tag{3.266}$$

w.p.1 for all n, where (3.265) follows from (3.264), and (3.266) from (3.253). Thus,

$$\sum_{j=1}^{n} E\left[\underline{\psi}^{\mathrm{T}}(\underline{x}_j - \underline{T}_j^R)\, A_j^{\mathrm{T}} A_j\, \underline{\psi}(\underline{x}_j - \underline{T}_j^R) \right] \leq \operatorname{tr}\left[S_0 \sum_{j=1}^{n} A_j^{\mathrm{T}} A_j \right] \tag{3.267}$$

$$< \infty \tag{3.268}$$

w.p.1 for $n \to \infty$, from (3.255) and the finiteness by hypothesis of S_0. Moreover,

$$E\left[(\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \right]$$

$$= E\left[E_P\left[(\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_n^R \right] \right] \tag{3.269}$$

$$= E\left[(\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\xi}(\underline{T}_n^R) \right] \tag{3.270}$$

$$< 0 \tag{3.271}$$

w.p.1 for all n, from (3.252) and the positivity by hypothesis of A_n. Thus,

$$\sum_{j=1}^{n} E\left[(\underline{T}_j^R - \underline{T}^*)^{\mathrm{T}} A_j\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \right] < 0 \tag{3.272}$$

w.p.1 for all n. But since

$$E\left[(\underline{T}_{n+1}^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_{n+1}^R - \underline{T}^*) \right] \geq 0 \tag{3.273}$$

because the term is in quadratic form, and

$$(\underline{T}_1^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_1^R - \underline{T}^*) < \infty \tag{3.274}$$

by hypothesis, it follows that

$$\sum_{j=1}^{\infty} E\left[(\underline{T}_j^R - \underline{T}^*)^{\mathrm{T}} A_j\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \right] = \sum_{j=1}^{\infty} E\left[(\underline{T}_j^R - \underline{T}^*)^{\mathrm{T}} A_j\, \underline{\xi}(\underline{T}_j^R) \right] \tag{3.275}$$

must be bounded from below w.p.1.

Define now

$$Y_n := E\left[(\underline{T}_{n+1}^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_{n+1}^R - \underline{T}^*) - (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right], \tag{3.276}$$

and consider the sequence

$$\left\{ (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) - \sum_{j=1}^{n-1} Y_j \right\}. \tag{3.277}$$

Since

$$E\left[(\underline{T}_{n+1}^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_{n+1}^R - \underline{T}^*) - \sum_{j=1}^{n} Y_j \;\Big|\; \underline{T}_1^R, \cdots, \underline{T}_n^R \right]$$

$$= E\left[(\underline{T}_{n+1}^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_{n+1}^R - \underline{T}^*) - (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) \right.$$

$$\left. + (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) - \sum_{j=1}^{n} Y_j \;\Big|\; \underline{T}_1^R, \cdots, \underline{T}_n^R \right] \tag{3.278}$$

$$= Y_n + (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) - \sum_{j=1}^{n} Y_j \tag{3.279}$$

$$= (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*) - \sum_{j=1}^{n-1} Y_j \tag{3.280}$$

w.p.1, it follows that (3.277) is a martingale.

Squaring (3.260) and taking conditional expectations,

$$E\left[(\underline{T}_{n+1}^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_{n+1}^R - \underline{T}^*) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right]$$

$$= (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*)$$

$$+ 2\, E\left[(\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right]$$

$$+ E\left[\underline{\psi}^{\mathrm{T}}(\underline{x}_n - \underline{T}_n^R)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right] \tag{3.281}$$

$$= (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*)$$

$$+ 2\, (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\xi}(\underline{T}_n^R)$$

$$+ E\left[\underline{\psi}^{\mathrm{T}}(\underline{x}_n - \underline{T}_n^R)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right] \tag{3.282}$$

w.p.1; it then follows from (3.276) that

$$Y_n = 2\, (\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} A_n\, \underline{\xi}(\underline{T}_n^R)$$

$$+ E\left[\underline{\psi}^{\mathrm{T}}(\underline{x}_n - \underline{T}_n^R)\, A_n^{\mathrm{T}} A_n\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) \mid \underline{T}_1^R, \cdots, \underline{T}_n^R \right]. \tag{3.283}$$

Thus,

$$\sum_{j=1}^{n} Y_j = 2 \sum_{j=1}^{n} (\underline{T}_j^R - \underline{T}^*)^{\mathrm{T}} A_j\, \underline{\xi}(\underline{T}_j^R)$$

$$+ \sum_{j=1}^{n} E\left[\underline{\psi}^{\mathrm{T}}(\underline{x}_j - \underline{T}_j^R)\, A_j^{\mathrm{T}} A_j\, \underline{\psi}(\underline{x}_j - \underline{T}_j^R) \mid \underline{T}_1^R, \cdots, \underline{T}_j^R \right] \tag{3.284}$$

is a sum of two terms whose expectations are monotone and bounded: the first is negative a.s., from (3.252), and was shown to be bounded from below; the second is nonnegative since it is in quadratic form, and obeys (3.268). (In both cases, the positivity by hypothesis of A_j is utilized.) Thus, the expectation of (3.284) converges almost surely to a finite limit as $n \to \infty$ a.s. Hence,

$$E\left[\,|\,(\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_n^R-\underline{T}^*\,)\,|\,\right]$$

$$= E\left[(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_n^R-\underline{T}^*\,)\right] \tag{3.285}$$

$$= (\,\underline{T}_1^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_1^R-\underline{T}^*\,)$$

$$+ \sum_{j=1}^{n-1} E\left[(\,\underline{T}_{j+1}^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_{j+1}^R-\underline{T}^*\,)\right.$$

$$\left. - (\,\underline{T}_j^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_j^R-\underline{T}^*\,)\right] \tag{3.286}$$

$$= (\,\underline{T}_1^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_1^R-\underline{T}^*\,) + \sum_{j=1}^{n-1} E[\,Y_j\,] \tag{3.287}$$

$$< \infty \tag{3.288}$$

a.s. for $n \to \infty$, where (3.285) holds because the term is in quadratic form and hence non-negative, and (3.288) follows from the finite convergence of the expectation of (3.284), and from the finiteness by hypothesis of $\underline{T}_1^R$. Thus,

$$\lim_{n\to\infty} E\left[\,\left|\,(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_n^R-\underline{T}^*\,) - \sum_{j=1}^{n-1} Y_j\,\right|\,\right]$$

$$\leq \lim_{n\to\infty} E\left[\,(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_n^R-\underline{T}^*\,)\,\right]$$

$$- \lim_{n\to\infty} E\left[2\sum_{j=1}^{n}(\,\underline{T}_j^R-\underline{T}^*\,)^{\mathrm{T}} A_j\, \underline{\xi}(\underline{T}_j^R)\right]$$

$$+ \lim_{n\to\infty} E\left[\sum_{j=1}^{n} \underline{\psi}^{\mathrm{T}}(\,\underline{x}_j-\underline{T}_j^R\,)\,A_j^{\mathrm{T}} A_j\, \underline{\psi}(\,\underline{x}_j-\underline{T}_j^R\,)\right] \tag{3.289}$$

$$< \infty \tag{3.290}$$

a.s. from (3.288) and the finite convergence of (3.284). Using a martingale convergence theorem due to Doob (1953, pp.319-323), it then follows that the sequence (3.277) converges a.s.; moreover, since the second term in (3.284) is monotone and bounded, and the first is monotone and appears in (3.277) as a positive quantity, and $(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}\,(\,\underline{T}_n^R-\underline{T}^*\,) \geq 0$ also due to its quadratic form, it follows that this latter also converges a.s. There is therefore a T_0 such that

$$\lim_{n\to\infty}\,(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}}(\,\underline{T}_n^R-\underline{T}^*\,) = T_0 \tag{3.291}$$

w.p.1. But the boundedness from below of (3.275) implies that

$$\limsup_{n\to\infty}\; E\left[(\,\underline{T}_n^R-\underline{T}^*\,)^{\mathrm{T}} A_n\, \underline{\xi}(\underline{T}_n^R)\right] = 0 \tag{3.292}$$

or, since the expression is negative a.s. from (3.252),

$$\liminf_{n\to\infty} \; E\left[\, \left|\, (\, \underline{T}_n^R - \underline{T}^* \,)^{\mathrm{T}} A_n \, \xi(\underline{T}_n^R) \,\right| \,\right] \;=\; 0, \tag{3.293}$$

implying that there exists a subsequence $\{n_k\}$ such that

$$\lim_{k\to\infty} \; E\left[\, \left|\, (\, \underline{T}_{n_k}^R - \underline{T}^* \,)^{\mathrm{T}} A_{n_k} \, \xi(\underline{T}_{n_k}^R) \,\right| \,\right] \;=\; 0. \tag{3.294}$$

Hence, by the Chebychev inequality (Loève, 1963, p.11),

$$\lim_{k\to\infty} \; (\, \underline{T}_{n_k}^R - \underline{T}^* \,)^{\mathrm{T}} A_{n_k} \, \xi(\underline{T}_{n_k}^R) \;=\; 0 \tag{3.295}$$

w.p.1, which in turn implies, from (3.252), that

$$\lim_{k\to\infty} \; (\, \underline{T}_{n_k}^R - \underline{T}^* \,) \;=\; 0 \tag{3.296}$$

w.p.1. Combining (3.291) and (3.296) establishes that

$$\lim_{n\to\infty} \; (\, \underline{T}_n^R - \underline{T}^* \,) \;=\; 0 \tag{3.297}$$

w.p.1, which proves the consistency of the estimator $\underline{T}_n^R$.

To prove asymptotic normality, note first that by hypothesis, $\xi(\underline{T})$ is continuous and differentiable in a neighborhood of $\underline{T}^*$, say $\| \underline{T} - \underline{T}^* \| < \delta_1$, and $\xi(\underline{T}^*) = 0$. Thus,

$$\xi(\underline{T}) \;=\; J(\underline{T}^*)\,(\, \underline{T} - \underline{T}^* \,) \;+\; O(\, \| \underline{T} - \underline{T}^* \|^2 \,) \tag{3.298}$$

for $\| \underline{T} - \underline{T}^* \| < \delta_1$. Moreover, since $\underline{T}_n^R \to \underline{T}^*$ w.p.1 (as proved above), there exists a large enough $n(\delta_1)$ such that $\| \underline{T}_n^R - \underline{T}^* \| < \delta_1$ w.p.1 for all $n \ge n(\delta_1)$. It follows that (3.260) may be rewritten as

$$\underline{T}_{n+1}^R \,-\, \underline{T}^* \;=\; \underline{T}_n^R \,-\, \underline{T}^* \;+\; A_n \left[\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) - \xi(\underline{T}_n^R) \,\right] \;+\; A_n \, \xi(\underline{T}_n^R) \tag{3.299}$$

$$= \left[\, I + A_n \, J(\underline{T}^*) + A_n \, O_p(\, \| \underline{T}_n^R \,-\, \underline{T}^* \| \,) \,\right] (\, \underline{T}_n^R \,-\, \underline{T}^* \,)$$

$$+ \; A_n \left[\, \underline{\psi}(\underline{x}_n - \underline{T}_n^R) - \xi(\underline{T}_n^R) \,\right] \tag{3.300}$$

w.p.1 for $n \ge n(\delta_1)$, so that $(\, \underline{T}_{n+1}^R \,-\, \underline{T}^* \,)$ is the sum of a sequence of zero-mean random variables (plus some higher-order terms). Note that

$$A_n \; O_p(\, \| \underline{T}_n^R \,-\, \underline{T}^* \| \,) \;=\; o_p(n^{-1}) \tag{3.301}$$

at least, in view of (3.256) (which implies that $A_n = O(n^{-1})$) and (3.297) (which implies that $O_p(\, \| \underline{T}_n^R \,-\, \underline{T}^* \| \,) = o_p(1)$ or less).

Next, define for some $\delta_2 > 0$ the set

$$A(\, n \,, \delta_2 , \underline{T} \,) \;:=\; \left\{ \underline{x} : \; \| \, \underline{\psi}(\underline{x} - \underline{T}) - \xi(\underline{T}) \, \|^2 \;\ge\; \delta_2 \, n \, \right\}. \tag{3.302}$$

Since $\xi(\underline{T}^*) = 0$ by hypothesis, $\xi(\underline{T}_n^R) < \infty$ w.p.1 for all $n \ge n(\delta_1)$ by virtue of continuity. Together with (3.253), this implies that

$$\| \, \underline{\psi}(\underline{x} - \underline{T}_n^R) - \xi(\underline{T}_n^R) \, \|^2 \;<\; \infty \tag{3.303}$$

w.p.1 for $n \geq n(\delta_1)$, so that

$$\lim_{n \to \infty} A(n, \delta_2, \underline{T}_n^R) = \emptyset \qquad (3.304)$$

(or possibly a set of measure 0). It then follows that

$$\lim_{n \to \infty} \int_{A(n, \delta_2, \underline{T}_n^R)} \| \underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \|^2 \, dP(\underline{x}) = 0 \qquad (3.305)$$

w.p.1 for any $\delta_2 > 0$. This is analogous to Lindeberg's condition for asymptotic normality, and is used in the proof below.

The characteristic function of the update in (3.300) is defined as

$$\zeta_n^{\Psi}(\underline{s}) := E\left[e^{i \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x}_n - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))} \right] \qquad (3.306)$$

$$= E\left[E_P \left[e^{i \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))} \mid \underline{T}_n^R \right] \right] \qquad (3.307)$$

w.p.1, since $\{\underline{x}_n\}$ are independent and identically distributed. Using Taylor's theorem yields

$$E_P \left[e^{i \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))} \mid \underline{T}_n^R \right]$$

$$= E_P \Big[1 + i \, \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right)$$

$$- \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right) \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right)^{\mathrm{T}} A_n^{\mathrm{T}} \underline{s}$$

$$+ R_n \mid \underline{T}_n^R \Big] \qquad (3.308)$$

$$= 1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}_n^R) \, A_n^{\mathrm{T}} \underline{s} + E_P[R_n \mid \underline{T}_n^R], \qquad (3.309)$$

where R_n denotes the remainder, from (3.229) and (3.230). Since the truncation error is dominated by the first omitted term in the Taylor series (see, for instance, Feller, 1966, vol.2, p.485),

$$| E_P[R_n \mid \underline{T}_n^R] |$$

$$\leq E_P[|R_n| \mid \underline{T}_n^R] \qquad (3.310)$$

$$\leq \left| -\frac{i}{6} \right| \int_{-\infty}^{\infty} | \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right) |^3 \, dP(\underline{x}) \qquad (3.311)$$

$$= \frac{1}{6} \int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))| \leq \delta_3} | \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right) |^3 \, dP(\underline{x})$$

$$+ \frac{1}{6} \int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))| > \delta_3} | \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right) |^3 \, dP(\underline{x}) \qquad (3.312)$$

for some $\delta_3 > 0$. Now:

$$\int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))| \leq \delta_3} | \underline{s}^{\mathrm{T}} A_n \left(\underline{\psi}(\underline{x} - \underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R) \right) |^3 \, dP(\underline{x})$$

$$\leq \delta_3 \int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x}-\underline{T}_n^R)-\underline{\xi}(\underline{T}_n^R))| \leq \delta_3} | \underline{s}^{\mathrm{T}} A_n \, (\underline{\psi}(\underline{x}-\underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R)) \, |^2 \, dP(\underline{x}) \tag{3.313}$$

$$\leq \delta_3 \int \underline{s}^{\mathrm{T}} A_n \, (\underline{\psi}(\underline{x}-\underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R)) \, (\underline{\psi}(\underline{x}-\underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))^{\mathrm{T}} A_n^{\mathrm{T}} \underline{s} \, dP(\underline{x}) \tag{3.314}$$

$$= \delta_3 \, \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}_n^R) \, A_n^{\mathrm{T}} \underline{s}, \tag{3.315}$$

where δ_3 can be made arbitrarily small. For the second integral in (3.312), it is more convenient to bound the error by the next lower term, i.e.

$$\int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x}-\underline{T}_n^R)-\underline{\xi}(\underline{T}_n^R))| > \delta_3} \underline{s}^{\mathrm{T}} A_n \, (\underline{\psi}(\underline{x}-\underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R)) \, (\underline{\psi}(\underline{x}-\underline{T}_n^R) - \underline{\xi}(\underline{T}_n^R))^{\mathrm{T}} A_n^{\mathrm{T}} \underline{s} \, dP(\underline{x})$$
$$= o_p(1) \, \underline{s}^{\mathrm{T}} A_n \, A_n^{\mathrm{T}} \underline{s} \tag{3.316}$$

w.p.1, from (3.305). It follows, combining (3.312), (3.315) (letting $\delta_3 \downarrow 0$), (3.316), and the fact that $A_n = O(n^{-1})$ from (3.256), that

$$| \, E_P[\, R_n \mid \underline{T}_n^R \,] \, | \leq o_p(n^{-2}) \, \| \underline{s} \|^2. \tag{3.317}$$

Denote the characteristic function of $\underline{T}_n^R - \underline{T}^*$ by

$$\zeta_n^T(\underline{s}) := E[\, e^{i \, \underline{s}^{\mathrm{T}} (\underline{T}_n^R - \underline{T}^*)} \,], \tag{3.318}$$

and define, for economy of notation, the matrix sequence

$$B_n := I + A_n \, J(\underline{T}^*) + o_p(n^{-1}) \tag{3.319}$$

(compare with (3.300)-(3.301)) and the recursion

$$\zeta_{n+1}(\underline{s}) = \zeta_n(B_n^{\mathrm{T}} \underline{s}) \left[1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) \, A_n^{\mathrm{T}} \underline{s} \right] \tag{3.320}$$

with

$$\zeta_1(\underline{s}) := \zeta_1^T = e^{i \, \underline{s}^{\mathrm{T}} (\underline{T}_1^R - \underline{T}^*)} \tag{3.321}$$

(since $\underline{T}_1^R$ is a given constant). Note that $\zeta_n(\underline{s})$ is essentially an approximation for $\zeta_n^T(\underline{s})$ obtained by substituting an approximation for $\zeta_n^\Psi(\underline{s})$; this becomes clear when (3.300) is used to write

$$\zeta_{n+1}^T(\underline{s}) = \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \, \zeta_n^\Psi(\underline{s}), \tag{3.322}$$

and a comparison is made with (3.309) and (3.320). It can be shown that $\zeta_n(\underline{s})$ and $\zeta_n^T(\underline{s})$ are asymptotically equivalent by noting that

$$\begin{aligned} | \, \zeta_{n+1}^T(\underline{s}) - \zeta_{n+1}(\underline{s}) \, | &= \Bigg| \, \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \, \zeta_n^\Psi(\underline{s}) \\ &\qquad - \zeta_n(B_n^{\mathrm{T}} \underline{s}) \left[1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) \, A_n^{\mathrm{T}} \underline{s} \right] \Bigg| \qquad (3.323) \\ &= \Bigg| \, \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \, \zeta_n^\Psi(\underline{s}) + \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \left[1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) \, A_n^{\mathrm{T}} \underline{s} \right] \end{aligned}$$

$$- \zeta_n^T(B_n^{\mathrm{T}}\underline{s}) \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right]$$

$$- \zeta_n(B_n^{\mathrm{T}}\underline{s}) \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right] \Bigg| \qquad (3.324)$$

$$= \Bigg| (\zeta_n^T(B_n^{\mathrm{T}}\underline{s}) - \zeta_n(B_n^{\mathrm{T}}\underline{s})) \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right]$$

$$+ \zeta_n^T(B_n^{\mathrm{T}}\underline{s}) \Bigg[\zeta_n^{\Psi}(\underline{s}) - 1$$

$$+ \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \Bigg] \Bigg| \qquad (3.325)$$

$$\leq \left| 1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right| \left| \zeta_n^T(B_n^{\mathrm{T}}\underline{s}) - \zeta_n(B_n^{\mathrm{T}}\underline{s}) \right|$$

$$+ \left| \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \right| \Bigg| \zeta_n^{\Psi}(\underline{s}) - 1$$

$$+ \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \Bigg|, \qquad (3.326)$$

where (3.323) follows from (3.320) and (3.322). But

$$| \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) | \leq 1 \qquad (3.327)$$

for all $\underline{s}$ (a property of all characteristic functions; see Feller, 1966, vol.2, p.473), while

$$\left| \zeta_n^{\Psi}(\underline{s}) - 1 + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right|$$

$$= \left| E\left[E_P \left[e^{i \, \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{x} - \underline{T}_n^R) - \xi(\underline{T}_n^R))} \mid \underline{T}_n^R \right] \right] - 1 + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right| \qquad (3.328)$$

$$= \left| E\left[E_P [R_n \mid \underline{T}_n^R] - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \left[\Sigma(\underline{T}_n^R) - \Sigma(\underline{T}^*) \right] A_n^{\mathrm{T}} \underline{s} \right] \right| \qquad (3.329)$$

$$\leq o_p(n^{-2}) \, \|\underline{s}\|^2 \qquad (3.330)$$

w.p.1, where (3.328) follows from (3.306), (3.329) from (3.309), and (3.330) from (3.317), (3.256), the fact that $\underline{T}_n^R \to \underline{T}^*$ w.p.1, and the continuity and boundedness of $\Sigma(\underline{T})$ in a neighborhood of $\underline{T}^*$ by hypothesis. It therefore follows that

$$\left| \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \right| \left| \zeta_n^{\Psi}(\underline{s}) - 1 + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right| = o_p(n^{-2}) \, \|\underline{s}\|^2. \qquad (3.331)$$

Similarly, again using (3.256),

$$\left| 1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) A_n^{\mathrm{T}} \underline{s} \right| = \left| 1 + O(n^{-2}) \, \|\underline{s}\|^2 \right| \qquad (3.332)$$

and

$$\| B_n \| = O(1) + o_p(1) \tag{3.333}$$

from (3.319).

Thus, defining

$$\Delta_n(\underline{s}) := | \zeta_n^T(\underline{s}) - \zeta_n(\underline{s}) |, \tag{3.334}$$

it follows that

$$\Delta_{n+1}(\underline{s}) \le \left| 1 + O(n^{-2}) \|\underline{s}\|^2 \right| \Delta_n(\,(O(1)+o_p(1))\,\underline{s}\,) + o_p(n^{-2}) \|\underline{s}\|^2, \tag{3.335}$$

from (3.326) with (3.331)-(3.333). Now: from (3.321), $\Delta_1(\underline{s}) = 0$ for all $\underline{s}$, so that (3.335) yields

$$\Delta_2(\underline{s}) \le o_p(n^{-2}) \|\underline{s}\|^2. \tag{3.336}$$

Assume by the induction hypothesis that

$$\Delta_n(\underline{s}) \le \sum_{k=1}^{n-1} o_p(n^{-2k}) \|\underline{s}\|^{2k}. \tag{3.337}$$

holds for some n. Then,

$$\Delta_{n+1}(\underline{s}) \le \left| 1 + O(n^{-2}) \|\underline{s}\|^2 \right| \sum_{k=1}^{n-1} o_p(n^{-2k}) \|\underline{s}\|^{2k} + o_p(n^{-2}) \|\underline{s}\|^2 \tag{3.338}$$

$$\le \sum_{k=1}^{n-1} o_p(n^{-2k}) \|\underline{s}\|^{2k} + \sum_{k=1}^{n-1} o_p(n^{-2k-2}) \|\underline{s}\|^{2k+2} + o_p(n^{-2}) \|\underline{s}\|^2 \tag{3.339}$$

$$= \sum_{k=1}^{n} o_p(n^{-2k}) \|\underline{s}\|^{2k}, \tag{3.340}$$

which establishes (3.337) for all n. Thus,

$$\lim_{n \to \infty} \Delta_n(\underline{s}) = 0 \tag{3.341}$$

for all finite $\underline{s}$, proving that $\zeta_n^T(\underline{s})$ and $\zeta_n(\underline{s})$ are asymptotically equivalent. Hence, it is sufficient to seek the limit of the recursion (3.320)-(3.321) in order to find the characteristic function of the limiting distribution of $\underline{T}_n^R$.

It follows from (3.320) and (3.321) that

$$\zeta_2(\underline{s}) = e^{i \underline{s}^T B_1 (\underline{T}_1^R - \underline{T}^*)} \left[1 - \frac{1}{2} \underline{s}^T A_1 \Sigma(\underline{T}^*) A_1^T \underline{s} \right], \tag{3.342}$$

so that

$$\log \zeta_2(\underline{s}) = i \underline{s}^T B_1 (\underline{T}_1^R - \underline{T}^*) + \log \left[1 - \frac{1}{2} \underline{s}^T A_1 \Sigma(\underline{T}^*) A_1^T \underline{s} \right]. \tag{3.343}$$

Assume by the induction hypothesis that

$$\log \zeta_n(\underline{s}) = i \underline{s}^T \left[\prod_{j=1}^{n-1} B_j \right] (\underline{T}_1^R - \underline{T}^*)$$

$$+ \sum_{j=1}^{n-1} \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j \, \Sigma(\underline{T}^*) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}} \underline{s} \right] \tag{3.344}$$

(where sums and products are replaced by additive and multiplicative identities, respectively, if the limits of their indices overlap, and matrix products are ordered by descending index) holds for some n. Then, (3.320) yields

$$\log \zeta_{n+1}(\underline{s}) = \log \zeta_n (B_n^{\mathrm{T}} \underline{s}) + \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) \, A_n^{\mathrm{T}} \underline{s} \right] \tag{3.345}$$

$$\begin{aligned} &= i \, \underline{s}^{\mathrm{T}} B_n \left[\prod_{j=1}^{n-1} B_j \right] (\underline{T}_1^R - \underline{T}^*) \\ &\quad + \sum_{j=1}^{n-1} \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} B_n \left[\prod_{k=j+1}^{n-1} B_k \right] A_j \, \Sigma(\underline{T}^*) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}} B_n^{\mathrm{T}} \underline{s} \right] \\ &\quad + \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(\underline{T}^*) \, A_n^{\mathrm{T}} \underline{s} \right] \end{aligned} \tag{3.346}$$

$$\begin{aligned} &= i \, \underline{s}^{\mathrm{T}} \left[\prod_{j=1}^{n} B_j \right] (\underline{T}_1^R - \underline{T}^*) \\ &\quad + \sum_{j=1}^{n} \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right] A_j \, \Sigma(\underline{T}^*) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right]^{\mathrm{T}} \underline{s} \right], \end{aligned} \tag{3.347}$$

establishing (3.344) for all n. But since $\| B_n \| < 1$ w.p.1 for large enough n (from (3.256) and (3.319), and the monotonicity by hypothesis of $\underline{\xi}(\underline{T})$ in a neighborhood of $\underline{T}^*$, which implies that $J(\underline{T}^*) < 0$), it follows that

$$\lim_{n \to \infty} \prod_{j=1}^{n} B_j = 0, \tag{3.348}$$

so that the first term in (3.344) vanishes as $n \to \infty$. Moreover,

$$\begin{aligned} &\log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right] A_j \, \Sigma(\underline{T}^*) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right]^{\mathrm{T}} \underline{s} \right] \\ &\qquad = \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right] A_j \, \Sigma(\underline{T}^*) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right]^{\mathrm{T}} \underline{s} \end{aligned} \tag{3.349}$$

approximately (using a first-order Taylor expansion for the logarithm) at least for large n, since (3.348) and (3.256) imply that the term on the right-hand side of (3.349) vanishes as $n \to \infty$. Thus,

$$\lim_{n \to \infty} \zeta_n(\underline{s}) - e^{-\frac{1}{2}\underline{s}^{\mathrm{T}} \left[\sum_{j=1}^{n-1} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j \, \Sigma(\underline{T}^*) A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}} \right] \underline{s}} = 0, \tag{3.350}$$

i.e. $\zeta_n(\underline{s})$ asymptotically has the form of the characteristic function of a normal distribution (e.g. Kendall and Stuart, 1977, vol.1, p.62). Since, by uniqueness, the limit of the characteristic functions of a sequence of distributions is the characteristic function of the limiting distribution (see, for instance, Loève, 1963, pp.189-193), it has thus been shown that $(\underline{T}_n^R - \underline{T}^*)$ is asymptotically normal, and there only remains to prove that its covariance approaches that in (3.257).

Finding the limit of the exponent in (3.350) is not trivial. Consider instead the recursion

$$Q_{n+1} = \left[I + A_n \, J(\underline{T}^*) + o_p(n^{-1}) \right] Q_n \left[I + A_n \, J(\underline{T}^*) + o_p(n^{-1}) \right]^{\mathrm{T}} + A_n \, \Sigma(\underline{T}_n^R) \, A_n^{\mathrm{T}} \tag{3.351}$$

(from (3.300) with (3.230) and (3.301), noting that the cross term vanishes), where

$$Q_n := E\left[(\underline{T}_n^R - \underline{T}^*)(\underline{T}_n^R - \underline{T}^*)^{\mathrm{T}} \right]. \tag{3.352}$$

Note first that by (3.256), there is for any $\delta > 0$ a large enough $n(\delta)$ such that

$$\| \, n \, A_n - A \, \| < \delta \tag{3.353}$$

for all $n > n(\delta)$. It follows that

$$n^2 A_n \, \Sigma(\underline{T}_n^R) \, A_n^{\mathrm{T}} = A \, \Sigma(\underline{T}_n^R) \, A^{\mathrm{T}} + O(\delta) \tag{3.354}$$

$$= A \, \Sigma(\underline{T}^*) \, A^{\mathrm{T}} + O(\delta) + o(1) \tag{3.355}$$

w.p.1 for $n > n(\delta)$, where (3.355) holds by virtue of the continuity of $\Sigma(\underline{T})$ in a neighborhood of $\underline{T}^*$ and the convergence w.p.1 of $\underline{T}_n^R$ to $\underline{T}^*$. Similarly,

$$n \, A_n \, J(\underline{T}^*) = A \, J(\underline{T}^*) + O(\delta) \tag{3.356}$$

for $n > n(\delta)$, so that (3.351) can be rewritten as

$$Q_{n+1} = \left[I + n^{-1} \left(A \, J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right] Q_n \left[I + n^{-1} \left(A \, J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right]^{\mathrm{T}} + n^{-2} \left[A \, \Sigma(\underline{T}^*) \, A^{\mathrm{T}} + O(\delta) + o(1) \right] \tag{3.357}$$

for $n > n(\delta)$. To enable a coordinatewise application of Lemma 3.4, it is necessary to diagonalize the matrices pre- and post-multiplying Q_n in (3.357). To this end, define

$$\tilde{Q}_n := V^{\mathrm{T}} Q_n \, V, \tag{3.358}$$

where V is defined in (3.259), and substitute into (3.357):

$$V \tilde{Q}_{n+1} V^{\mathrm{T}} = \left[I + n^{-1} \left(A \, J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right] V \tilde{Q}_n \, V^{\mathrm{T}} \left[I + n^{-1} \left(A \, J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right]^{\mathrm{T}}$$

$$+ \; n^{-2} \left[A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} + O(\delta) + o(1) \right], \tag{3.359}$$

or (by the orthogonality of V)

$$\begin{aligned} \tilde{Q}_{n+1} \; = \; & V^{\mathrm{T}} \left[I + n^{-1} \left(A \; J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right] V \; \tilde{Q}_n \\ & \qquad V^{\mathrm{T}} \left[I + n^{-1} \left(A \; J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right]^{\mathrm{T}} V \\ & + \; n^{-2} \, V^{\mathrm{T}} \left[A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} + O(\delta) + o(1) \right] V. \end{aligned} \tag{3.360}$$

But

$$V^{\mathrm{T}} \left[I + n^{-1} \left(A \; J(\underline{T}^*) + O(\delta) \right) + o_p(n^{-1}) \right] V \; = \; I - n^{-1} \left[\Lambda + O(\delta) + o_p(1) \right] \tag{3.361}$$

by (3.259) and the orthogonality of V, and

$$V^{\mathrm{T}} \left[A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} + O(\delta) + o(1) \right] V \; = \; V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} \, V + O(\delta) + o(1), \tag{3.362}$$

whence it follows that

$$\begin{aligned} \tilde{Q}_{n+1} \; = \; & \left[I - n^{-1} \left(\Lambda + O(\delta) + o_p(1) \right) \right] \tilde{Q}_n \; \left[I - n^{-1} \left(\Lambda + O(\delta) + o_p(1) \right) \right]^{\mathrm{T}} \\ & + \; n^{-2} \left[V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} \, V + O(\delta) + o(1) \right]. \end{aligned} \tag{3.363}$$

Consider now the element $\tilde{q}_{ij}^{(n)}$ of $\tilde{Q}_n$: equation (3.363) yields

$$\begin{aligned} \tilde{q}_{ij}^{(n+1)} \; = \; & \left[1 - n^{-1} \left(\lambda_i + O(\delta) + o_p(1) \right) \right] \left[1 - n^{-1} \left(\lambda_j + O(\delta) + o_p(1) \right) \right] \tilde{q}_{ij}^{(n)} \\ & + \; n^{-2} \left[\left(O(\delta) + o_p(1) \right)^2 \; \| \; \tilde{Q}_n \; \| \right. \\ & \qquad \left. + \; [\, V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} \, V \,]_{ij} + O(\delta) + o(1) \right]. \end{aligned} \tag{3.364}$$

But

$$\begin{aligned} & \left[1 - n^{-1} \left(\lambda_i + O(\delta) + o_p(1) \right) \right] \left[1 - n^{-1} \left(\lambda_j + O(\delta) + o_p(1) \right) \right] \\ & \qquad = \; \left[1 - n^{-1} \left(\lambda_i + \lambda_j + O(\delta) + o_p(1) + O(n^{-1}) \right) \right] \end{aligned} \tag{3.365}$$

and, letting $\delta \downarrow 0$,

$$\lim_{n \to \infty} \; \lambda_i + \lambda_j + O(\delta) + o_p(1) + O(n^{-1}) \; = \; \lambda_i + \lambda_j \; > \; 1, \tag{3.366}$$

where the inequality follows from that in (3.256), which implies that

$$-A \; J(\underline{T}^*) \; > \; \frac{1}{2} \, I, \tag{3.367}$$

or

$$\Lambda \; = \; -V \, A \; J(\underline{T}^*) \, V^{\mathrm{T}} \tag{3.368}$$

$$> \; \frac{1}{2} \, I \tag{3.369}$$

by the orthogonality of V. Moreover, again letting $\delta \downarrow 0$,

$$\lim_{n \to \infty} \; (O(\delta) + o_p(1))^2 \, \| \, \tilde{Q}_n \, \| + [\, V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} V \,]_{ij} + O(\delta) + o(1)$$

$$= [\, V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} V \,]_{ij} . \qquad (3.370)$$

(Note here that $\| \, \tilde{Q}_n \, \|$ is bounded because $\underline{T}_1^R < \infty$ by hypothesis, (3.255) holds, and $\| B_n \| < 1$ w.p.1 for large enough n.) Thus, by Lemma 3.4 (with $p = 1$),

$$\lim_{n \to \infty} \; n \; \tilde{q}_{ij}^{(n)} \; = \; \frac{[\, V^{\mathrm{T}} A \; \Sigma(\underline{T}^*) \, A^{\mathrm{T}} V \,]_{ij}}{\lambda_i + \lambda_j - 1} , \qquad (3.371)$$

and using (3.358) establishes (3.257)-(3.258). (The proof of consistency follows, with modifications, that of Blum, 1954b; on convergence proofs for stochastic approximation procedures *via* martingales, see also Métivier, 1982, pp. 66-72, 75; the proof of asymptotic normality follows, with modifications, that of Fabian, 1968; both results are also discussed by Wasan, 1969, pp.77-79, 106-110.) ■

Remark It is easy to verify that the results of Theorem 3.2 agree in the scalar case with those of Theorem 3.1. Since $J(\underline{T}^*) = \xi'(T^*)$, $P = 1$, and $\Lambda = - \, a \;\, \xi'(T^*)$, it follows that

$$V \, Q \, V^{\mathrm{T}} \; = \; - \, \frac{a^2 \; \Sigma(T^*)}{2 \, a \;\, \xi'(T^*) + 1} \qquad (3.372)$$

in the scalar case, which is the expression in (3.43). Note, however, that some earlier conditions (most notably (3.7)) are no longer necessary in this proof; while they are not unduly restrictive, as discussed earlier, they will henceforth not be required to hold.

Corollary 3.4 Theorem 3.2 holds also if $\underline{T}_1^R$ is a random variable, provided that

$$E \left[\; \underline{T}_1^R \; (\underline{T}_1^R)^{\mathrm{T}} \; \right] \; < \; \infty , \qquad (3.373)$$

and $\underline{T}_1^R$ is independent of $\underline{x}_n$, $n = 2, 3, \cdots$. If, moreover, $\underline{T}_1^R$ is a translation-invariant function of $\underline{x}_1$, then $\underline{T}_n^R$ is translation invariant.

Proof The proof of the first part of the corollary follows that of Theorem 3.2 identically. In the proof of consistency, the condition of independence is required for equations (3.266), (3.270), and (3.282) to hold. Furthermore, the product $(\underline{T}_1^R - \underline{T}^*)^{\mathrm{T}} (\underline{T}_1^R - \underline{T}^*)$ is replaced by its expectation in (3.262), (3.274), (3.286), and (3.287), and (3.373) is required for (3.274) and (3.288) (in their modified form) to hold. In the proof of asymptotic normality, independence is required for (3.309) and (3.329) to hold. The exponential in (3.321) and (3.342) is replaced by its expectation, while the first terms on the right-hand sides of equations (3.343), (3.344), (3.346), and (3.347) are replaced by the logarithm of the expectation of their exponentials. Equation (3.373) is required for the first term in (3.344) to vanish as $n \to \infty$. The proof of translation invariance is identical with that in Corollary 3.1, and is omitted. ■

Corollary 3.5 Under the conditions of Theorem 3.2, the recursive minimax robust estimator $\underline{T}_n^R$ has minimum asymptotic variance for the choice

$$A^* = -J^{-1}(\underline{T}^*), \tag{3.374}$$

in which case

$$\mathbf{L}(\sqrt{n}\ (\underline{T}_n^R - \underline{T}^*)) \rightarrow \mathbf{N}\left[0,\ J^{-1}(\underline{T}^*)\,\Sigma(\underline{T}^*)\,(J^{-1}(\underline{T}^*))^{\mathrm{T}}\right]. \tag{3.375}$$

Proof Let λ denote the largest diagonal element of the matrix Λ defined in (3.259), and note that from (3.258),

$$V\,Q\,V^{\mathrm{T}} \geq \frac{1}{2\,\lambda - 1}\ V\,V^{\mathrm{T}} A\ \Sigma(\underline{T}^*)\,A^{\mathrm{T}}\,V\,V^{\mathrm{T}} \tag{3.376}$$

$$= \frac{1}{2\,\lambda - 1}\ A\ \Sigma(\underline{T}^*)\,A^{\mathrm{T}} \tag{3.377}$$

$$= \frac{1}{2\,\lambda - 1}\ (-A\ J(\underline{T}^*))\,J^{-1}(\underline{T}^*)\,\Sigma(\underline{T}^*)\,(J^{-1}(\underline{T}^*))^{\mathrm{T}}(-A\ J(\underline{T}^*))^{\mathrm{T}}, \tag{3.378}$$

where (3.377) follows from the orthogonality of V. Since λ is the largest eigenvalue of $A\ J(\underline{T}^*)$, the infimum of (3.378) may be found in two steps. Defining

$$X := -A\ J(\underline{T}^*), \tag{3.379}$$

$$X' := \frac{1}{\lambda}\ X, \tag{3.380}$$

and

$$M := J^{-1}(\underline{T}^*)\,\Sigma(\underline{T}^*)\,(J^{-1}(\underline{T}^*))^{\mathrm{T}}, \tag{3.381}$$

the following two problems may be solved independently:

$$\inf_{X'}\ X'\,M\,X'^{\mathrm{T}} \tag{3.382}$$

subject to

$$\|\,X'\,\| = 1, \tag{3.383}$$

(from (3.380)) and

$$X' > \alpha\,I \tag{3.384}$$

for some $\alpha > 0$ (from (3.367)); and

$$\inf_{\lambda}\ \frac{\lambda^2}{2\,\lambda - 1}. \tag{3.385}$$

Since (3.382) is in quadratic form with $M > 0$, the infimum exists and occurs at the constrained infimum of X'; combining (3.383) and (3.384), it follows that

$$X' = I \tag{3.386}$$

minimizes (3.382). In the case of (3.385), on the other hand,

$$\frac{d}{d\,\lambda}\ \frac{\lambda^2}{2\,\lambda - 1} = \frac{2\,\lambda\,(\,2\,\lambda - 1\,)\ -\ 2\,\lambda^2}{(\,2\,\lambda - 1\,)^2} \tag{3.387}$$

$$= 0 \tag{3.388}$$

at the extremum, yielding the solution

$$\lambda = 1. \tag{3.389}$$

Note that this solution satisfies (3.369), and is the only solution that does so. Moreover,

$$\frac{d^2}{d\lambda^2}\ \frac{\lambda^2}{2\lambda - 1} = \frac{2}{(2\lambda - 1)^3} \tag{3.390}$$

$$> 0 \tag{3.391}$$

for all λ satisfying (3.369), confirming that (3.389) corresponds to a global minimum.

It follows from (3.386) and (3.389) that $X = I$, implying (3.374). Moreover, substituting (3.374) into (3.257)-(3.259) yields $V = \Lambda = I$, so that

$$V\,Q\,V^{\mathrm{T}} = J^{-1}(\underline{T}^*)\,\Sigma(\underline{T}^*)\,(J^{-1}(\underline{T}^*))^{\mathrm{T}}. \tag{3.392}$$

Thus, the lower bound on (3.378) is in fact achieved by the choice (3.374), establishing (3.375) and completing the proof. ■

Definition 3.1 In the multivariate case, the Fisher information matrix of the density $f_{\underline{\theta}}(\underline{x})$ at $\underline{\theta}$, $\underline{\theta} \in \Theta$, is defined as

$$I(f_{\underline{\theta}}) = E_{f_{\underline{\theta}}}\left[\left[\underline{\nabla}_{\underline{\theta}} \log f_{\underline{\theta}}(\underline{x}) \right] \left[\underline{\nabla}_{\underline{\theta}} \log f_{\underline{\theta}}(\underline{x}) \right]^{\mathrm{T}} \right] \tag{3.393}$$

$$= E_{f_{\underline{\theta}}}\left[\left[\frac{1}{f_{\underline{\theta}}(\underline{x})}\ \underline{\nabla}_{\underline{\theta}} f_{\underline{\theta}}(\underline{x}) \right] \left[\frac{1}{f_{\underline{\theta}}(\underline{x})}\ \underline{\nabla}_{\underline{\theta}} f_{\underline{\theta}}(\underline{x}) \right]^{\mathrm{T}} \right], \tag{3.394}$$

provided these expressions exist.

Corollary 3.6 For a given family of symmetric distributions with location parameter $\underline{\theta}$, let the least favorable distribution $f_{\underline{\theta}}(\underline{x}) = f(\underline{x} - \underline{\theta})$ be such that the corresponding influence-bounding function $\underline{\psi}(\underline{x} - \underline{\theta})$ satisfies the conditions of Theorem 3.2. Let $\underline{T}_n^R$ be the recursive minimax robust estimator of $\underline{\theta}$ defined by equation (3.228), with coefficients $\{A_n\}$ satisfying the conditions of Theorem 3.2 as well as (3.374). If the true underlying distribution is $f_{\underline{\theta}^*}(\underline{x})$, then

$$\mathbf{L}(\sqrt{n}\ (\underline{T}_n^R - \underline{\theta}^*)) \rightarrow \mathbf{N}\left[0,\ I^{-1}(f_{\underline{\theta}^*})\right] \tag{3.395}$$

as $n \rightarrow \infty$ (i.e. $\underline{T}_n^R$ is *asymptotically efficient*). In that case,

$$A^* = I^{-1}(f_{\underline{\theta}^*}). \tag{3.396}$$

Proof In analogy with equations (2.131)-(2.135), (3.231) yields

$$J_{ij}(\underline{\theta}^*) = \frac{\partial}{\partial t_j}\ \xi_i(\underline{t})\ \Big|_{\underline{t} = \underline{\theta}^*} \tag{3.397}$$

$$= \frac{\partial}{\partial t_j} \int \psi_i(\underline{x}-\underline{t})\, f(\underline{x}-\underline{\theta}^*)\, d\underline{x} \,\Big|_{\underline{t}=\underline{\theta}^*} \tag{3.398}$$

$$= \frac{\partial}{\partial t_j} \int \psi_i(\underline{x}-\underline{\theta}^*)\, f(\underline{x}+\underline{t}-2\underline{\theta}^*)\, d\underline{x} \,\Big|_{\underline{t}=\underline{\theta}^*} \tag{3.399}$$

$$= \int \psi_i(\underline{x}-\underline{\theta}^*)\, \frac{\partial}{\partial t_j} f(\underline{x}+\underline{t}-2\underline{\theta}^*) \,\Big|_{\underline{t}=\underline{\theta}^*} d\underline{x}, \tag{3.400}$$

where (3.398) follows from (3.229), (3.399) from a change of variable (with $\underline{x}$ replaced by $\underline{x} - \underline{\theta}^* + \underline{t}$), and (3.400) holds by virtue of the Lebesgue dominated convergence theorem, provided that $f_{\underline{\theta}}$ is bounded and differentiable in a neighborhood of $\underline{\theta}^*$. Postponing for a moment the justification of this step, (3.400) yields

$$J_{ij}(\underline{\theta}^*) = \int \left[-\frac{1}{f(\underline{x}-\underline{\theta}^*)} \; \frac{\partial}{\partial t_i} f(\underline{x}-\underline{t}) \,\Big|_{\underline{t}=\underline{\theta}^*} \right] \left[\frac{1}{f(\underline{x}-\underline{\theta}^*)} \; \frac{\partial}{\partial t_j} f(\underline{x}+\underline{t}-2\underline{\theta}^*) \,\Big|_{\underline{t}=\underline{\theta}^*} \right] f(\underline{x}-\underline{\theta}^*)\, d\underline{x} \tag{3.401}$$

a.s. from (3.227), and a comparison with (3.394) establishes that

$$J(\underline{\theta}^*) = -I(f_{\underline{\theta}^*}). \tag{3.402}$$

Thus, the restriction $\|J(\underline{\theta}^*)\| < \infty$ (imposed by Theorem 3.2) implies finite Fisher information and hence bounded and differentiable $f_{\underline{\theta}}$, justifying (3.400).

Combining (3.374) and (3.402) establishes (3.397). Moreover, in the special case when the underlying distribution is indeed the least favorable one, a comparison of (3.227), (3.230), and (3.394) reveals that

$$\Sigma(\underline{\theta}^*) = I(f_{\underline{\theta}^*}), \tag{3.403}$$

so that (3.375) reduces to (3.395), proving the assertion. ■

Theorem 3.2 and Corollaries 3.4-3.6 show that all the desirable properties of the recursive minimax robust estimator extend to the multivariate case. Given a least favorable distribution with location parameter $\underline{\theta}$, the vector-valued influence-bounding function $\underline{\psi}(\underline{x}-\underline{\theta})$ can be found from (3.226)-(3.227), and the resulting estimator $\underline{T}_n^R$ is consistent and asymptotically normal under fairly weak conditions. There moreover exists a choice of A minimizing the asymptotic variance, and this choice results in an asymptotically efficient estimator when the true underlying distribution is the least favorable one.

There is nevertheless one problem. So far, in the present section, matrices were ordered in the usual way -- specifically, 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. This is not a lattice ordering, however. 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.

Huber touches on the multivariate case only very briefly (1972; 1977, p.35; 1981, pp.211, 222-223). He proposes to consider spherically symmetric distributions, and to apply non-degenerate affine transformations of the form $\underline{x} \rightarrow W(\underline{x} + \underline{\theta})$ to obtain parametric families of "elliptic" distributions. This, however, brings forth the problem of determining W when, as is usually the case, it is not known *a priori*. Indeed, nothing has so far been said about the problem of "scale estimation": Sections 2.2 and 3.1 assumed unit (or known) scale -- an assumption implicit, for instance, in the definition of the ε-contaminated normal neighborhood as a set of perturbations of the standard normal distribution. Huber addresses the issue of simultaneous location and scale estimation in the scalar case, and also offers some methods for estimating W (Huber, 1981, pp.215-223). This problem is resolved later, in Section 4.4, where an estimated covariance based on theory is used; for the present, it is assumed that W is known.

Given the measure space ($\mathbf{X}$, $\mathbf{B}$, μ) as before, let $\{ \underline{x}_1, \cdots, \underline{x}_n \}$ be a sample of independent random variates taking values in $\mathbf{X}$, with a common spherically symmetric distribution function P; let $\mathbf{P} := \{ P_{\underline{\theta}} : \underline{\theta} \in \Theta \}$, Θ as before, be a family of spherically symmetric probability measures on ($\mathbf{X}$, $\mathbf{B}$) such that for all $\underline{\theta} \in \Theta$, $P_{\underline{\theta}}$ is absolutely continuous with respect to μ and admits the density $f_{\underline{\theta}}$ in accordance with the Radon-Nikodym theorem. Define the linear transformation

$$\underline{z}_n := W \underline{x}_n, \tag{3.404}$$

$n = 1, 2, \cdots$, where $W \in \mathbf{R}^{q \times q}$ is a known matrix, with $W > 0$.

Let $f^*_{\underline{\theta}} \in \mathbf{P}$ be the least favorable distribution, and consider the recursion

$$\underline{T}^W_{n+1} = \underline{T}^W_n + W A_n \underline{\psi}\left[W^{-1} (\underline{z}_n - \underline{T}^W_n) \right], \tag{3.405}$$

$n = 1, 2, \cdots$, where $\{A_n\}$ is a given matrix sequence with $A_n \in \mathbf{R}^{q \times q}$, $\underline{T}^W_1$ is an arbitrary (possibly random) starting point, and $\underline{\psi}(\underline{x} - \underline{\theta})$ is related to $f^*_{\underline{\theta}}(\underline{x}) = f^*(\underline{x} - \underline{\theta})$ through equations (3.226)-(3.227). Note in passing that the W matrix premultiplying A_n is there primarily for purposes of normalization, and that similar ideas are used in Masreliez and Martin (1974, 1977) to design a one-step multivariate robust estimator. The following result holds:

Corollary 3.7 Under the consistency conditions of Theorem 3.2, $\underline{T}^W_n \rightarrow W \underline{T}^*$ as $n \rightarrow \infty$ a.s. Under the asymptotic normality conditions of Theorem 2.7,

$$\mathbf{L}(\sqrt{n} \; (\underline{T}^W_n - W \underline{T}^*)) \rightarrow \mathbf{N}\left[0, \; W V Q V^{\mathrm{T}} W^{\mathrm{T}} \right], \tag{3.406}$$

where Q and V are defined by equations (3.258)-(3.259).

Proof Premultiplying equation (3.405) by W^{-1} yields

$$W^{-1} \underline{T}^W_{n+1} = W^{-1} \underline{T}^W_n + A_n \underline{\psi}\left[W^{-1} (\underline{z}_n - \underline{T}^W_n) \right], \tag{3.407}$$

or, defining

$$\tilde{\underline{T}}_n := W^{-1} \underline{T}_n^W, \tag{3.408}$$

it follows that

$$\tilde{\underline{T}}_{n+1} = \tilde{\underline{T}}_n + A_n \, \underline{\psi}(W^{-1} \underline{z}_n - \tilde{\underline{T}}_n). \tag{3.409}$$

From (3.404) and by hypothesis, $\{\underline{z}_n\}$ is a sample of independent random variates with a common distribution function, say P_z, so that

$$P_z(\underline{z}) = P(\underline{x}). \tag{3.410}$$

Thus,

$$E_{P_z}\left[\underline{\psi}(W^{-1} \underline{z} - \underline{T}) \right] = E_P \left[\underline{\psi}(\underline{x} - \underline{T}) \right] \tag{3.411}$$

$$= \underline{\xi}(\underline{T}), \tag{3.412}$$

where (3.411) follows from (3.404) and (3.410), and (3.412) from (3.229). A similar argument establishes that

$$E_{P_z}\left[\underline{\psi}(W^{-1} \underline{z} - \underline{T}) \, \underline{\psi}^{\mathrm{T}}(W^{-1} \underline{z} - \underline{T}) \right] \leq S_0 \tag{3.413}$$

for some $S_0 < \infty$, from (3.253). Thus, under the conditions of Theorem 3.2, $\tilde{\underline{T}}_n \to \underline{T}^*$ as $n \to \infty$ a.s., whence it follows by (3.408) that $\underline{T}_n^W \to W \underline{T}^*$ as $n \to \infty$ a.s.

An argument similar to (3.411)-(3.412) also establishes that

$$E_{P_z}\left[(\underline{\psi}(W^{-1} \underline{z} - \underline{T}) - \underline{\xi}(\underline{T})) \, (\underline{\psi}(W^{-1} \underline{z} - \underline{T}) - \underline{\xi}(\underline{T}))^{\mathrm{T}} \right] = \Sigma(\underline{T}), \tag{3.414}$$

so that under the conditions of Theorem 3.2,

$$\mathbf{L}(\sqrt{n} \, (\tilde{\underline{T}}_n - \underline{T}^*)) \to \mathbf{N}\left[0, \ V \, Q \, V^{\mathrm{T}} \right], \tag{3.415}$$

where Q and V are given by (3.258)-(3.259), and (3.408) establishes (3.406), completing the proof. ■

Clearly, statements analogous to Corollaries 3.4-3.6 can be made for the recursion (3.405). Thus, in the absence of knowledge of the least favorable distribution in an arbitrary neighborhood of probability measures, the multivariate minimax robust estimation problem can still be solved at least if the observation can be expressed as a linear transformation of a random variable with a spherically symmetric distribution.

3.3 The Time-Variant Case

So far, only the time-invariant case has been addressed: the sample of observations was assumed to be not only independent but also identically distributed, and the location parameter of the common distribution function was sought.

A generalization of these results concerns the case where the parameter to be estimated changes over time. This has been considered by Burkholder (1956) and Fabian (1968), both of whom analyze the case where the observations are not distributed identically but according to a *converging sequence* of probability distributions. An alternative model is one in which the sequence of distributions does not approach a limit, but is a known time function of an unknown but constant parameter. This case is discussed first in the present section.

Let $(\mathbf{X}, \mathbf{B}, \mu)$ be a measure space, as before, and let $\{ \underline{v}_1, \cdots, \underline{v}_n \}$ be a sample of independent random variates taking values in $\mathbf{X}$, with a common spherically symmetric distribution function P centered at the origin; let $\mathbf{P} := \{ P_{\underline{\theta}} : \underline{\theta} \in \Theta \}$, Θ as before, be a family of spherically symmetric probability measures on $(\mathbf{X}, \mathbf{B})$ such that for all $\underline{\theta} \in \Theta$, $P_{\underline{\theta}}$ is absolutely continuous with respect to μ and admits the density $f_{\underline{\theta}}$ in accordance with the Radon-Nikodym theorem. Define the transformation

$$\underline{z}_n := \underline{\theta}_n + \underline{v}_n, \tag{3.416}$$

$n = 1, 2, \cdots$, where

$$\underline{\theta}_{n+1} = F_n \, \underline{\theta}_n, \tag{3.417}$$

$\{F_n\}$ is a known sequence of non-singular matrices with $F_n \in \mathbf{R}^{q \times q}$, and $\underline{\theta}_0$ is an unknown (but fixed and finite) parameter.

Let $f^*_{\underline{\theta}} \in \mathbf{P}$ be the least favorable distribution, and consider the recursion

$$\underline{T}^F_{n+1} = F_n \, \underline{T}^F_n + A_n \, \underline{\psi}(\underline{z}_{n+1} - F_n \, \underline{T}^F_n), \tag{3.418}$$

$n = 0, 1, \cdots$, where $\{A_n\}$ is a given matrix sequence with $A_n \in \mathbf{R}^{q \times q}$, $\underline{T}^F_0$ is an arbitrary (possibly random) starting point, and $\underline{\psi}(\underline{v} - \underline{\theta})$ is related to $f^*_{\underline{\theta}}(\underline{v}) = f^*(\underline{v} - \underline{\theta})$ through equations (3.226)-(3.227). Define $\xi(\underline{T})$, $\Sigma(\underline{T})$, and $J(\underline{T})$ as in (3.229)-(3.231), provided these expressions exist. The following is a generalization of Theorem 3.2.

Theorem 3.3 Let $\xi(\underline{T})$ exist for all $\underline{T}$, and for any $\delta > 0$ and all $q \times q$ matrices $M > 0$, let

$$\sup_{\delta \le \|\underline{T}\|} \underline{T}^{\mathrm{T}} M \, \xi(\underline{T}) < 0. \tag{3.419}$$

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

$$E_P \left[\underline{\psi}(\underline{v} - \underline{T}) \, \underline{\psi}^{\mathrm{T}}(\underline{v} - \underline{T}) \right] \le S_0 \tag{3.420}$$

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

$$\sum_{n=1}^{\infty} A_n = \infty, \tag{3.421}$$

and

$$\sum_{n=1}^{\infty} A_n^{\mathrm{T}} A_n < \infty. \tag{3.422}$$

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

$$\left[\prod_{j=m}^{n} F_j \right]^{\mathrm{T}} \left[\prod_{j=m}^{n} F_j \right] < \alpha I \tag{3.423}$$

(where products are ordered by descending index), then, given any $\underline{T}_0^F < \infty$, $\underline{T}_n^F - \underline{\theta}_n \to 0$ as $n \to \infty$ a.s. (i.e. $\underline{T}_n^F$ 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

$$\limsup_{n\to\infty} \; n \, A_n \; < \; \infty, \tag{3.424}$$

then

$$\mathrm{L}(\, \Sigma_n^{-1/2} \, (\underline{T}_n^F - \underline{\theta}_n \,) \,) \;\to\; \mathrm{N}(\, 0, I \,), \tag{3.425}$$

where

$$\Sigma_n \;=\; \left[I + A_{n-1} \, J(0) \right] F_{n-1} \, \Sigma_{n-1} \, F_{n-1}^{\mathrm{T}} \; \left[I + A_{n-1} \, J(0) \right]^{\mathrm{T}} \;+\; A_{n-1} \, \Sigma(0) \, A_{n-1}^{\mathrm{T}} \tag{3.426}$$

with

$$\Sigma_0 \;=\; 0 \tag{3.427}$$

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

Proof The proof follows in part that of Theorem 3.2, and some intermediate steps are omitted for brevity. Note first that, denoting the distribution of the observation $\underline{z}$ at time n by P_n, (3.416) implies

$$P_n(\, \underline{z} \,) \;=\; P(\, \underline{v} \,), \tag{3.428}$$

so that

$$E_{P_n} \left[\, \underline{\psi}(\underline{z} - \underline{T} \,) \, \right] \;=\; E_P \left[\, \underline{\psi}(\, \underline{v} + \underline{\theta}_n - \underline{T} \,) \, \right] \tag{3.429}$$

$$= \; \underline{\xi}(\, \underline{T} - \underline{\theta}_n \,). \tag{3.430}$$

A similar argument establishes that

$$E_{P_n} \left[\, \underline{\psi}(\underline{z} - \underline{T} \,) \, \underline{\psi}^{\mathrm{T}}(\underline{z} - \underline{T} \,) \, \right] \;\le\; S_0 \tag{3.431}$$

for all $\underline{T}$, from (3.420).

Rewriting (3.418) as

$$\underline{T}_{n+1}^F \;-\; \underline{\theta}_{n+1} \;=\; F_n \, \underline{T}_n^F \;-\; \underline{\theta}_{n+1} \;+\; A_n \, \underline{\psi}(\, \underline{z}_{n+1} - F_n \, \underline{T}_n^F \,), \tag{3.432}$$

it follows, upon squaring and taking expectations, that

$$\begin{aligned} E\left[\, (\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1} \,)^{\mathrm{T}} \, (\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1} \,) \, \right] & \\ = E\left[\, (\, \underline{T}_n^F - \underline{\theta}_n \,)^{\mathrm{T}} \, F_n^{\mathrm{T}} \, F_n \, (\, \underline{T}_n^F - \underline{\theta}_n \,) \, \right] & \\ + \; 2 \, E\left[\, (\, F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \,)^{\mathrm{T}} \, A_n \, \underline{\psi}(\, \underline{z}_{n+1} - F_n \, \underline{T}_n^F \,) \, \right] & \\ + \; E\left[\, \underline{\psi}^{\mathrm{T}}(\, \underline{z}_{n+1} - F_n \, \underline{T}_n^F \,) \, A_n^{\mathrm{T}} \, A_n \, \underline{\psi}(\, \underline{z}_{n+1} - F_n \, \underline{T}_n^F \,) \, \right] & \end{aligned} \tag{3.433}$$

$$
\begin{aligned}
&= (\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} \left[\prod_{j=0}^{n} F_j \right]^{\mathrm{T}} \left[\prod_{j=0}^{n} F_j \right] (\underline{T}_0^F - \underline{\theta}_0) \\
&\quad + 2 \sum_{j=0}^{n} E\Bigg[(F_j\, \underline{T}_j^F - \underline{\theta}_{j+1})^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \\
&\qquad\qquad \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \Bigg] \\
&\quad + \sum_{j=0}^{n} E\Bigg[\underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \\
&\qquad\qquad \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \Bigg] \qquad (3.434)
\end{aligned}
$$

(products are replaced by the multiplicative identity if the limits of their indices overlap, and are ordered by descending index), where use is made of (3.417) in (3.433). Noting that (3.434) is scalar, an argument analogous to (3.263)-(3.264) yields

$$
\begin{aligned}
&E\left[\underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \right] \\
&\qquad = \mathrm{tr}\Bigg[E\Bigg[\underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F)\, \underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \Bigg] \\
&\qquad\qquad A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j \Bigg] \qquad (3.435) \\
&\qquad = \mathrm{tr}\Bigg[E\Bigg[E_{P_{j+1}}\Bigg[\underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F)\, \underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \;\Big|\; \underline{T}_j^F \Bigg] \Bigg] \\
&\qquad\qquad A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j \Bigg] \qquad (3.436) \\
&\qquad \le \alpha\, \mathrm{tr}\left[S_0\, A_j^{\mathrm{T}} A_j \right] \qquad (3.437)
\end{aligned}
$$

w.p.1 for all j, where (3.437) follows from (3.423) and (3.431). Thus,

$$
\begin{aligned}
&\sum_{j=0}^{n} E\left[\underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \right] \\
&\qquad \le \alpha\, \mathrm{tr}\left[S_0 \sum_{j=0}^{n} A_j^{\mathrm{T}} A_j \right] \qquad (3.438) \\
&\qquad < \infty \qquad (3.439)
\end{aligned}
$$

w.p.1 as $n \to \infty$, from (3.422), the finiteness by hypothesis of S_0, and (3.423). Moreover,

$$
E\left[(F_j\, \underline{T}_j^F - \underline{\theta}_{j+1})^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \right]
$$

$$= E\left[E_{P_{j+1}}\left[(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,)^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\Psi}(\, \underline{z}_{j+1} - F_j\, \underline{T}_j^F\,)\ \middle|\ \underline{T}_j^F \right]\right] \tag{3.440}$$

$$= E\left[(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,)^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\xi}(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,) \right] \tag{3.441}$$

$$< 0 \tag{3.442}$$

w.p.1 for all j, from (3.419), the non-negativity of the quadratic product, and that of A_j by hypothesis. Thus,

$$\sum_{j=0}^{n} E\left[(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,)^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\Psi}(\, \underline{z}_{j+1} - F_j\, \underline{T}_j^F\,) \right] < 0 \tag{3.443}$$

w.p.1 for all n. But since

$$E\left[(\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1}\,)^{\mathrm{T}} (\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1}\,) \right] \geq 0 \tag{3.444}$$

because the term is in quadratic form, and

$$(\, \underline{T}_0^F - \underline{\theta}_0\,)^{\mathrm{T}} \left[\prod_{j=0}^{n} F_j \right]^{\mathrm{T}} \left[\prod_{j=0}^{n} F_j \right] (\, \underline{T}_0^F - \underline{\theta}_0\,) \leq \alpha\, (\, \underline{T}_0^F - \underline{\theta}_0\,)^{\mathrm{T}} (\, \underline{T}_0^F - \underline{\theta}_0\,) \tag{3.445}$$

$$< \infty \tag{3.446}$$

as $n \to \infty$, since $\underline{T}_0 < \infty$ by hypothesis and from (3.423), it follows that

$$\sum_{j=0}^{n} E\left[(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,)^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\Psi}(\, \underline{z}_{j+1} - F_j\, \underline{T}_j^F\,) \right]$$
$$= \sum_{j=0}^{n} E\left[(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,)^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right]^{\mathrm{T}} \left[\prod_{k=j+1}^{n} F_k \right] A_j\, \underline{\xi}(\, F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}\,) \right] \tag{3.447}$$

must be bounded from below w.p.1 as $n \to \infty$.

Defining

$$Y_n := E\left[(\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1}\,)^{\mathrm{T}} (\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1}\,) - (\, \underline{T}_n^F - \underline{\theta}_n\,)^{\mathrm{T}} (\, \underline{T}_n^F - \underline{\theta}_n\,)\ \middle|\ \underline{T}_1^F, \cdots, \underline{T}_n^F \right], \tag{3.448}$$

it can be shown in a manner analogous to (3.278)-(3.280) that the sequence

$$\left\{ (\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} (\underline{T}_n^F - \underline{\theta}_n) - \sum_{j=1}^{n-1} Y_j \right\} \tag{3.449}$$

is a martingale. Squaring (3.432) and taking conditional expectations,

$$\begin{aligned} E\Big[& (\underline{T}_{n+1}^F - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1}^F - \underline{\theta}_{n+1}) \mid \underline{T}_1^F, \cdots, \underline{T}_n^F \Big] \\ &= (\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} F_n^{\mathrm{T}} F_n (\underline{T}_n^F - \underline{\theta}_n) \\ &\qquad + 2\, E_{P_{n+1}}\Big[(F_n \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} A_n \underline{\psi}(\underline{z}_{n+1} - F_n \underline{T}_n^F) \\ &\qquad\qquad \mid \underline{T}_1^F, \cdots, \underline{T}_n^F \Big] \\ &\qquad + E_{P_{n+1}}\Big[\underline{\psi}^{\mathrm{T}}(\underline{z}_{n+1} - F_n \underline{T}_n^F) A_n^{\mathrm{T}} A_n \underline{\psi}(\underline{z}_{n+1} - F_n \underline{T}_n^F) \\ &\qquad\qquad \mid \underline{T}_1^F, \cdots, \underline{T}_n^F \Big] \end{aligned} \tag{3.450}$$

$$\begin{aligned} &= (\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} F_n^{\mathrm{T}} F_n (\underline{T}_n^F - \underline{\theta}_n) \\ &\qquad + 2 (F_n \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} A_n \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \\ &\qquad + E_{P_{n+1}}\Big[\underline{\psi}^{\mathrm{T}}(\underline{z}_{n+1} - F_n \underline{T}_n^F) A_n^{\mathrm{T}} A_n \underline{\psi}(\underline{z}_{n+1} - F_n \underline{T}_n^F) \\ &\qquad\qquad \mid \underline{T}_1^F, \cdots, \underline{T}_n^F \Big] \end{aligned} \tag{3.451}$$

w.p.1, where use is made of (3.430); it then follows from (3.448) that

$$\begin{aligned} Y_n &= (\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} (F_n^{\mathrm{T}} F_n - I) (\underline{T}_n^F - \underline{\theta}_n) \\ &\qquad + 2 (F_n \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} A_n \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \\ &\qquad + E_{P_{n+1}}\Big[\underline{\psi}^{\mathrm{T}}(\underline{z}_{n+1} - F_n \underline{T}_n^F) A_n^{\mathrm{T}} A_n \underline{\psi}(\underline{z}_{n+1} - F_n \underline{T}_n^F) \\ &\qquad\qquad \mid \underline{T}_1^F, \cdots, \underline{T}_n^F \Big] \end{aligned} \tag{3.452}$$

w.p.1, and thus,

$$\begin{aligned} \sum_{j=0}^{n} Y_j &= \sum_{j=0}^{n} (\underline{T}_j^F - \underline{\theta}_j)^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) (\underline{T}_j^F - \underline{\theta}_j) \\ &\qquad + 2 \sum_{j=0}^{n} (F_j \underline{T}_j^F - \underline{\theta}_{j+1})^{\mathrm{T}} A_j \underline{\xi}(F_j \underline{T}_j^F - \underline{\theta}_{j+1}) \\ &\qquad + \sum_{j=0}^{n} E_{P_{j+1}}\Big[\underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j \underline{T}_j^F) A_j^{\mathrm{T}} A_j \underline{\psi}(\underline{z}_{j+1} - F_j \underline{T}_j^F) \\ &\qquad\qquad \mid \underline{T}_1^F, \cdots, \underline{T}_j^F \Big]. \end{aligned} \tag{3.453}$$

Now: from (3.432) and (3.417),

$$(\underline{T}_j^F - \underline{\theta}_j)^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) (\underline{T}_j^F - \underline{\theta}_j)$$

$$\begin{aligned}
&= (\underline{T}_{j-1}^{F} - \underline{\theta}_{j-1})^{\mathrm{T}} F_{j-1}^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) F_{j-1} (\underline{T}_{j-1}^{F} - \underline{\theta}_{j-1}) \\
&\qquad + 2 (F_{j-1} \underline{T}_{j-1}^{F} - \underline{\theta}_{j})^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) A_{j-1} \underline{\psi}(\underline{z}_j - F_{j-1} \underline{T}_{j-1}^{F}) \\
&\qquad + \underline{\psi}^{\mathrm{T}}(\underline{z}_j - F_{j-1} \underline{T}_{j-1}^{F}) A_{j-1}^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \\
&\qquad\qquad\qquad A_{j-1} \underline{\psi}(\underline{z}_j - F_{j-1} \underline{T}_{j-1}^{F})
\end{aligned} \tag{3.454}$$

$$\begin{aligned}
&= (\underline{T}_{0}^{F} - \underline{\theta}_{0})^{\mathrm{T}} \left[\prod_{k=0}^{j-1} F_k \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \left[\prod_{k=0}^{j-1} F_k \right] (\underline{T}_{0}^{F} - \underline{\theta}_{0}) \\
&\qquad + 2 \sum_{k=0}^{j-1} (F_k \underline{T}_{k}^{F} - \underline{\theta}_{k+1})^{\mathrm{T}} \left[\prod_{i=k+1}^{j-1} F_i \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \\
&\qquad\qquad\qquad \left[\prod_{i=k+1}^{j-1} F_i \right] A_k \underline{\psi}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F}) \\
&\qquad + \sum_{k=0}^{j-1} \underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F}) A_k^{\mathrm{T}} \left[\prod_{i=k+1}^{j-1} F_i \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \\
&\qquad\qquad\qquad \left[\prod_{i=k+1}^{j-1} F_i \right] A_k \underline{\psi}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F})
\end{aligned} \tag{3.455}$$

(where sums are replaced by the additive identity if the limits of their indices overlap), so that

$$\begin{aligned}
&\sum_{j=0}^{n} (\underline{T}_{j}^{F} - \underline{\theta}_{j})^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) (\underline{T}_{j}^{F} - \underline{\theta}_{j}) \\
&\quad = \sum_{j=0}^{n} (\underline{T}_{0}^{F} - \underline{\theta}_{0})^{\mathrm{T}} \left[\prod_{k=0}^{j-1} F_k \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \left[\prod_{k=0}^{j-1} F_k \right] (\underline{T}_{0}^{F} - \underline{\theta}_{0}) \\
&\qquad + 2 \sum_{j=0}^{n} \sum_{k=0}^{j-1} (F_k \underline{T}_{k}^{F} - \underline{\theta}_{k+1})^{\mathrm{T}} \left[\prod_{i=k+1}^{j-1} F_i \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \\
&\qquad\qquad\qquad \left[\prod_{i=k+1}^{j-1} F_i \right] A_k \underline{\psi}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F}) \\
&\qquad + \sum_{j=0}^{n} \sum_{k=0}^{j-1} \underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F}) A_k^{\mathrm{T}} \left[\prod_{i=k+1}^{j-1} F_i \right]^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I) \\
&\qquad\qquad\qquad \left[\prod_{i=k+1}^{j-1} F_i \right] A_k \underline{\psi}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F})
\end{aligned} \tag{3.456}$$

$$\begin{aligned}
&\quad = (\underline{T}_{0}^{F} - \underline{\theta}_{0})^{\mathrm{T}} \sum_{j=0}^{n} \left[\left[\prod_{k=0}^{j} F_k \right]^{\mathrm{T}} \left[\prod_{k=0}^{j} F_k \right] \right. \\
&\qquad\qquad \left. - \left[\prod_{k=0}^{j-1} F_k \right]^{\mathrm{T}} \left[\prod_{k=0}^{j-1} F_k \right] \right] (\underline{T}_{0}^{F} - \underline{\theta}_{0}) \\
&\qquad + 2 \sum_{k=0}^{n-1} (F_k \underline{T}_{k}^{F} - \underline{\theta}_{k+1})^{\mathrm{T}} \sum_{j=k+1}^{n} \left[\left[\prod_{i=k+1}^{j} F_i \right]^{\mathrm{T}} \left[\prod_{i=k+1}^{j} F_i \right] \right. \\
&\qquad\qquad \left. - \left[\prod_{i=k+1}^{j-1} F_i \right]^{\mathrm{T}} \left[\prod_{i=k+1}^{j-1} F_i \right] \right] A_k \underline{\psi}(\underline{z}_{k+1} - F_k \underline{T}_{k}^{F})
\end{aligned}$$

$$+ \sum_{k=0}^{n-1} \underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} \sum_{j=k+1}^{n} \Bigg[\Bigg[\prod_{i=k+1}^{j} F_i \Bigg]^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{j} F_i \Bigg]$$

$$- \Bigg[\prod_{i=k+1}^{j-1} F_i \Bigg]^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{j-1} F_i \Bigg] \Bigg] A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) \quad (3.457)$$

$$= (\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} \Bigg[\Bigg[\prod_{k=0}^{n} F_k \Bigg]^{\mathrm{T}} \Bigg[\prod_{k=0}^{n} F_k \Bigg] - I \Bigg] (\underline{T}_0^F - \underline{\theta}_0)$$

$$+ 2 \sum_{k=0}^{n-1} (F_k \, \underline{T}_k^F - \underline{\theta}_{k+1})^{\mathrm{T}} \Bigg[\Bigg[\prod_{i=k+1}^{n} F_i \Bigg]^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{n} F_i \Bigg] - I \Bigg]$$

$$A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F)$$

$$+ \sum_{k=0}^{n-1} \underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} \Bigg[\Bigg[\prod_{i=k+1}^{n} F_i \Bigg]^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{n} F_i \Bigg] - I \Bigg]$$

$$A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F), \qquad (3.458)$$

where (3.456) follows from (3.432), and (3.458) holds by virtue of cancellation. Thus, from (3.453) and (3.458),

$$E \Bigg[\sum_{j=0}^{n} Y_j \Bigg] = (\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} \Bigg[\Bigg[\prod_{k=0}^{n} F_k \Bigg]^{\mathrm{T}} \Bigg[\prod_{k=0}^{n} F_k \Bigg] - I \Bigg] (\underline{T}_0^F - \underline{\theta}_0)$$

$$+ 2 \sum_{k=0}^{n-1} E \Bigg[E_{P_{k+1}} \Bigg[(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1})^{\mathrm{T}} \Bigg[\Bigg[\prod_{i=k+1}^{n} F_i \Bigg]^{\mathrm{T}}$$

$$\Bigg[\prod_{i=k+1}^{n} F_i \Bigg] - I \Bigg] A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) \Bigg| \, \underline{T}_k^F \Bigg] \Bigg]$$

$$+ \sum_{k=0}^{n-1} E \Bigg[\underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} \Bigg[\Bigg[\prod_{i=k+1}^{n} F_i \Bigg]^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{n} F_i \Bigg] - I \Bigg]$$

$$A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) \Bigg]$$

$$+ 2 \sum_{k=0}^{n} E \Big[(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1})^{\mathrm{T}} A_k \, \underline{\xi}(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1}) \Big]$$

$$+ \sum_{k=0}^{n} E \Big[E_{P_{k+1}} \Big[\underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F)$$

$$\Big| \, \underline{T}_1^F, \cdots, \underline{T}_k^F \Big] \Big] \qquad (3.459)$$

$$= (\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} \Bigg[\Bigg[\prod_{k=0}^{n} F_k \Bigg]^{\mathrm{T}} \Bigg[\prod_{k=0}^{n} F_k \Bigg] - I \Bigg] (\underline{T}_0^F - \underline{\theta}_0)$$

$$+ 2 \sum_{k=0}^{n} E \Bigg[(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1})^{\mathrm{T}} \Bigg[\prod_{i=k+1}^{n} F_i \Bigg]^{\mathrm{T}}$$

$$\left. \left[\prod_{i=k+1}^{n} F_i \right] A_k \, \underline{\xi}(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1}) \right]$$

$$+ \sum_{k=0}^{n} E \left[\underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} \left[\prod_{i=k+1}^{n} F_i \right]^{\mathrm{T}} \right.$$

$$\left. \left[\prod_{i=k+1}^{n} F_i \right] A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) \right] \qquad (3.460)$$

w.p.1 for all n (noting that $k = n$ terms in the first two sums on the right-hand side of (3.459) would be identically zero and can thus be added in at will), where each term is bounded by virtue of (3.423), (3.439), and the finiteness of (3.447). Hence,

$$\sup_n \; E \left[\left| (\underline{T}_{n+1}^F - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1}^F - \underline{\theta}_{n+1}) - \sum_{j=1}^{n} Y_j \right| \right]$$

$$\leq \sup_n \; E \left[(\underline{T}_{n+1}^F - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1}^F - \underline{\theta}_{n+1}) \right]$$

$$+ \sup_n \; \left| (\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} \left[\left[\prod_{k=0}^{n} F_k \right]^{\mathrm{T}} \right. \right.$$

$$\left. \left. \left[\prod_{k=0}^{n} F_k \right] - I \right] (\underline{T}_0^F - \underline{\theta}_0) \right|$$

$$- \inf_n \; 2 \sum_{k=0}^{n} E \left[(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1})^{\mathrm{T}} \left[\prod_{i=k+1}^{n} F_i \right]^{\mathrm{T}} \right.$$

$$\left. \left[\prod_{i=k+1}^{n} F_i \right] A_k \, \underline{\xi}(F_k \, \underline{T}_k^F - \underline{\theta}_{k+1}) \right]$$

$$+ \sup_n \; \sum_{k=0}^{n} E \left[\underline{\psi}^{\mathrm{T}}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) A_k^{\mathrm{T}} \left[\prod_{i=k+1}^{n} F_i \right]^{\mathrm{T}} \right.$$

$$\left. \left[\prod_{i=k+1}^{n} F_i \right] A_k \, \underline{\psi}(\underline{z}_{k+1} - F_k \, \underline{T}_k^F) \right] \qquad (3.461)$$

$$< \infty \qquad (3.462)$$

a.s., where use is made of the positivity of the first and last terms in (3.461), and the negativity of the third; the finiteness of the first term in (3.461) follows from the fact that the right-hand side of (3.434) was shown earlier to be finite. Note in passing that the infimum and supremum (respectively) of the last two terms in (3.461) are their limits as $n \to \infty$, by virtue of monotonicity. Using a martingale convergence theorem (see the version in Loève, 1963, pp.393-394), it then follows that the sequence (3.449) converges almost surely. It remains to show that each term in (3.449) does so as well.

Since

$$\sum_{j=0}^{n} E \left[E_{P_{j+1}} \left[\underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j \, \underline{T}_j^F) A_j^{\mathrm{T}} A_j \, \underline{\psi}(\underline{z}_{j+1} - F_j \, \underline{T}_j^F) \;\middle|\; \underline{T}_j^F \right] \right]$$

$$= \mathrm{tr}\left[\sum_{j=0}^{n} E\left[E_{P_{j+1}}\left[\underline{\psi}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F)\, \underline{\psi}^{\mathrm{T}}(\underline{z}_{j+1} - F_j\, \underline{T}_j^F) \;\Big|\; \underline{T}_j^F \right] A_j^{\mathrm{T}} A_j \right]\right] \tag{3.463}$$

$$\leq \mathrm{tr}\left[S_0 \sum_{j=0}^{n} A_j^{\mathrm{T}} A_j \right] \tag{3.464}$$

$$< \infty \tag{3.465}$$

(from (3.422) and the finiteness by hypothesis of S_0), and moreover the sum on the left-hand side of (3.463) is monotone (each term is of quadratic form and hence non-negative), it converges. Thus, the convergence of (3.449) implies (from (3.453)) that there is a C such that

$$\lim_{n\to\infty} \left[(\underline{T}_{n+1}^F - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1}^F - \underline{\theta}_{n+1}) - \sum_{j=0}^{n} (\underline{T}_j^F - \underline{\theta}_j)^{\mathrm{T}} (F_j^{\mathrm{T}} F_j - I)(\underline{T}_j^F - \underline{\theta}_j) + 2 \sum_{j=0}^{n} (F_j\, \underline{T}_j^F - \underline{\theta}_{j+1})^{\mathrm{T}} A_j\, \underline{\xi}(F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}) \right]$$

$$= \lim_{n\to\infty} \left[\sum_{j=0}^{n+1} (\underline{T}_j^F - \underline{\theta}_j)^{\mathrm{T}} (\underline{T}_j^F - \underline{\theta}_j) - \sum_{j=0}^{n} (\underline{T}_j^F - \underline{\theta}_j)^{\mathrm{T}} F_j^{\mathrm{T}} F_j (\underline{T}_j^F - \underline{\theta}_j) + 2 \sum_{j=0}^{n} (F_j\, \underline{T}_j^F - \underline{\theta}_{j+1})^{\mathrm{T}} A_j\, \underline{\xi}(F_j\, \underline{T}_j^F - \underline{\theta}_{j+1}) \right] \tag{3.466}$$

$$= C, \tag{3.467}$$

whence it follows that

$$\lim_{n\to\infty} \left[(\underline{T}_{n+1}^F - \underline{\theta}_{n+1})^{\mathrm{T}} (\underline{T}_{n+1}^F - \underline{\theta}_{n+1}) - (\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} F_n^{\mathrm{T}} F_n (\underline{T}_n^F - \underline{\theta}_n) + 2 (F_n\, \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} A_n\, \underline{\xi}(F_n\, \underline{T}_n^F - \underline{\theta}_{n+1}) \right] = 0 \tag{3.468}$$

w.p.1. But since (3.447) is bounded from below,

$$\limsup_{n\to\infty} E\left[(F_n\, \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} \left[\prod_{k=n+1}^{\infty} F_k \right]^{\mathrm{T}} \left[\prod_{k=n+1}^{\infty} F_k \right] A_n\, \underline{\xi}(F_n\, \underline{T}_n^F - \underline{\theta}_{n+1}) \right] = 0 \tag{3.469}$$

or, since the expression is negative a.s. from (3.419),

$$\liminf_{n\to\infty} E\left[\left| (F_n\, \underline{T}_n^F - \underline{\theta}_{n+1})^{\mathrm{T}} \left[\prod_{k=n+1}^{\infty} F_k \right]^{\mathrm{T}} \left[\prod_{k=n+1}^{\infty} F_k \right] A_n\, \underline{\xi}(F_n\, \underline{T}_n^F - \underline{\theta}_{n+1}) \right| \right]$$

$$= 0, \tag{3.470}$$

implying that there exists a subsequence $\{ n_m \}$ such that

$$\lim_{m \to \infty} E \left[\left| (F_{n_m} \underline{T}^F_{n_m} - \underline{\theta}_{n_m + 1})^{\mathrm{T}} \left[\prod_{k = n_m + 1}^{\infty} F_k \right]^{\mathrm{T}} \right. \right.$$

$$\left. \left. \left[\prod_{k = n_m + 1}^{\infty} F_k \right] A_{n_m} \underline{\xi}(F_{n_m} \underline{T}^F_{n_m} - \underline{\theta}_{n_m + 1}) \right| \right] = 0. \tag{3.471}$$

Hence, by the Chebychev inequality,

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

$$= 0 \tag{3.472}$$

w.p.1, which in turn implies, from (3.419) and the non-singularity of F_n for all n, that

$$\lim_{m \to \infty} (\underline{T}^F_{n_m} - \underline{\theta}_{n_m}) = 0 \tag{3.473}$$

w.p.1. Thus, for any $\delta_1 > 0$, there is a large enough $m(\delta_1)$ such that

$$\left| (\underline{T}^F_{n_m} - \underline{\theta}_{n_m})^{\mathrm{T}} F^{\mathrm{T}}_{n_m} F_{n_m} (\underline{T}^F_{n_m} - \underline{\theta}_{n_m}) \right.$$

$$\left. + \; 2 (F_{n_m} \underline{T}^F_{n_m} - \underline{\theta}_{n_m + 1})^{\mathrm{T}} A_{n_m} \underline{\xi}(F_{n_m} \underline{T}^F_{n_m} - \underline{\theta}_{n_m + 1}) \right| < \delta_1 \tag{3.474}$$

w.p.1 for all $m > m(\delta_1)$. Substituting into (3.468), noting that this implies (3.474) with n_m replaced by $n_m + 1$ and δ_1 replaced by $\delta_2 = O(\delta_1)$ or less (by (3.422)-(3.423) and (3.472)), and letting $\delta_1 \downarrow 0$, implies

$$\lim_{n \to \infty} (\underline{T}^F_n - \underline{\theta}_n) = 0, \tag{3.475}$$

which is the desired result.

To prove asymptotic normality, note first that by hypothesis, $\underline{\xi}(\underline{T})$ is continuous and differentiable in a neighborhood of 0, say $\| \underline{T} \| < \delta_1$, and $\underline{\xi}(0) = 0$. Thus,

$$\underline{\xi}(\underline{T}) = J(0) \, \underline{T} + O(\| \underline{T} \|^2) \tag{3.476}$$

for $\| \underline{T} \| < \delta_1$. Moreover, since $\underline{T}^F_n - \underline{\theta}_n \to 0$ w.p.1 (as proved above), there exists a large enough $n(\delta_1)$ such that $\| F_n \underline{T}^F_n - \underline{\theta}_{n+1} \| < \delta_1$ w.p.1 for all $n \geq n(\delta_1)$. It follows that (3.432) may be rewritten as

$$\underline{T}^F_{n+1} - \underline{\theta}_{n+1} = F_n (\underline{T}^F_n - \underline{\theta}_n) + A_n \left[\underline{\Psi}(\underline{z}_{n+1} - F_n \underline{T}^F_n) - \underline{\xi}(F_n \underline{T}^F_n - \underline{\theta}_{n+1}) \right]$$

$$+ \; A_n \underline{\xi}(F_n \underline{T}^F_n - \underline{\theta}_{n+1}) \tag{3.477}$$

$$= \left[I + A_n J(0) + A_n O_p(\| F_n \underline{T}^F_n - \underline{\theta}_{n+1} \|) \right] F_n (\underline{T}^F_n - \underline{\theta}_n)$$

$$+ \; A_n \left[\underline{\Psi}(\underline{z}_{n+1} - F_n \underline{T}^F_n) - \underline{\xi}(F_n \underline{T}^F_n - \underline{\theta}_{n+1}) \right] \tag{3.478}$$

w.p.1 for $n \geq n(\delta_1)$, so that $(\, \underline{T}_{n+1}^F - \underline{\theta}_{n+1})$ is the sum of a sequence of zero-mean random variables (plus some higher-order terms). Note that

$$A_n \; O_p(\; \| F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \| \;) \; = \; o_p(n^{-1}) \tag{3.479}$$

at least, in view of (3.424) (which implies that $A_n = O(n^{-1})$ or less) and (3.475) (which implies, using (3.423), that $O_p(\; \| F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \| \;) = o_p(1)$ or less).

Next, define for some $\delta_2 > 0$ the set

$$A(\, n,\, \delta_2,\, \underline{T},\, \underline{\theta}\,) \; := \; \left\{ \underline{z} : \; \| \; \underline{\psi}(\underline{z} - \underline{T}) - \underline{\xi}(\underline{T} - \underline{\theta}) \; \|^2 \; \geq \; \delta_2 \, n \; \right\}. \tag{3.480}$$

Since $\underline{\xi}(0) = 0$ by hypothesis, $\underline{\xi}(\, F_n \, \underline{T}_n^F - \underline{\theta}_{n+1}\,) < \infty$ w.p.1 for all $n \geq n(\delta_1)$ by virtue of continuity. Together with (3.420), this implies that

$$\| \; \underline{\psi}(\, \underline{z} - F_n \, \underline{T}_n^F \,) - \underline{\xi}(\, F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \,) \; \|^2 \; < \; \infty \tag{3.481}$$

w.p.1 for $n \geq n(\delta_1)$, so that

$$\lim_{n \to \infty} \; A(\, n,\, \delta_2,\, F_n \, \underline{T}_n^F,\, \underline{\theta}_{n+1} \,) \; = \; \varnothing \tag{3.482}$$

(or possibly a set of measure 0). It then follows that

$$\lim_{n \to \infty} \; \int_{A(n,\, \delta_2,\, F_n \underline{T}_n^F,\, \underline{\theta}_{n+1})} \; \| \; \underline{\psi}(\, \underline{z} - F_n \, \underline{T}_n^F \,) - \underline{\xi}(\, F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \,) \; \|^2 \, dP_n(\underline{z}) \; = \; 0 \tag{3.483}$$

w.p.1 for any $\delta_2 > 0$. This is analogous to Lindeberg's condition for asymptotic normality, and is used in the proof below.

The characteristic function of the update in (3.478) is defined as

$$\zeta_n^{\Psi}(\underline{s}) \; := \; E\left[e^{\, i\, \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{z}_{n+1} - F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))} \right] \tag{3.484}$$

$$= \; E\left[E_P \left[e^{\, i\, \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v} + \underline{\theta}_{n+1} - F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))} \mid \underline{T}_n^F \right] \right]^{\cdot} \tag{3.485}$$

w.p.1, from (3.416) and because $\{\underline{v}_n\}$ are independent and identically distributed. Using Taylor's theorem yields (see (3.308))

$$E_P \left[e^{\, i\, \underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v} + \underline{\theta}_{n+1} - F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))} \mid \underline{T}_n^F \right]$$

$$= \; 1 \; - \; \frac{1}{2} \, \underline{s}^{\,\mathrm{T}} A_n \; \Sigma(\, F_n \, \underline{T}_n^F - \underline{\theta}_{n+1} \,) \, A_n{}^{\mathrm{T}} \, \underline{s} \; + \; E_P[\, R_n \mid \underline{T}_n^F \,], \tag{3.486}$$

where R_n denotes the remainder. Since the truncation error is dominated by the first omitted term in the Taylor series,

$$| \; E_P[\, R_n \mid \underline{T}_n^F \,] \; |$$

$$\leq \; \frac{1}{6} \; \int_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v} + \underline{\theta}_{n+1} - F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))| \leq \delta_3} \; | \; \underline{s}^{\,\mathrm{T}} A_n \; (\; \underline{\psi}(\underline{v} + \underline{\theta}_{n+1} - F_n \underline{T}_n^F)$$

$$- \; \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \;) \; |^3 \, dP(\underline{v})$$

$$+ \ \frac{1}{6} \int\limits_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))| > \delta_3} | \ \underline{s}^{\mathrm{T}} A_n \ (\ \underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F)$$

$$- \ \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \ ) \ |^{\,3} \ dP(\underline{v}) \quad (3.487)$$

for some $\delta_3 > 0$, from (3.310)-(3.311). But since

$$\int\limits_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))| \le \delta_3} | \ \underline{s}^{\mathrm{T}} A_n \ (\ \underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \ ) \ |^{\,3} \ dP(\underline{v})$$

$$\le \ \delta_3 \, \underline{s}^{\mathrm{T}} A_n \ \Sigma(F_n \, \underline{T}_n^F - \underline{\theta}_{n+1}) \, A_n^{\mathrm{T}} \, \underline{s}, \qquad (3.488)$$

where δ_3 can be made arbitrarily small, from (3.313)-(3.314), and (bounding the other part of the remainder by the next lower term for convenience)

$$\int\limits_{|\underline{s}^{\mathrm{T}} A_n (\underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}))| > \delta_3} \underline{s}^{\mathrm{T}} A_n \ (\ \underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \ )$$

$$(\ \underline{\psi}(\underline{v}+\underline{\theta}_{n+1}-F_n \underline{T}_n^F) - \underline{\xi}(F_n \underline{T}_n^F - \underline{\theta}_{n+1}) \ )^{\mathrm{T}} A_n^{\mathrm{T}} \, \underline{s} \ dP(\underline{v})$$

$$= \ o_p(1) \, \underline{s}^{\mathrm{T}} A_n \ A_n^{\mathrm{T}} \, \underline{s} \qquad (3.489)$$

w.p.1, from (3.483). It follows, combining (3.487), (3.488) (letting $\delta_3 \downarrow 0$), (3.489), and the fact that $A_n = O(n^{-1})$ or less from (3.424), that

$$| \ E_P[\, R_n \ | \ \underline{T}_n^F \,] \ | \ \le \ o_p(n^{-2}) \ \|\underline{s}\|^2. \qquad (3.490)$$

Denote the characteristic function of $\underline{T}_n^F - \underline{\theta}_n$ by

$$\zeta_n^T(\underline{s}) \ := \ E[\, e^{\,i\,\underline{s}^{\mathrm{T}}(\underline{T}_n^F - \underline{\theta}_n)} \,], \qquad (3.491)$$

and define the matrix sequence

$$B_n \ := \ (\, I \ + \ A_n \ J(0) \ + \ o_p(n^{-1}) \,) \, F_n \qquad (3.492)$$

and the recursion

$$\zeta_{n+1}(\underline{s}) \ = \ \zeta_n(B_n^{\mathrm{T}} \underline{s}) \left[1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \ \Sigma(0) \, A_n^{\mathrm{T}} \, \underline{s} \right] \qquad (3.493)$$

with

$$\zeta_0(\underline{s}) \ := \ \zeta_0^T \ = \ e^{\,i\,\underline{s}^{\mathrm{T}}(\underline{T}_0^F - \underline{\theta}_0)} \qquad (3.494)$$

(since $\underline{T}_0^F$ is a given constant). It can be shown that $\zeta_n(\underline{s})$ and $\zeta_n^T(\underline{s})$ are asymptotically equivalent by noting that

$$| \ \zeta_{n+1}^T(\underline{s}) - \zeta_{n+1}(\underline{s}) \ | \ = \ \Big| \ \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) \, \zeta_n^{\Psi}(\underline{s})$$

$$- \ \zeta_n(B_n^{\mathrm{T}} \underline{s}) \left[1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \ \Sigma(0) \, A_n^{\mathrm{T}} \, \underline{s} \right] \Big| \qquad (3.495)$$

$$\le \ \left| \, 1 - \frac{1}{2} \, \underline{s}^{\mathrm{T}} A_n \ \Sigma(0) \, A_n^{\mathrm{T}} \, \underline{s} \, \right| \ \left| \, \zeta_n^T(B_n^{\mathrm{T}} \underline{s}) - \zeta_n(B_n^{\mathrm{T}} \underline{s}) \, \right|$$

$$+ \left| \zeta_n^T(B_n{}^{\mathrm{T}} \underline{s}) \right| \left| \zeta_n^{\Psi}(\underline{s}) - 1 \right.$$

$$\left. + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(0) \, A_n{}^{\mathrm{T}} \underline{s} \right| , \quad (3.496)$$

from (3.324)-(3.325). But

$$\left| \zeta_n^{\Psi}(\underline{s}) - 1 + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(0) \, A_n{}^{\mathrm{T}} \underline{s} \right|$$

$$= \left| E \left[E_P[R_n \mid \underline{T}_n^F] - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \left[\Sigma(F_n \, \underline{T}_n^F - \underline{\theta}_{n+1}) - \Sigma(0) \right] A_n{}^{\mathrm{T}} \underline{s} \right] \right| \quad (3.497)$$

$$\leq o_p(n^{-2}) \, \| \underline{s} \|^2 \quad (3.498)$$

w.p.1, where (3.497) follows from (3.486), and (3.498) from (3.424), (3.490), the fact that $\underline{T}_n^F - \underline{\theta}_n \to 0$ w.p.1, and the continuity and boundedness of $\Sigma(\underline{T})$ in a neighborhood of 0 by hypothesis. It therefore follows, using (3.327), that

$$\left| \zeta_n^T(B_n{}^{\mathrm{T}} \underline{s}) \right| \left| \zeta_n^{\Psi}(\underline{s}) - 1 + \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(0) \, A_n{}^{\mathrm{T}} \underline{s} \right| = o_p(n^{-2}) \, \| \underline{s} \|^2 . \quad (3.499)$$

Similarly, again using (3.424),

$$\left| 1 - \frac{1}{2} \underline{s}^{\mathrm{T}} A_n \, \Sigma(0) \, A_n{}^{\mathrm{T}} \underline{s} \right| = \left| 1 + O(n^{-2}) \, \| \underline{s} \|^2 \right| \quad (3.500)$$

and

$$\| B_n \| = O(1) + o_p(1) \quad (3.501)$$

or less, from (3.492) and (3.423) (which implies that $F_n = O(1)$ or less). Thus, equations (3.334)-(3.341) hold, proving that $\zeta_n^T(\underline{s})$ and $\zeta_n(\underline{s})$ are asymptotically equivalent. Moreover, it can be shown by induction (see equations (3.342)-(3.347)) that

$$\log \zeta_n(\underline{s}) = i \, \underline{s}^{\mathrm{T}} \left[\prod_{j=0}^{n-1} B_j \right] (\underline{T}_0^F - \underline{\theta}_0)$$

$$+ \sum_{j=0}^{n-1} \log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j \, \Sigma(0) \, A_j{}^{\mathrm{T}} \right.$$

$$\left. \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}} \underline{s} \right] . \quad (3.502)$$

But since

$$\| I + A_n \, J(0) + o_p(n^{-1}) \| < 1 \quad (3.503)$$

w.p.1 for large enough n (from (3.424) and the monotonicity by hypothesis of $\underline{\xi}(\underline{T})$ in a neighborhood of 0, which implies that $J(0) < 0$), it follows by (3.423) that

$$\lim_{n\to\infty} \left\| \prod_{j=0}^{n} B_i \right\| \leq \sqrt{\alpha} \lim_{n\to\infty} \prod_{j=0}^{n} \left\| I + A_n\, J(0) + o_p(n^{-1}) \right\| \tag{3.504}$$

$$= 0, \tag{3.505}$$

so that

$$\lim_{n\to\infty} \prod_{j=0}^{n} B_j = 0, \tag{3.506}$$

and the first term in (3.502) vanishes as $n \to \infty$. Moreover,

$$\log \left[1 - \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right] A_j\, \Sigma(0)\, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right]^{\mathrm{T}} \underline{s} \right]$$

$$= \frac{1}{2} \underline{s}^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right] A_j\, \Sigma(0)\, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n} B_k \right]^{\mathrm{T}} \underline{s} \tag{3.507}$$

approximately (using a first-order Taylor expansion for the logarithm) at least for large n, since (3.506) and (3.422) imply that the term on the right-hand side of (3.507) vanishes as $n \to \infty$. Thus,

$$\lim_{n\to\infty} \zeta_n(\underline{s}) - e^{-\frac{1}{2} \underline{s}^{\mathrm{T}} \left[\sum_{j=0}^{n-1} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j\, \Sigma(0)\, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}} \right] \underline{s}} = 0, \tag{3.508}$$

i.e. $\zeta_n(\underline{s})$ asymptotically has the form of the characteristic function of a normal distribution, and hence $T_n^F - \underline{\theta}_n$ is asymptotically normal as well. There only remains to derive the limiting distribution; since the sequence $\{F_n\}$ is not required to approach a limit, however, this last step necessitates some form of normalization.

Define

$$\Sigma'_n := \sum_{j=0}^{n-1} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j\, \Sigma(0)\, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_k \right]^{\mathrm{T}}, \tag{3.509}$$

and note that (3.508) implies

$$\lim_{n\to\infty} \left[E\left[(\underline{T}_n^F - \underline{\theta}_n)(\underline{T}_n^F - \underline{\theta}_n)^{\mathrm{T}} \right] - \Sigma'_n \right] = 0. \tag{3.510}$$

It is easy to verify by inspection that (3.509) yields

$$\Sigma'_{n+1} = B_n\, \Sigma'_n\, B_n^{\mathrm{T}} + A_n\, \Sigma(0)\, A_n^{\mathrm{T}}, \tag{3.511}$$

with

$$\Sigma'_0 = 0. \tag{3.512}$$

Defining the matrix sequence $\{\Sigma_n\}$ as in equations (3.426)-(3.427), and setting

$$\Delta_n := \Sigma_n - \Sigma'_n, \tag{3.513}$$

it follows (using (3.492) and the fact that $F_n = O(1)$ or less) that

$$\Delta_{n+1} = \left[I + A_n\, J(0) \right] F_n\, \Delta_n\, F_n^{\mathrm{T}} \left[I + A_n\, J(0) \right]^{\mathrm{T}} + o_p(n^{-1})\, \Sigma'_n \tag{3.514}$$

$$= \left[\prod_{j=0}^{n} (I + A_j J(0)) F_j \right] \Delta_0 \left[\prod_{j=0}^{n} (I + A_j J(0)) F_j \right]^{\mathrm{T}}$$

$$+ \sum_{j=0}^{n} o_p(j^{-1}) \left[\prod_{k=j+1}^{n} (I + A_k J(0)) F_k \right] \Sigma'_n \tag{3.515}$$

$$= \sum_{j=0}^{n} o_p(j^{-1}) \left[\prod_{k=j+1}^{n} (I + A_k J(0)) F_k \right] \Sigma'_n , \tag{3.516}$$

since $\Delta_0 = 0$ from (3.427) and (3.512). But

$$\lim_{n \to \infty} \prod_{k=j+1}^{n} (I + A_k J(0)) F_k = 0, \tag{3.517}$$

as argued previously, and moreover (3.511) yields

$$\Sigma'_n = \left[\prod_{j=0}^{n-1} B_j \right] \Sigma'_0 \left[\prod_{j=0}^{n-1} B_j \right]^{\mathrm{T}}$$

$$+ \sum_{j=0}^{n-1} \left[\prod_{k=j+1}^{n-1} B_k \right] A_j \, \Sigma(0) \, A_j^{\mathrm{T}} \left[\prod_{k=j+1}^{n-1} B_j \right]^{\mathrm{T}} , \tag{3.518}$$

where the first term vanishes by (3.512), and the second is bounded by virtue of (3.420) (which, together with the finiteness of $\underline{\xi}(0)$, implies that $\Sigma(0)$ is bounded), (3.517), and (3.422). It therefore follows that

$$\lim_{n \to \infty} (\Sigma_n - \Sigma'_n) = 0, \tag{3.519}$$

and the variance of ($\underline{T}_n^F - \underline{\theta}_n$) approaches Σ_n as $n \to \infty$. This implies (3.425) and completes the proof of the theorem. ■

Corollary 3.8 Theorem 3.1 holds also if $\underline{T}_0^F$ is a random variable, provided that

$$E \left[\underline{T}_0^F \, (\underline{T}_0^F)^{\mathrm{T}} \right] < \infty, \tag{3.520}$$

and $\underline{T}_0^F$ is independent of $\underline{z}_n$, $n = 2, 3, \cdots$. If, moreover, $\underline{T}_0^F$ is a translation-invariant function of $\underline{z}_1$, then $\underline{T}_n^F$ is translation invariant.

Proof The proof of the first part of the corollary follows that of Theorem 3.3 identically. In the proof of consistency, the condition of independence is required for equations (3.437), (3.441), and (3.451) to hold. Furthermore, products of the form $(\underline{T}_0^F - \underline{\theta}_0)^{\mathrm{T}} (\underline{T}_0^F - \underline{\theta}_0)$ (with or without norming matrices) are replaced by their expectations in (3.434), (3.445), and (3.459)-(3.461), and (3.520) is required for (3.446) and (3.462) (in their modified form) to hold. In the proof of asymptotic normality, independence is required for (3.486) and (3.497) to hold. The exponential in (3.494) and (3.342) is replaced by its expectation, while the first terms on the right-hand sides of equations (3.343), (3.344), (3.346), and (3.347) are replaced by the logarithm of the expectation of their exponentials. Equation (3.520) is required for the first term in (3.502) to vanish as $n \to \infty$. The proof of translation invariance is identical with that in Corollary 3.1, and is omitted.

■

A further generalization concerns the case where the observations $\{\underline{z}_n\}$ are corrupted by noise that is independent but not necessarily identically distributed or spherically symmetric, but that can be expressed as a time-varying linear transformation of a sample $\{\underline{v}_n\}$ of independent identically distributed random variables with a common spherically symmetric distribution function P. In other words, let

$$\underline{z}_n := \underline{\theta}_n + D_n \underline{v}_n, \tag{3.521}$$

$n = 1, 2, \cdots$, where $\{D_n\}$ is a known sequence of non-singular matrices, $D_n \in \mathbf{R}^{q \times q}$, and $\underline{\theta}_n$ obeys (3.417) with $\underline{\theta}_0$ an unknown (but fixed and finite) parameter. Define the recursion

$$\underline{T}^D_{n+1} = F_n \underline{T}^D_n + D_{n+1} A_n \underline{\psi}\left[D_{n+1}^{-1} (\underline{z}_{n+1} - F_n \underline{T}^D_n) \right], \tag{3.522}$$

$n = 0, 1, \cdots$, where $\underline{T}^D_0$ is an arbitrary (possibly random) starting point, and $\{A_n\}$ and $\underline{\psi}$ are as defined earlier.

Corollary 3.9 Under the consistency conditions of Theorem 3.3, if there is a $\beta_1 > 0$ and a $\beta_2 < \infty$ such that

$$\beta_1 I < D_n < \beta_2 I \tag{3.523}$$

for all n, then, given any $\underline{T}^D_0 < \infty$, $\underline{T}^D_n - \underline{\theta}_n \to 0$ as $n \to \infty$ a.s. Under the asymptotic normality conditions of Theorem 3.3 and (3.523),

$$\mathrm{L}(\Sigma_n^{-1/2} (\underline{T}^D_n - \underline{\theta}_n)) \to \mathrm{N}(0, I), \tag{3.524}$$

where

$$\Sigma_n = D_n \left[I + A_{n-1} J(0) \right] (D_n^{-1} F_{n-1}) \Sigma_{n-1} (D_n^{-1} F_{n-1})^{\mathrm{T}} \left[I + A_{n-1} J(0) \right]^{\mathrm{T}} D_n^{\mathrm{T}}$$
$$+ D_n A_{n-1} \Sigma(0) A_{n-1}^{\mathrm{T}} D_n^{\mathrm{T}} \tag{3.525}$$

with

$$\Sigma_0 = 0. \tag{3.526}$$

Proof Letting

$$\tilde{\underline{T}}_n := D_n^{-1} \underline{T}^D_n, \tag{3.527}$$

equation (3.522) yields

$$D_{n+1} \tilde{\underline{T}}_{n+1} = F_n D_n \tilde{\underline{T}}_n + D_{n+1} A_n \underline{\psi}\left[D_{n+1}^{-1} (\underline{z}_{n+1} - F_n D_n \tilde{\underline{T}}_n) \right], \tag{3.528}$$

or, premultiplying by D_{n+1}^{-1} (which exists and is positive, by (3.523)),

$$\tilde{\underline{T}}_{n+1} = D_{n+1}^{-1} F_n D_n \tilde{\underline{T}}_n + A_n \underline{\psi}\left[D_{n+1}^{-1} \underline{z}_{n+1} - D_{n+1}^{-1} F_n D_n \tilde{\underline{T}}_n \right]. \tag{3.529}$$

Similarly, (3.521) may be rewritten as

$$D_n^{-1} \underline{z}_n = D_n^{-1} \underline{\theta}_n + \underline{v}_n, \qquad (3.530)$$

or, defining

$$\tilde{\underline{\theta}}_n := D_n^{-1} \underline{\theta}_n, \qquad (3.531)$$

it follows that

$$D_n^{-1} \underline{z}_n = \tilde{\underline{\theta}}_n + \underline{v}_n, \qquad (3.532)$$

with

$$\tilde{\underline{\theta}}_{n+1} = D_{n+1}^{-1} F_n \underline{\theta}_n \qquad (3.533)$$

$$= D_{n+1}^{-1} F_n D_n \tilde{\underline{\theta}}_n \qquad (3.534)$$

from (3.531) and (3.417). Thus, defining

$$\tilde{F}_n := D_{n+1}^{-1} F_n D_n, \qquad (3.535)$$

where $D_0 := I$ for convenience, it follows that

$$\tilde{\underline{T}}_{n+1} = \tilde{F}_n \tilde{\underline{T}}_n + A_n \underline{\psi}\left[D_{n+1}^{-1} \underline{z}_{n+1} - \tilde{F}_n \tilde{\underline{T}}_n \right], \qquad (3.536)$$

and moreover,

$$\left[\prod_{j=m}^{n} \tilde{F}_j \right]^{\mathrm{T}} \left[\prod_{j=m}^{n} \tilde{F}_j \right] = D_m^{\mathrm{T}} \left[\prod_{j=m}^{n} F_j \right]^{\mathrm{T}} (D_{n+1}^{-1})^{\mathrm{T}} D_{n+1}^{-1} \left[\prod_{j=m}^{n} F_j \right] D_m \qquad (3.537)$$

$$< \alpha \frac{\beta_2^2}{\beta_1^2} I \qquad (3.538)$$

from (3.523) and (3.423). Finally, from (3.530) and by hypothesis, $\{ D_n^{-1} (\underline{z}_n - \underline{\theta}_n) \}$ is a sample of independent random variates with a common distribution function P, so that

$$E\left[\underline{\psi}(D_n^{-1} \underline{z} - \underline{T}) \right] = E_P \left[\underline{\psi}(\underline{v} + D_n^{-1} \underline{\theta}_n - \underline{T}) \right] \qquad (3.539)$$

$$= \underline{\xi}(\underline{T} - D_n^{-1} \underline{\theta}_n), \qquad (3.540)$$

and similarly,

$$E\left[\underline{\psi}(D_n^{-1} \underline{z} - \underline{T}) \underline{\psi}^{\mathrm{T}}(D_n^{-1} \underline{z} - \underline{T}) \right] \le S_0 \qquad (3.541)$$

for some $S_0 < \infty$, from (3.420). Thus, under the conditions of Theorem 3.3, $\tilde{\underline{T}}_n - \tilde{\underline{\theta}}_n \to 0$ as $n \to \infty$ a.s., whence it follows by (3.527) and (3.531) that $\underline{T}_n^D - \underline{\theta}_n \to 0$ as $n \to \infty$ a.s.

An argument similar to (3.539)-(3.540) also establishes that

$$E\left[\left[\underline{\psi}(D_n^{-1} \underline{z} - \underline{T}) - \underline{\xi}(\underline{T} - D_n^{-1} \underline{\theta}_n) \right] \left[\underline{\psi}(D_n^{-1} \underline{z} - \underline{T}) - \underline{\xi}(\underline{T} - D_n^{-1} \underline{\theta}_n) \right]^{\mathrm{T}} \right]$$

$$= \Sigma(\underline{T} - D_n^{-1} \underline{\theta}_n), \qquad (3.542)$$

so that under the conditions of Theorem 3.3,

$$\mathbf{L}(\, \tilde{\Sigma}_n^{-1/2} \, (\tilde{\underline{T}}_n - \tilde{\underline{\theta}}_n \,)\,) \;\rightarrow\; \mathbf{N}(\, 0, I \,), \qquad . \quad (3.543)$$

with $\tilde{\Sigma}_n$ given by

$$\tilde{\Sigma}_n \;=\; \Big[I + A_{n-1} \, J(0) \Big] \, (\, D_n^{-1} \, F_{n-1} \, D_{n-1} \,) \, \tilde{\Sigma}_{n-1} \, (D_n^{-1} \, F_{n-1} \, D_{n-1} \,)^{\mathrm{T}} \, \Big[I + A_{n-1} \, J(0) \, \Big]^{\mathrm{T}}$$

$$+ \; A_{n-1} \, \Sigma(0) \, A_{n-1}^{\mathrm{T}} \qquad (3.544)$$

and

$$\tilde{\Sigma}_0 \;=\; 0. \qquad (3.545)$$

Thus, setting

$$\Sigma_n \;:=\; D_n^{-1} \, \tilde{\Sigma}_n \, (\, D_n^{-1} \,)^{\mathrm{T}}, \qquad (3.546)$$

equations (3.527) and (3.531) establish (3.524), completing the proof. ■

The models in equations (3.416) and (3.521) both correspond to the case where the *full* state $\underline{\theta}_n$ of the dynamic system is observed, though corrupted by noise. While this is rare in practice, consistency results such as those of Theorem 3.3 and Corollary 3.9 are obviously not possible if the observations $\underline{z}_n$ do not span the entire space of the system state. In that case, different goals must be set, such as minimizing the asymptotic error variance rather than seeking to drive it to zero as $n \rightarrow \infty$. The following generalization is perhaps of limited practical use, since it still assumes that the full state can be observed, but is given here for completeness.

Consider the model

$$\underline{z}_n \;=\; H_n \, \underline{\theta}_n \;+\; \underline{v}_n \,, \qquad (3.547)$$

$n = 1, 2, \cdots$, where $\{H_n\}$ is a known sequence of matrices, $H_n \in \mathbf{R}^{p \times q}$ with $p \geq q$, and $\underline{\theta}_n$ obeys (3.417) with $\underline{\theta}_0$ an unknown (but fixed and finite) parameter. Define the recursion

$$\underline{T}_{n+1}^H \;=\; F_n \, \underline{T}_n^H \;+\; (\, H_{n+1}^{\mathrm{T}} \, H_{n+1} \,)^{-1} \, H_{n+1}^{\mathrm{T}} \, A_n \, \underline{\psi}(\, \underline{z}_{n+1} - H_{n+1} \, F_n \, \underline{T}_n^H \,), \qquad (3.548)$$

$n = 0, 1, \cdots$ (provided the inverse exists), where $\{A_n\}$ is a given matrix sequence with $A_n \in \mathbf{R}^{p \times p}$, $\underline{T}_0^H$ is an arbitrary (possibly random) starting point, and $\underline{\psi}: \mathbf{R}^p \rightarrow \mathbf{R}^p$ is related to the least favorable distribution of $\underline{v}$ in the manner discussed earlier.

Corollary 3.10 Under the consistency conditions of Theorem 3.3, if there is an $\gamma_1 > 0$ and an $\gamma_2 < \infty$ such that

$$\gamma_1 \, I \;<\; H_n^{\mathrm{T}} \, H_n \;<\; \gamma_2 \, I \qquad (3.549)$$

for all n (implying, among other things, that $\mathrm{rank}[H_n] = q$ for all n), then, given any $\underline{T}_0^H < \infty$, $\underline{T}_n^H - \underline{\theta}_n \rightarrow 0$ as $n \rightarrow \infty$ a.s. Under the asymptotic normality conditions of Theorem 3.3 and (3.549),

$$\mathbf{L}(\, \Sigma_n^{-1/2} \, (\underline{T}_n^H - \underline{\theta}_n \,)\,) \;\rightarrow\; \mathbf{N}(\, 0, I \,), \qquad (3.550)$$

where

$$\begin{aligned}\Sigma_n \;=\; & (H_n^{\mathrm{T}} H_n)^{-1} H_n^{\mathrm{T}} \left[I + A_{n-1} J(0) \right] H_n \, F_{n-1} \, \Sigma_{n-1} \\ & \qquad F_{n-1}^{\mathrm{T}} \, H_n^{\mathrm{T}} \left[I + A_{n-1} J(0) \right]^{\mathrm{T}} H_n \, (H_n^{\mathrm{T}} H_n)^{-1} \\ & + (H_n^{\mathrm{T}} H_n)^{-1} H_n^{\mathrm{T}} A_{n-1} \, \Sigma(0) \, A_{n-1}^{\mathrm{T}} \, H_n \, (H_n^{\mathrm{T}} H_n)^{-1} \end{aligned} \tag{3.551}$$

with

$$\Sigma_0 \;=\; 0. \tag{3.552}$$

Proof Letting

$$\tilde{\underline{\theta}}_n \;:=\; H_n \, \underline{\theta}_n , \tag{3.553}$$

equation (3.417) implies that

$$\begin{aligned}\tilde{\underline{\theta}}_{n+1} \;&=\; H_{n+1} \, \underline{\theta}_{n+1} & (3.554)\\ &=\; H_{n+1} \, F_n \, \underline{\theta}_n & (3.555)\\ &=\; H_{n+1} \, F_n \, (H_n^{\mathrm{T}} H_n)^{-1} H_n^{\mathrm{T}} \, \tilde{\underline{\theta}}_n & (3.556)\end{aligned}$$

(where the inverse exists and is positive by virtue of (3.549)). Thus, defining

$$\tilde{F}_n \;:=\; H_{n+1} \, F_n \, (H_n^{\mathrm{T}} H_n)^{-1} H_n^{\mathrm{T}}, \tag{3.557}$$

it follows from (3.549) and (3.423) that

$$\begin{aligned}& \left[\prod_{j=m}^{n} \tilde{F}_j \right]^{\mathrm{T}} \left[\prod_{j=m}^{n} \tilde{F}_j \right] \\ &\quad = H_m \, (H_m^{\mathrm{T}} H_m)^{-1} \left[\prod_{j=m}^{n} F_j \right]^{\mathrm{T}} H_{n+1}^{\mathrm{T}} \, H_{n+1} \left[\prod_{j=m}^{n} F_j \right] (H_m^{\mathrm{T}} H_m)^{-1} \, H_m^{\mathrm{T}} & (3.558)\\ &\quad < \alpha \, \frac{\gamma_2^2}{\gamma_1^2} \, I & (3.559)\end{aligned}$$

Hence, under the consistency conditions of Theorem 3.3, the recursion

$$\tilde{\underline{T}}_{n+1} \;=\; \tilde{F}_n \, \tilde{\underline{T}}_n \; + \; A_n \, \underline{\psi}(\, \underline{z}_{n+1} - \tilde{F}_n \, \tilde{\underline{T}}_n \,) \tag{3.560}$$

is consistent and $\tilde{\underline{T}}_n - \tilde{\underline{\theta}}_n \to 0$ as $n \to \infty$ a.s. Multiplying (3.560) through by $(\, H_{n+1}^{\mathrm{T}} \, H_{n+1} \,)^{-1} \, H_{n+1}^{\mathrm{T}}$, setting

$$\tilde{\underline{T}}_n \;:=\; H_n \, \underline{T}_n^H , \tag{3.561}$$

and substituting (3.557) establishes that $\underline{T}_n^H - \underline{\theta}_n \to 0$ as $n \to \infty$ a.s.

Similarly, under the asymptotic normality conditions of Theorem 3.3,

$$\mathbf{L}(\, \tilde{\Sigma}_n^{-1/2} \, (\tilde{\underline{T}}_n - \tilde{\underline{\theta}}_n) \,) \;\to\; \mathrm{N}(\, 0, I \,), \tag{3.562}$$

with $\tilde{\Sigma}_n$ given by

$$\tilde{\Sigma}_n = \Big[I + A_{n-1} J(0) \Big] H_n \, F_{n-1} (H_{n-1}^{\mathrm{T}} H_{n-1})^{-1} H_{n-1}^{\mathrm{T}} \tilde{\Sigma}_{n-1}$$

$$H_{n-1} (H_{n-1}^{\mathrm{T}} H_{n-1})^{-1} F_{n-1}^{\mathrm{T}} H_n^{\mathrm{T}} \Big[I + A_{n-1} J(0) \Big]^{\mathrm{T}}$$

$$+ \; A_{n-1} \Sigma(0) A_{n-1}^{\mathrm{T}} \tag{3.563}$$

and

$$\tilde{\Sigma}_0 = 0. \tag{3.564}$$

Thus, setting

$$\tilde{\Sigma}_n = H_n \Sigma_n H_n^{\mathrm{T}} \tag{3.565}$$

(from (3.561)) and noting that this implies

$$\Sigma_n = (H_n^{\mathrm{T}} H_n)^{-1} H_n^{\mathrm{T}} \tilde{\Sigma}_n H_n (H_n^{\mathrm{T}} H_n)^{-1} \tag{3.566}$$

establishes (3.551), completing the proof. ■

Remark Two special cases of Corollary 3.10 are of interest:

(i) If $p = q$, then (3.549) implies that H_n^{-1} exists for all n, so that (3.548) reduces to

$$\underline{T}_{n+1}^{H} = F_n \underline{T}_n^{H} + H_{n+1}^{-1} A_n \underline{\Psi}(\underline{z}_{n+1} - H_{n+1} F_n \underline{T}_n^{H}). \tag{3.567}$$

(ii) The case

$$\underline{z}_n = H_n \underline{\theta}_n + D_n \underline{\nu}_n \tag{3.568}$$

follows trivially from Corollary 3.10, since (3.568) may be multiplied through by D_n^{-1}, yielding the recursion

$$\underline{T}_{n+1} = F_n \underline{T}_n + \Big[(D_{n+1}^{-1} H_{n+1})^{\mathrm{T}} (D_{n+1}^{-1} H_{n+1}) \Big]^{-1}$$

$$(D_{n+1}^{-1} H_{n+1})^{\mathrm{T}} A_n \underline{\Psi} \Big[D_{n+1}^{-1} (\underline{z}_{n+1} - H_{n+1} F_n \underline{T}_n) \Big] \tag{3.569}$$

by analogy to (3.548). Indeed, setting $H_n = I$ for all n in (3.569) shows that Corollary 3.9 is only a special case of Corollary 3.10.

4. Approximate Conditional Mean Estimators

The recursive estimators discussed in Section 3 correspond to the linear dynamic model discussed in Section 1.1 when there is no process noise. In other words, they are estimators of location parameters which are either fixed or vary 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. This section therefore extends previous results to the case where process noise is pressent.

Consider the model

$$\underline{\theta}_{n+1} = F_n \, \underline{\theta}_n + \underline{w}_n \tag{4.1}$$

$$\underline{z}_n = \underline{\theta}_n + \underline{v}_n , \tag{4.2}$$

$n = 0, 1, \cdots$, where $\{F_n\}$ is a sequence of non-singular matrices, $\underline{\theta}_0$ is a random variable with $L(\underline{\theta}_0) = N(\overline{\underline{\theta}}_0, M_0)$, where $0 < M_0 < \infty$, and $\{\underline{w}_n\}$ is an independent random sequence with $L(\underline{w}_n) = N(0, Q_n)$, where $Q_n \geq 0$ for all n. Assume, moreover, that $\{\underline{v}_n\}$ is a sample of independent random variates with a common spherically symmetric distribution function P centered at the origin, and that $\underline{\theta}_0$, $\{\underline{w}_n\}$, and $\{\underline{v}_n\}$ are mutually independent.

Here, the "location parameter" $\underline{\theta}_n$ is itself random and time-variant, necessitating different conditions and a somewhat different approach than those discussed so far; in that sense, Corollary 3.10 represents in a way the end of the road for a recursive estimator of a purely Robbins-Monro type. For example, it is clear that if $\underline{\theta}_n$ changes randomly over time, the gains $\{A_n\}$ cannot always be required to vanish as $n \to \infty$ since observations must continue to be taken into account in order to track the trajectory of $\underline{\theta}_n$.

Furthermore, results that can be obtained in the presence of process noise are somewhat weaker than those of Theorem 3.3 and related corollaries. In particular, since $\underline{\theta}_n$ is now randomly varying, the estimator cannot be consistent, i.e. the estimation error variance does not vanish, except in some special cases. Indeed, only using asymptotic performance measures makes little sense except in the special case where the process noise vanishes w.p.1 as $n \to \infty$. Instead, it is necessary to seek other performance criteria, measuring short-term performance as well.

As in the case of non-robust recursive estimation (the Kalman Filter), an appropriate criterion in the robust case is unbiasedness and minimum variance. It is well known that the *conditional mean estimator* fullfils these conditions (see for example Anderson and Moore, 1979, pp.26-28). The first derivation of a robust approximate conditional mean estimator of the state $\underline{\theta}_n$ of the linear dynamic system (4.1)-(4.2) in the presence of heavy-tailed observation noise $\{\underline{v}_n\}$ 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 estimate based on past observations $\{ \underline{z}_0, \cdots, \underline{z}_{n-1} \}$ is zero-mean normal. This assumption allows rather clever algebraic manipulation that yields 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 $P = \mathrm{N}(0, R)$ (see Spall and Wall, 1984), which is clearly of no interest here. No analytical results have been published to bolster empirical findings, and the resulting *ad hoc* application of this assumption has therefore not been uniformly accepted in the literature.

In Section 4.1, a first-order approximation to the conditional distribution prior to updating is derived for the case where P belongs to the ε-contaminated normal family. In Section 4.2, this distribution is used in an extension of Masreliez's theorem to derive a first-order approximation to a robust conditional mean estimator. Some related simplifications and approximations are then given in Section 4.3, and a brief discussion of minimax issues follows in Section 4.4.

4.1 A First-Order Approximation to the Conditional Prior Distribution

As stated above, the method pioneered by Masreliez and Martin is crucially dependent on the assumption that the estimate immediately prior to updating is conditionally normal. While this is never exactly satisfied in the presence of non-normal noise, it is shown in this section that, in fact, the zeroeth-order term in a Taylor series representation of the distribution is indeed normal. Furthermore, a first-order approximation is derived, and the error is shown to be bounded as $n \to \infty$, provided certain conditions are satisfied. The small parameter around which the Taylor series is constructed involves ε, the fraction of contamination.

It is first shown 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, pp.234-243), Hager and Horowitz (1976), and Moore and Anderson (1980). Hager and Horowitz (1976) have proposed relaxing the conditions of *controllability* and *observability*, used below, to *detectability* and *stabilizability*, but have only provided results for the time-invariant case; while they claim that extension to the time-variant case is direct, this is not obvious. Moore and Anderson (1980) promise the extension in a future paper, and investigate these conditions further in Anderson and Moore (1981).

The stability theorem discussed below follows Moore and Anderson (1980). Although it is required here that $\{F_n\}$ be non-singular, this condition is relaxed by Moore and Anderson. The following simple lemma will be used:

Lemma 4.1 Let $A \in \mathbf{R}^{n \times n}$, $B \in \mathbf{R}^{m \times m}$, and $C \in \mathbf{R}^{n \times m}$ be such that $A > 0$, $B > 0$, and furthermore $A = A^{\mathrm{T}}$ and $B = B^{\mathrm{T}}$. Then,

$$A - C B C^{\mathrm{T}} \geq 0 \tag{4.3}$$

if and only if

$$B^{-1} - C^{\mathrm{T}} A^{-1} C \;\geq\; 0. \tag{4.4}$$

Proof It is easy to verify that

$$\begin{bmatrix} A - C\,B\,C^{\mathrm{T}} & 0 \\ 0 & B^{-1} \end{bmatrix} = \begin{bmatrix} I & -C\,B \\ 0 & I \end{bmatrix} \begin{bmatrix} A & C \\ C^{\mathrm{T}} & B^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -B\,C^{\mathrm{T}} & I \end{bmatrix}. \tag{4.5}$$

and similarly,

$$\begin{bmatrix} B^{-1} - C^{\mathrm{T}} A^{-1} C & 0 \\ 0 & A \end{bmatrix} = \begin{bmatrix} I & -C^{\mathrm{T}} A^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} B^{-1} & C^{\mathrm{T}} \\ C & A \end{bmatrix} \begin{bmatrix} I & 0 \\ -A^{-1} C & I \end{bmatrix}. \tag{4.6}$$

Then, it follows from (4.5) that (4.3) holds if and only if

$$\begin{bmatrix} A & C \\ C^{\mathrm{T}} & B^{-1} \end{bmatrix} \geq 0 \tag{4.7}$$

and likewise, from (4.6), that (4.4) holds if and only if

$$\begin{bmatrix} B^{-1} & C^{\mathrm{T}} \\ C & A \end{bmatrix} \geq 0. \tag{4.8}$$

Comparing equations (4.7) and (4.8), and noting that one can be obtained by pre- and post-multiplying the other by a rotation matrix, proves the lemma. (Moore and Anderson, 1980.) ■

The exponential asymptotic stability of the Kalman Filter recursions is now established. The following generalization of equation (4.2) is utilized, since the more general results are used further on:

$$\underline{z}_n \;=\; H_n\, \underline{\theta}_n \;+\; D_n\, \underline{v}_n\,, \tag{4.9}$$

where $\{H_n\}$ and $\{D_n\}$ are sequences of matrices of appropriate dimensions, and D_n is non-singular for all n. Moreover, the notation

$$R_n \;:=\; D_n\, R\, D_n^{\mathrm{T}} \tag{4.10}$$

is used for brevity, where $R := E_P\,[\underline{v}_n\, \underline{v}_n^{\mathrm{T}}]$.

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

$$\sum_{i=n}^{n+t} \left[\prod_{j=n}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n}^{i-1} F_j \right] > \alpha I \tag{4.11}$$

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

$$\sum_{i=n-s}^{n} \left[\prod_{j=i+1}^{n} F_j \right] Q_i \left[\prod_{j=i+1}^{n} F_j \right]^{\mathrm{T}} > \beta I \tag{4.12}$$

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

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

$$\tilde{\underline{\theta}}_{n+1} = (I - K_{n+1} H_{n+1}) F_n \, \tilde{\underline{\theta}}_n , \tag{4.13}$$

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

$$|| \, \tilde{\underline{\theta}}_n \, || < \lambda \, \delta^n , \tag{4.14}$$

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

Proof Define first the Lyapunov function (see for example Kalman and Bertram, 1960; La Salle and Lefschetz, 1961, pp.33-36; Hahn, 1963, pp.14-18, Willems, 1970, pp.170-185)

$$V_n = \tilde{\underline{\theta}}_n^{\mathrm{T}} P_n^{-1} \tilde{\underline{\theta}}_n , \tag{4.15}$$

where P_n is the Kalman covariance defined in equation (1.8) (where it is denoted by Σ_n). Note that

$$\hat{\underline{\theta}}_n - \underline{\theta}_n = F_{n-1} \hat{\underline{\theta}}_{n-1} + K_n (\underline{z}_n - H_n F_{n-1} \hat{\underline{\theta}}_{n-1}) - (F_{n-1} \underline{\theta}_{n-1} + \underline{w}_{n-1}) \tag{4.16}$$

$$= F_{n-1} \hat{\underline{\theta}}_{n-1} + K_n (H_n \underline{\theta}_n + D_n \underline{v}_n - H_n F_{n-1} \hat{\underline{\theta}}_{n-1}) - (F_{n-1} \underline{\theta}_{n-1} + \underline{w}_{n-1}) \tag{4.17}$$

$$\begin{aligned} = F_{n-1} \hat{\underline{\theta}}_{n-1} + K_n & \Big[H_n (F_{n-1} \underline{\theta}_{n-1} + \underline{w}_{n-1}) + D_n \underline{v}_n - H_n F_{n-1} \hat{\underline{\theta}}_{n-1} \Big] \\ & - (F_{n-1} \underline{\theta}_{n-1} + \underline{w}_{n-1}) \end{aligned} \tag{4.18}$$

$$= (I - K_n H_n) F_{n-1} (\hat{\underline{\theta}}_{n-1} - \underline{\theta}_{n-1}) + K_n D_n \underline{v}_n - (I - K_n H_n) \underline{w}_{n-1} , \tag{4.19}$$

where $\hat{\underline{\theta}}_n$ is the Kalman estimate defined in equation (1.3), (4.17) follows from (4.9), and (4.18) from (4.1). Thus,

$$\begin{aligned} P_n = (I - K_n H_n) F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}} & + K_n R_n K_n^{\mathrm{T}} \\ & + (I - K_n H_n) Q_{n-1} (I - K_n H_n)^{\mathrm{T}} \end{aligned} \tag{4.20}$$

(by independence), so that

$$\begin{aligned} P_n - K_n R_n K_n^{\mathrm{T}} &= (I - K_n H_n) F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}} \\ & \quad + (I - K_n H_n) Q_{n-1} (I - K_n H_n)^{\mathrm{T}} \end{aligned} \tag{4.21}$$

$$\geq (I - K_n H_n) F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}} , \tag{4.22}$$

since the last term in (4.21) is of quadratic form with $Q_{n-1} \geq 0$ by hypothesis. Furthermore,

$$K_n = M_n H_n^{\mathrm{T}} (H_n M_n H_n^{\mathrm{T}} + R_n)^{-1} \tag{4.23}$$

$$= M_n H_n^{\mathrm{T}} \Big[R_n^{-1} - R_n^{-1} H_n (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} H_n^{\mathrm{T}} R_n^{-1} \Big] \tag{4.24}$$

$$= M_n H_n^{\mathrm{T}} R_n^{-1} - M_n H_n^{\mathrm{T}} R_n^{-1} H_n (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} H_n^{\mathrm{T}} R_n^{-1} \tag{4.25}$$

$$\begin{aligned} = M_n (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n) & (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} H_n^{\mathrm{T}} R_n^{-1} \\ & - M_n H_n^{\mathrm{T}} R_n^{-1} H_n (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} H_n^{\mathrm{T}} R_n^{-1} \end{aligned} \tag{4.26}$$

$$= M_n M_n^{-1} (M_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} H_n^{\mathrm{T}} R_n^{-1} \tag{4.27}$$

$$= \Big[M_n - M_n H_n^{\mathrm{T}} (R_n + H_n M_n H_n^{\mathrm{T}})^{-1} H_n M_n \Big] H_n^{\mathrm{T}} R_n^{-1} \tag{4.28}$$

$$= P_n H_n^{\mathrm{T}} R_n^{-1}, \tag{4.29}$$

where (4.23) follows from (1.5)-(1.6), (4.24) and (4.28) from the Sherman-Morrison-Woodbury theorem (see for example Householder, 1964, pp.123-124) -- the existence of R_n^{-1} is guaranteed by the fact that the sequence $\{R_n\}$ is bounded below by hypothesis, and that of M_n^{-1} by the non-singularity of $\{F_n\}$, the fact that $M_0 > 0$ by hypothesis, and the structure of equation (1.7) -- and (4.29) follows from (1.8) and (1.5)-(1.6). Thus,

$$P_n H_n^{\mathrm{T}} R_n^{-1} H_n P_n = P_n H_n^{\mathrm{T}} R_n^{-1} R_n R_n^{-1} H_n P_n \tag{4.30}$$

$$= K_n R_n K_n \tag{4.31}$$

from (4.29). Finally, it follows from the Sherman-Morrison-Woodbury theorem that

$$(P_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} = P_n - P_n H_n^{\mathrm{T}} (R_n + H_n P_n H_n^{\mathrm{T}})^{-1} H_n P_n \tag{4.32}$$

$$\geq P_n - P_n H_n^{\mathrm{T}} R_n^{-1} H_n P_n \tag{4.33}$$

$$= P_n - K_n R_n K_n \tag{4.34}$$

$$\geq (I - K_n H_n) F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}}, \tag{4.35}$$

where (4.33) follows from the fact that $P_n \geq 0$ so that

$$H_n P_n H_n^{\mathrm{T}} \geq 0, \tag{4.36}$$

$$R_n + H_n P_n H_n^{\mathrm{T}} \geq R_n, \tag{4.37}$$

and therefore

$$(R_n + H_n P_n H_n^{\mathrm{T}})^{-1} \leq R_n^{-1}, \tag{4.38}$$

(4.34) follows from (4.31), and (4.35) from (4.22). Thus,

$$(P_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)^{-1} - (I - K_n H_n) F_{n-1} P_{n-1} F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}} \geq 0, \tag{4.39}$$

which implies that

$$P_{n-1}^{-1} - F_{n-1}^{\mathrm{T}} (I - K_n H_n)^{\mathrm{T}} (P_n^{-1} + H_n^{\mathrm{T}} R_n^{-1} H_n)(I - K_n H_n) F_{n-1} \geq 0 \tag{4.40}$$

by virtue of Lemma 4.1.

From (4.15), therefore,

$$V_n - V_{n+1} = \tilde{\underline{\theta}}_n^{\mathrm{T}} P_n^{-1} \tilde{\underline{\theta}}_n - \tilde{\underline{\theta}}_{n+1}^{\mathrm{T}} P_{n+1}^{-1} \tilde{\underline{\theta}}_{n+1} \tag{4.41}$$

$$= \tilde{\underline{\theta}}_n^{\mathrm{T}} \Big[P_n^{-1} - F_n^{\mathrm{T}} (I - K_{n+1} H_{n+1})^{\mathrm{T}} P_{n+1}^{-1} (I - K_{n+1} H_{n+1}) F_n \Big] \tilde{\underline{\theta}}_n \tag{4.42}$$

$$\geq \tilde{\underline{\theta}}_n^{\mathrm{T}} F_n^{\mathrm{T}} (I - K_{n+1} H_{n+1})^{\mathrm{T}} H_{n+1}^{\mathrm{T}} R_{n+1}^{-1} H_{n+1} (I - K_{n+1} H_{n+1}) F_n \tilde{\underline{\theta}}_n, \tag{4.43}$$

where (4.42) follows from (4.13), and (4.43) from (4.40) with $n+1$ substituted for n. Thus,

$$V_{n-1} - V_{n+t} = \sum_{i=n}^{n+t} (V_{i-1} - V_i) \tag{4.44}$$

$$\geq \sum_{i=n}^{n+t} \underline{\tilde{\theta}}_{i-1}^{\mathrm{T}} F_{i-1}^{\mathrm{T}} (I - K_i H_i)^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i (I - K_i H_i) F_{i-1} \underline{\tilde{\theta}}_{i-1} \tag{4.45}$$

$$= \underline{\tilde{\theta}}_n^{\mathrm{T}} \left[\sum_{i=n}^{n+t} \left[\prod_{j=n+1}^{i} (I - K_j H_j) F_{j-1} \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n+1}^{i} (I - K_j H_j) F_{j-1} \right] \right] \underline{\tilde{\theta}}_n \tag{4.46}$$

from (4.43) and (4.13). It is now necessary to show that this quantity is bounded below by a positive number.

Note first that under the conditions of this theorem, P_n is bounded above. This can be proven by considering the (suboptimal) moving average estimate

$$\underline{\overline{\theta}}_n := \left[\prod_{j=n-t}^{n-1} F_j \right] \left[\sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \right]^{-1} \sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} \underline{z}_i \tag{4.47}$$

for $n \geq t$, where the existence of the inverse of the sum is guaranteed by (4.11) with n replaced by $n-t$. From (4.1) and (4.9),

$$\underline{z}_i = H_i \underline{\theta}_i + D_i \underline{v}_i \tag{4.48}$$

$$= H_i F_{i-1} \underline{\theta}_{i-1} + H_i \underline{w}_{i-1} + D_i \underline{v}_i \tag{4.49}$$

$$= H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \underline{\theta}_{n-t} + H_i \sum_{k=n-t}^{i-1} \left[\prod_{j=k+1}^{i-1} F_j \right] \underline{w}_k + D_i \underline{v}_i , \tag{4.50}$$

so that

$$\begin{aligned} \underline{\overline{\theta}}_n = & \left[\prod_{j=n-t}^{n-1} F_j \right] \left[\sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \right]^{-1} \\ & \Bigg(\left[\sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \right] \underline{\theta}_{n-t} \\ & + \sum_{i=n-t+1}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \sum_{k=n-t}^{i-1} \left[\prod_{j=k+1}^{i-1} F_j \right] \underline{w}_k \\ & + \sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} D_i \underline{v}_i \Bigg] \end{aligned} \tag{4.51}$$

$$\begin{aligned} = & \left[\prod_{j=n-t}^{n-1} F_j \right] \underline{\theta}_{n-t} \\ & + \left[\prod_{j=n-t}^{n-1} F_j \right] \left[\sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \right]^{-1} \end{aligned}$$

$$\left[\sum_{i=n-t+1}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \sum_{k=n-t}^{i-1} \left[\prod_{j=k+1}^{i-1} F_j \right] \underline{w}_k \right.$$

$$\left. + \sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} D_i \, \underline{v}_i \right]. \tag{4.52}$$

Similarly, from (4.1),

$$\underline{\theta}_n = \left[\prod_{j=n-t}^{n-1} F_j \right] \underline{\theta}_{n-t} + \sum_{k=n-t}^{n-1} \left[\prod_{j=k+1}^{n-1} F_j \right] \underline{w}_k \tag{4.53}$$

so that, subtracting (4.53) from (4.52), the $\underline{\theta}_{n-t}$ terms cancel, leaving

$$\overline{\underline{\theta}}_n - \underline{\theta}_n = \left[\prod_{j=n-t}^{n-1} F_j \right] \left[\sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n-t}^{i-1} F_j \right] \right]^{-1}$$

$$\left[\sum_{i=n-t+1}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \sum_{k=n-t}^{i-1} \left[\prod_{j=k+1}^{i-1} F_j \right] \underline{w}_k \right.$$

$$\left. + \sum_{i=n-t}^{n} \left[\prod_{j=n-t}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} D_i \, \underline{v}_i \right]$$

$$- \sum_{k=n-t}^{n-1} \left[\prod_{j=k+1}^{n-1} F_j \right] \underline{w}_k . \tag{4.54}$$

Since the matrix sequences $\{F_n\}$, $\{H_n\}$, $\{Q_n\}$, and $\{R_n\}$ are bounded above by hypothesis, and likewise $\{R_n\}$ is also bounded below, and by (4.11) which ensures that the inverse of the sum in (4.54) is bounded above, it follows that there exists a γ_1 obeying $0 < \gamma_1 < \infty$ such that

$$E\left[(\overline{\underline{\theta}}_n - \underline{\theta}_n)(\overline{\underline{\theta}}_n - \underline{\theta}_n)^{\mathrm{T}} \right] \leq \gamma_1 I. \tag{4.55}$$

But since $\overline{\underline{\theta}}_n$ is suboptimal,

$$P_n \leq E\left[(\overline{\underline{\theta}}_n - \underline{\theta}_n)(\overline{\underline{\theta}}_n - \underline{\theta}_n)^{\mathrm{T}} \right] \tag{4.56}$$

$$\leq \gamma_1 I \tag{4.57}$$

from (4.55), establishing that the sequence $\{P_n\}$ is bounded above, and so, by (4.29), is $\{K_n\}$.

Note furthermore that from (1.5)-(1.6),

$$(I - K_n H_n) = (I - M_n H_n^{\mathrm{T}} (H_n M_n H_n^{\mathrm{T}} + R_n)^{-1} H_n) \tag{4.58}$$

$$= M_n^{1/2} (I - M_n^{1/2} H_n^{\mathrm{T}} (H_n M_n H_n^{\mathrm{T}} + R_n)^{-1} H_n M_n^{1/2}) M_n^{-1/2} \tag{4.59}$$

$$= M_n^{1/2} \left[I - M_n^{1/2} H_n^{\mathrm{T}} \left[R_n^{-1} - R_n^{-1} H_n M_n^{1/2} (I + M_n^{1/2} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{1/2})^{-1} \right. \right.$$

$$\left. \left. M_n^{1/2} H_n^{\mathrm{T}} R_n^{-1} \right] H_n M_n^{1/2} \right] M_n^{-1/2} \tag{4.60}$$

$$= M_n^{1/2} \left[I - M_n^{1/2} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{1/2} \right.$$

$$+ M_n^{1/2} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{1/2} (I + M_n^{1/2} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{1/2})^{-1}$$

$$M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} \Big] M_n^{-\frac12} \tag{4.61}$$

$$= M_n^{\frac12} \Big[I - M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12}) + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} \Big] M_n^{-\frac12} \tag{4.62}$$

$$= M_n^{\frac12} \Big[I - M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} \Big] M_n^{-\frac12} \tag{4.63}$$

$$= M_n^{\frac12} \Big[(I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})(I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} - M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} \Big] M_n^{-\frac12} \tag{4.64}$$

$$= M_n^{\frac12} (I + M_n^{\frac12} H_n^{\mathrm{T}} R_n^{-1} H_n M_n^{\frac12})^{-1} M_n^{-\frac12} \tag{4.65}$$

$$\geq 0, \tag{4.66}$$

where (4.60) follows from the Sherman-Morrison-Woodbury theorem, and (4.66) from the positivity of the covariance matrices M_n and R_n, as well as the non-negativity of the quadratic form in (4.65).

Since $\{K_n\}$ is bounded above (as shown earlier), equation (4.66) implies that there exists a γ_2 obeying $0 < \gamma_2 < 1$ such that

$$(I - K_n H_n) \geq \gamma_2 I. \tag{4.67}$$

It follows that (4.46) can be rewritten as

$$V_{n-1} - V_{n+t} \geq \underline{\tilde{\theta}}_n^{\mathrm{T}} \left[\sum_{i=n}^{n+t} \left[\prod_{j=n+1}^{i} \gamma_2 F_{j-1} \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n+1}^{i} \gamma_2 F_{j-1} \right] \right] \underline{\tilde{\theta}}_n \tag{4.68}$$

$$= \underline{\tilde{\theta}}_n^{\mathrm{T}} \left[\sum_{i=n}^{n+t} \gamma_2^{2(i-n)} \left[\prod_{j=n}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n}^{i-1} F_j \right] \right] \underline{\tilde{\theta}}_n \tag{4.69}$$

$$\geq \gamma_2^{2t} \underline{\tilde{\theta}}_n^{\mathrm{T}} \left[\sum_{i=n}^{n+t} \left[\prod_{j=n}^{i-1} F_j \right]^{\mathrm{T}} H_i^{\mathrm{T}} R_i^{-1} H_i \left[\prod_{j=n}^{i-1} F_j \right] \right] \underline{\tilde{\theta}}_n \tag{4.70}$$

$$\geq \gamma_2^{2t} \alpha \underline{\tilde{\theta}}_n^{\mathrm{T}} \underline{\tilde{\theta}}_n, \tag{4.71}$$

where (4.68) follows from (4.67), (4.70) from the fact that $0 < \gamma_2 < 1$. and (4.71) from (4.11). Since the right-hand side of (4.71) is non-negative, this establishes, by the method due to Lyapunov, that

$$\lim_{n \to \infty} \| \underline{\tilde{\theta}}_n \| = 0, \tag{4.72}$$

i.e. the system (4.13) is asymptotically stable.

To prove exponential asymptotic stability, it is first necessary to show that the sequence $\{P_n\}$ is bounded below as well. Note that

$$P_n = M_n - M_n H_n^{\mathrm{T}}(H_n M_n H_n^{\mathrm{T}} + R_n)^{-1} H_n M_n \tag{4.73}$$

$$= M_n^{1/2} \left[I - M_n^{1/2} H_n^{\mathrm{T}}(H_n M_n H_n^{\mathrm{T}} + R_n)^{-1} H_n M_n^{1/2} \right] M_n^{1/2} \tag{4.74}$$

$$\geq \gamma_2 M_n , \tag{4.75}$$

where (4.75) follows from (4.67) and (4.59) postmultiplied by M_n. Thus, in particular,

$$P_0 \geq \gamma_2 M_0, \tag{4.76}$$

whence it follows, using (1.7), that

$$M_1 = F_0 P_0 F_0^{\mathrm{T}} + Q_0 \tag{4.77}$$

$$\geq \gamma_2 F_0 M_0 F_0^{\mathrm{T}} + Q_0. \tag{4.78}$$

Define the recursion

$$\overline{M}_{n+1} = \gamma_2 F_n \overline{M}_n F_n^{\mathrm{T}} + Q_n \tag{4.79}$$

with

$$\overline{M}_0 = M_0. \tag{4.80}$$

Then, assuming by the induction argument that

$$M_n \geq \overline{M}_n , \tag{4.81}$$

it follows that

$$M_{n+1} = F_n P_n F_n^{\mathrm{T}} + Q_n \tag{4.82}$$

$$\geq \gamma_2 F_n M_n F_n^{\mathrm{T}} + Q_n \tag{4.83}$$

$$\geq \gamma_2 F_n \overline{M}_n F_n^{\mathrm{T}} + Q_n \tag{4.84}$$

$$= \overline{M}_{n+1}, \tag{4.85}$$

where (4.83) follows from (4.75), (4.84) from (4.81), and (4.85) from (4.79), thus proving that (4.81) holds for all n. But (4.79) yields

$$\overline{M}_{n+1} = \gamma_2^{s+1} \left[\prod_{j=n-s}^{n} F_j \right] \overline{M}_{n-s} \left[\prod_{j=n-s}^{n} F_j \right]^{\mathrm{T}} + \sum_{i=n-s}^{n} \gamma_2^{n-i} \left[\prod_{j=i+1}^{n} F_j \right] Q_i \left[\prod_{j=i+1}^{n} F_j \right]^{\mathrm{T}} \tag{4.86}$$

$$\geq \gamma_2^{s+1} \left[\prod_{j=n-s}^{n} F_j \right] \overline{M}_{n-s} \left[\prod_{j=n-s}^{n} F_j \right]^{\mathrm{T}} + \gamma_2^{s} \sum_{i=n-s}^{n} \left[\prod_{j=i+1}^{n} F_j \right] Q_i \left[\prod_{j=i+1}^{n} F_j \right]^{\mathrm{T}} \tag{4.87}$$

$$\geq \gamma_2^{s+1}\left[\prod_{j=n-s}^{n} F_j\right]\bar{M}_{n-s}\left[\prod_{j=n-s}^{n} F_j\right]^{\mathrm{T}} + \gamma_2^{s}\beta I \tag{4.88}$$

$$\geq \gamma_2^{s}\beta I, \tag{4.89}$$

where (4.87) follows from the fact that $0 < \gamma_2 < 1$, (4.88) from (4.12), and (4.89) from the non-negativity of the quadratic form in (4.88). Thus, combining (4.81) and (4.89) -- which holds independently of n -- yields

$$M_n \geq \gamma_2^{s}\beta I, \tag{4.90}$$

whence it follows from (4.75) that

$$P_n \geq \gamma_2^{s+1}\beta I, \tag{4.91}$$

or

$$P_n^{-1} \leq \frac{1}{\gamma_2^{s+1}\beta} I. \tag{4.92}$$

Thus,

$$V_n \leq \frac{1}{\gamma_2^{s+1}\beta}\, \tilde{\underline{\theta}}_n^{\mathrm{T}}\tilde{\underline{\theta}}_n \tag{4.93}$$

from (4.15) and (4.92), so that (4.71) yields

$$V_{n-1} - V_{n+t} \geq \gamma_2^{2t}\alpha\gamma_2^{s+1}\beta\, V_n \tag{4.94}$$

$$\geq \gamma_2^{2t+s+1}\alpha\beta\, V_{n+t}, \tag{4.95}$$

where (4.95) follows from the monotonicity of V_n, evident from (4.43) and the non-negativity of the quadratic form. Rewriting (4.95) as

$$V_{n+t} \leq \frac{1}{1+\gamma_2^{2t+s+1}\alpha\beta} V_{n-1} \tag{4.96}$$

establishes the exponential asymptotic convergence of V_n, with

$$V_n \leq \left[\frac{1}{1+\gamma_2^{2t+s+1}\alpha\beta}\right]^{\frac{n}{t+1}} V_0. \tag{4.97}$$

Then, using the non-negativity of V_n (from (4.15) and the positivity of the covariance), equation (4.71) yields

$$\gamma_2^{2t}\alpha\tilde{\underline{\theta}}_n^{\mathrm{T}}\tilde{\underline{\theta}}_n \leq V_{n-1} - V_{n+t} \tag{4.98}$$

$$\leq V_{n-1} \tag{4.99}$$

$$\leq \left[\frac{1}{1+\gamma_2^{2t+s+1}\alpha\beta}\right]^{\frac{n-1}{t+1}} V_0, \tag{4.100}$$

from (4.97). The proof is concluded by taking the square root of (4.100), and setting

$$\lambda = \left[\frac{(1 + \gamma_2^{2t+s+1} \alpha \beta)^{t+1}}{\gamma_2^{2t} \alpha} \ \underline{\tilde{\theta}}_0^{\mathrm{T}} P_0^{-1} \underline{\tilde{\theta}}_0 \right]^{½} \tag{4.101}$$

and

$$\delta = \left[\frac{1}{1 + \gamma_2^{2t+s+1} \alpha \beta} \right]^{\frac{1}{2(t+1)}} . \tag{4.102}$$

(This is a special case, restricted to non-singular transition matrices, of the proof given in Moore and Anderson, 1980.) ■

This result is used in the following, slightly different form. Let $\mathrm{N}(\underline{x}; \underline{\mu}, \Sigma)$ denote the density associated with the normal distribution of a random variable $\underline{x}$, with mean $\underline{\mu}$ and covariance Σ.

Corollary 4.1 Let the conditions of Theorem 4.1 be satisfied for the system (4.1) and (4.9), and let a $0 < \phi < \infty$ exist such that for all n,

$$\left\| \prod_{j=1}^{n} F_j \right\| < \phi \tag{4.103}$$

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

$$\underline{\theta}_{n+1}^i = F_n \underline{\theta}_n^i + K_{n+1}^i (\underline{z}_{n+1} - H_{n+1} F_n \underline{\theta}_n^i) \tag{4.104}$$

$$K_n^i = M_n^i H_n^{\mathrm{T}} (H_n M_n^i H_n^{\mathrm{T}} + R_n)^{-1} \tag{4.105}$$

$$M_{n+1}^i = F_n P_n^i F_n^{\mathrm{T}} + Q_n \tag{4.106}$$

$$P_n^i = (I - K_n^i H_n) M_n^i \tag{4.107}$$

be two Kalman Filters with respective initial state estimates $\underline{\theta}_0^i$ and initial covariances M_0^i, $i = 1, 2$. Then, there is a $0 < \delta < 1$ such that for any finite $\underline{\theta}$,

$$\mathrm{N}(\underline{\theta}; \underline{\theta}_n^1, M_n^1) = \mathrm{N}(\underline{\theta}; \underline{\theta}_n^2, M_n^2) + O_p(\delta^n). \tag{4.108}$$

Proof Combining (4.107) and (4.106) with n replaced by $n-1$,

$$\begin{aligned} P_n^1 - P_n^2 &= (I - K_n^1 H_n)(F_{n-1} P_{n-1}^1 F_{n-1}^{\mathrm{T}} + Q_{n-1}) \\ &\qquad - (I - K_n^2 H_n)(F_{n-1} P_{n-1}^2 F_{n-1}^{\mathrm{T}} + Q_{n-1}) \qquad (4.109) \\ &= (I - K_n^1 H_n)(F_{n-1} P_{n-1}^1 F_{n-1}^{\mathrm{T}} + Q_{n-1}) \\ &\qquad - (F_{n-1} P_{n-1}^2 F_{n-1}^{\mathrm{T}} + Q_{n-1})(I - K_n^2 H_n)^{\mathrm{T}} \qquad (4.110) \\ &= (I - K_n^1 H_n) F_{n-1} (P_{n-1}^1 - P_{n-1}^2) F_{n-1}^{\mathrm{T}} (I - K_n^2 H_n)^{\mathrm{T}} \\ &\qquad + \left[(I - K_n^1 H_n) F_{n-1} P_{n-1}^1 F_{n-1}^{\mathrm{T}} + Q_n \right] (K_n^2 H_n)^{\mathrm{T}} \\ &\qquad - (K_n^1 H_n) \left[F_{n-1} P_{n-1}^2 F_{n-1}^{\mathrm{T}} (I - K_n^2 H_n) + Q_n \right] \qquad (4.111) \end{aligned}$$

$$= (I - K_n^1 H_n) F_{n-1} (P_{n-1}^1 - P_{n-1}^2) F_{n-1}^{\mathrm{T}} (I - K_n^2 H_n)^{\mathrm{T}}$$

$$+ P_n^1 (K_n^2 H_n)^{\mathrm{T}} - (K_n^1 H_n) P_n^2, \tag{4.112}$$

where (4.110) follows from symmetry, and (4.112) from (4.106)-(4.107). But from (4.29),

$$P_n^1 (K_n^2 H_n)^{\mathrm{T}} - (K_n^1 H_n) P_n^2 = P_n^1 (P_n^2 H_n^{\mathrm{T}} R_n^{-1} H_n)^{\mathrm{T}} - (P_n^1 H_n^{\mathrm{T}} R_n^{-1} H_n) P_n^2 \tag{4.113}$$

$$= 0 \tag{4.114}$$

by symmetry. Thus, (4.112) becomes

$$P_n^1 - P_n^2 = (I - K_n^1 H_n) F_{n-1} (P_{n-1}^1 - P_{n-1}^2) F_{n-1}^{\mathrm{T}} (I - K_n^2 H_n)^{\mathrm{T}} \tag{4.115}$$

$$= \left[\prod_{j=1}^{n} (I - K_j^1 H_j) F_{j-1} \right] (P_0^1 - P_0^2) \left[\prod_{j=1}^{n} (I - K_j^2 H_j) F_{j-1} \right]^{\mathrm{T}}. \tag{4.116}$$

But Theorem 4.1 implies that there exists a $0 < \delta < 1$ such that

$$\left\| \prod_{j=1}^{n+1} (I - K_j^i H_j) F_{j-1} \right\| = O(\delta^{n+1}). \tag{4.117}$$

It follows that

$$\| P_n^1 - P_n^2 \| = O(\delta^{2n}), \tag{4.118}$$

and hence, by (4.29) and the facts that $\{H_n\}$ is bounded above and $\{R_n\}$ is bounded below,

$$\| K_n^1 - K_n^2 \| = O(\delta^{2n}) \tag{4.119}$$

also. Similarly, (4.106) yields

$$M_{n+1}^1 - M_{n+1}^2 = F_n P_n^1 F_n^{\mathrm{T}} + Q_n - F_n P_n^2 F_n^{\mathrm{T}} - Q_n \tag{4.120}$$

$$= F_n (P_n^1 - P_n^2) F_n^{\mathrm{T}}, \tag{4.121}$$

so that, since F_n is bounded above by hypothesis,

$$\| M_n^1 - M_n^2 \| = O(\delta^{2n}), \tag{4.122}$$

from (4.118) and (4.121).

Now, equation (4.104) yields

$$\underline{\theta}_{n+1}^i = \left[\prod_{j=1}^{n+1} (I - K_j^i H_j) F_{j-1} \right] \underline{\theta}_0^i + \sum_{k=0}^{n+1} \left[\prod_{j=k+1}^{n+1} (I - K_j^i H_j) F_{j-1} \right] K_k^i \underline{z}_k . \tag{4.123}$$

It therefore follows that

$$\| \underline{\theta}_{n+1}^1 - \underline{\theta}_{n+1}^2 \| \leq O(\delta^{n+1}) \| \underline{\theta}_0^1 - \underline{\theta}_0^2 \|$$

$$+ \sum_{k=0}^{n+1} O(\delta^{n-k+1}) \| K_k^1 - K_k^2 \| \| \underline{z}_k \| \tag{4.124}$$

$$= O(\delta^{n+1}) \| \underline{\theta}_0^1 - \underline{\theta}_0^2 \| + O(\delta^{n+1}) \sum_{k=0}^{n+1} O(\delta^k) \| \underline{z}_k \| \tag{4.125}$$

$$= O_p(\,\delta^{n+1}\,), \tag{4.126}$$

where (4.124) follows from the Cauchy-Schwarz inequality, (4.125) from (4.119), and (4.126) from the fact that the transition matrix is bounded above, by (4.103), and therefore so w.p.1 is the output.

Finally, note that $\mathrm{N}(\,\underline{x};\,\underline{\mu},\,\Sigma\,)$ is everywhere continuously differentiable with respect to $\underline{\mu}$ and Σ except at $\Sigma = 0$. But equation (4.90) shows that M_n^i is bounded away from 0, so that it is possible to write a first-order Taylor series expansion as follows:

$$\begin{aligned}
\mathrm{N}(\,\underline{\theta};\,\underline{\theta}_n^1,\,M_n^1\,) \;=\; & \mathrm{N}(\,\underline{\theta};\,\underline{\theta}_n^2,\,M_n^2\,) \\
& + \sum_i\,[\,\underline{\theta}_n^1-\underline{\theta}_n^2\,]_i\,\frac{\partial}{\partial[\underline{\mu}]_i}\,\mathrm{N}(\,\underline{\theta};\,\underline{\mu},\,M_n^2\,)\,\Big|_{\underline{\mu}=\underline{\theta}_n^2} \\
& + \sum_{ij}\,[\,M_n^1-M_n^2\,]_{ij}\,\frac{\partial}{\partial[\Sigma]_{ij}}\,\mathrm{N}(\,\underline{\theta};\,\underline{\theta}_n^2,\,\Sigma\,)\,\Big|_{\Sigma=M_n^2} \\
& + \Delta,
\end{aligned} \tag{4.127}$$

where $[\;]_i$ and $[\;]_{ij}$ respectively denote vector and matrix elements, and Δ is the remainder term. Using (4.122) and (4.126) concludes the proof. (The use of equation (4.115) follows Jazwinski, 1970, pp.242-243.) ■

The following lemma is used repeatedly in the proof of Theorems 4.2 and 4.3:

Lemma 4.2 Let $\underline{\theta}$, $\underline{x}$, and $\underline{y} \in \mathbf{R}^n$ and A, B, and $C \in \mathbf{R}^{n\times n}$. Let $A > 0$, $C > 0$. and furthermore $A = A^{\mathrm{T}}$ and $C = C^{\mathrm{T}}$. Then,

$$\begin{aligned}
& \mathrm{N}(\,\underline{\theta};\,\underline{x},\,A\,)\,\mathrm{N}(\,B\,\underline{\theta};\,\underline{y},\,C\,) \\
& \quad = \mathrm{N}(\,\underline{\theta};\,\underline{x}+A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}(\underline{y}-B\,\underline{x}),\,A-A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\,) \\
& \qquad \mathrm{N}(\,\underline{y};\,B\,\underline{x},\,C+B\,A\,B^{\mathrm{T}}\,).
\end{aligned} \tag{4.128}$$

Proof Expanding the sum of the exponents on the left-hand-side of (4.128), and neglecting the $-½$ factor for simplicity, yields

$$\begin{aligned}
& \underline{\theta}^{\mathrm{T}}A^{-1}\underline{\theta} - 2\underline{\theta}^{\mathrm{T}}A^{-1}\underline{x} + \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{\theta}^{\mathrm{T}}B^{\mathrm{T}}C^{-1}B\,\underline{\theta} - 2\underline{\theta}^{\mathrm{T}}B^{\mathrm{T}}C^{-1}\underline{y} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y} \\
& \quad = \underline{\theta}^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B\,)\underline{\theta} - 2\underline{\theta}^{\mathrm{T}}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
& \qquad + \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y} \qquad\qquad (4.129) \\
& \quad = \underline{\theta}^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B\,)\underline{\theta} \\
& \qquad - 2\underline{\theta}^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B\,)(A^{-1}+B^{\mathrm{T}}C^{-1}B\,)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
& \qquad + (A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B\,)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})
\end{aligned}$$

$$\begin{aligned}
&- (A^{-1}\underline{x} + B^{\mathrm{T}}C^{-1}\underline{y})^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
&+ \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y} \qquad (4.130) \\
= &\left[\underline{\theta} - (A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})\right]^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1} \\
&\qquad \left[\underline{\theta} - (A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})\right] \\
&- (A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
&+ \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y}. \qquad (4.131)
\end{aligned}$$

It follows from the Sherman-Morrison-Woodbury theorem that

$$\begin{aligned}
&(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
&\quad = \left[A - A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\right](A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \qquad (4.132) \\
&\quad = \underline{x} - A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,\underline{x} \\
&\qquad\qquad + A\,B^{\mathrm{T}}C^{-1}\underline{y} - A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\,B^{\mathrm{T}}C^{-1}\underline{y}, \qquad (4.133)
\end{aligned}$$

where the existence of the inverse in (4.132) is guaranteed by the fact that $A > 0$ and $C > 0$, by hypothesis. But

$$\begin{aligned}
&A\,B^{\mathrm{T}}\left[I - (C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\,B^{\mathrm{T}}\right]C^{-1}\underline{y} \\
&\quad = A\,B^{\mathrm{T}}\left[(C+B\,A\,B^{\mathrm{T}})^{-1}(C+B\,A\,B^{\mathrm{T}})\right. \\
&\qquad\qquad \left. - (C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\,B^{\mathrm{T}}\right]C^{-1}\underline{y} \qquad (4.134) \\
&\quad = A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}C\,C^{-1}\underline{y} \qquad (4.135) \\
&\quad = A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}\underline{y}. \qquad (4.136)
\end{aligned}$$

Thus,

$$(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) = \underline{x} + A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}(\underline{y} - B\,\underline{x}). \qquad (4.137)$$

Moreover,

$$\begin{aligned}
&-(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})^{\mathrm{T}}(A^{-1}+B^{\mathrm{T}}C^{-1}B)^{-1}(A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) + \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y} \\
&\quad = (A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y})^{\mathrm{T}}\left[A - A\,B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\right](A^{-1}\underline{x}+B^{\mathrm{T}}C^{-1}\underline{y}) \\
&\qquad\qquad + \underline{x}^{\mathrm{T}}A^{-1}\underline{x} + \underline{y}^{\mathrm{T}}C^{-1}\underline{y} \qquad (4.138) \\
&\quad = \underline{x}^{\mathrm{T}}\left[-A^{-1} + B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B + A^{-1}\right]\underline{x} \\
&\qquad\qquad - 2\underline{x}^{\mathrm{T}}\left[B^{\mathrm{T}}C^{-1} - B^{\mathrm{T}}(C+B\,A\,B^{\mathrm{T}})^{-1}B\,A\,B^{\mathrm{T}}C^{-1}\right]\underline{y}
\end{aligned}$$

$$\begin{aligned} &+ \underline{y}^{\mathrm{T}} \Big[- C^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \\ &+ C^{-1} B\,A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} + C^{-1} \Big\} \underline{y}. \end{aligned} \tag{4.139}$$

Now:

$$\Big[B^{\mathrm{T}} C^{-1} - B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \Big]$$

$$= B^{\mathrm{T}} \Big[I \; - \; (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} \Big] C^{-1} \tag{4.140}$$

$$= B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1}, \tag{4.141}$$

as in (4.134)-(4.136), while

$$- C^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \; + \; C^{-1} B\,A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} + C^{-1}$$

$$\begin{aligned} = &- C^{-1} (C + B\,A\,B^{\mathrm{T}})(C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \\ &+ C^{-1} B\,A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \; + \; C^{-1} \end{aligned} \tag{4.142}$$

$$= - C^{-1} C \, (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \; + \; C^{-1} \tag{4.143}$$

$$\begin{aligned} = &- (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} C^{-1} \\ &+ \; (C + B\,A\,B^{\mathrm{T}})^{-1} (C + B\,A\,B^{\mathrm{T}}) C^{-1} \end{aligned} \tag{4.144}$$

$$= (C + B\,A\,B^{\mathrm{T}})^{-1} C \, C^{-1} \tag{4.145}$$

$$= (C + B\,A\,B^{\mathrm{T}})^{-1}. \tag{4.146}$$

Finally, note that

$$\frac{|\; C + B\,A\,B^{\mathrm{T}} \;|}{|\; A^{-1} + B^{\mathrm{T}} C^{-1} B \;|} = \frac{|\; C + B\,A\,B^{\mathrm{T}} \;|}{|\; A^{-1} + B^{\mathrm{T}} C^{-1} B \;|} \; \frac{|\; B\,B^{\mathrm{T}} \;|}{|\; B\,B^{\mathrm{T}} \;|} \tag{4.147}$$

$$= \frac{|\; B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B \;|}{|\; A^{-1} + B^{\mathrm{T}} C^{-1} B \;| \; |\; B\,B^{\mathrm{T}} \;|} \tag{4.148}$$

$$= \frac{|\; (A^{-1} + B^{\mathrm{T}} C^{-1} B)^{-1} B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B \;|}{|\; B\,B^{\mathrm{T}} \;|}. \tag{4.149}$$

But

$$(A^{-1} + B^{\mathrm{T}} C^{-1} B)^{-1} B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B$$

$$= \Big[A \; - \; A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A \Big] B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B \tag{4.150}$$

$$\begin{aligned} = &\; A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B \\ &- \; A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}}) B \end{aligned} \tag{4.151}$$

$$= A\,B^{\mathrm{T}} (C + B\,A\,B^{\mathrm{T}})^{-1} (C + B\,A\,B^{\mathrm{T}}) (C + B\,A\,B^{\mathrm{T}}) B$$

$$- A\,B^{\mathrm{T}}(C + B\,A\,B^{\mathrm{T}})^{-1} B\,A\,B^{\mathrm{T}}(C + B\,A\,B^{\mathrm{T}})B \quad (4.152)$$

$$= A\,B^{\mathrm{T}}(C + B\,A\,B^{\mathrm{T}})^{-1} C\,(C + B\,A\,B^{\mathrm{T}})B. \quad (4.153)$$

It follows that

$$\frac{|\,C + B\,A\,B^{\mathrm{T}}\,|}{|\,A^{-1} + B^{\mathrm{T}}C^{-1}B\,|} = \frac{|\,A\,B^{\mathrm{T}}(C + B\,A\,B^{\mathrm{T}})^{-1}C\,(C + B\,A\,B^{\mathrm{T}})B\,|}{|\,B\,B^{\mathrm{T}}\,|} \quad (4.154)$$

$$= \frac{|\,A\,|\,|\,C\,|\,|\,(C + B\,A\,B^{\mathrm{T}})^{-1}(C + B\,A\,B^{\mathrm{T}})\,|\,|\,B\,B^{\mathrm{T}}\,|}{|\,B\,B^{\mathrm{T}}\,|} \quad (4.155)$$

$$= |\,A\,|\,|\,C\,|. \quad (4.156)$$

Combining (4.137), (4.139), (4.141), (4.146), and (4.156) establishes (4.128), completing the proof. ■

In the sequel, it is assumed that $\mathbf{L}(\,\underline{v}_n\,) := P \in P_\varepsilon$, the ε-contaminated normal family defined in equation (2.135). Then, it is possible to write

$$P = (\,1 - \varepsilon\,)\,\mathrm{N}(\,0, R\,) + \varepsilon\,H \quad (4.157)$$

for some $H \in \mathbf{S}$. It is further assumed that H is absolutely continuous with respect to the Lebesgue measure, and admits the probability density h in accordance with the Radon-Nikodym theorem. A first-order approximation to the conditional probability distribution of the estimate of the state $\underline{\theta}_n$ based on past observations $\{\,\underline{z}_0, \cdots, \underline{z}_{n-1}\,\}$ 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 (4.1)-(4.2), and let δ be a real number for which (4.14) holds. Let ω be the smallest integer such that

$$\delta^{\omega} \le \varepsilon. \quad (4.158)$$

If

$$\omega\,\varepsilon < 1 \quad (4.159)$$

and if the distribution H has bounded moments, then

$$p(\,\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}\,)$$

$$= (\,1 - \varepsilon\,)^{\omega}\,\kappa_n\,\kappa_n^0\,\mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^0, M_n^0\,)$$

$$+ \varepsilon(\,1 - \varepsilon\,)^{\omega-1}\,\kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\,\mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i, M_n^i\,)$$

$$\int \mathrm{N}(\,\underline{z}_{i-1} - \xi;\,\underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,),\,W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\,h(\,\xi\,)\,d\xi$$

$$+ O_p(\,\omega^2\,\varepsilon^2\,) \quad (4.160)$$

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

$$\underline{\theta}_n^i = F_{n-1} \underline{\theta}_{n-1}^i + F_{n-1} M_{n-1}^i \Gamma_{n-1}^{i\ -1} (\underline{z}_{n-1} - \underline{\theta}_{n-1}^i) \tag{4.161}$$

$$M_n^i = F_{n-1} P_{n-1}^i F_{n-1}^{\mathrm{T}} + Q_{n-1} \tag{4.162}$$

$$P_n^i = M_n^i - M_n^i \Gamma_n^{i\ -1} M_n^i \tag{4.163}$$

$$\Gamma_n^i = M_n^i + R \tag{4.164}$$

$$V_n^i = V_{n-1}^i P_{n-1}^i F_{n-1}^{\mathrm{T}} M_n^{i\ -1} \tag{4.165}$$

$$\underline{v}_n^i = \underline{v}_{n-1}^i + V_{n-1}^i M_{n-1}^i \Gamma_{n-1}^{i\ -1} (\underline{z}_{n-1} - \underline{\theta}_{n-1}^i) \tag{4.166}$$

$$W_n^i = W_{n-1}^i - V_{n-1}^i M_{n-1}^i \Gamma_{n-1}^{i\ -1} M_{n-1}^i V_{n-1}^{i\ \mathrm{T}} \tag{4.167}$$

$$\kappa_n^i = \kappa_{n-1}^i \mathrm{N}(\underline{z}_{n-1}; \underline{\theta}_{n-1}^i, \Gamma_{n-1}^i) \tag{4.168}$$

subject to the initial conditions

$$\underline{\theta}_i^i = F_{i-1} \underline{\theta}_{i-1}^0 \tag{4.169}$$

$$M_i^i = F_{i-1} M_{i-1}^0 F_{i-1}^{\mathrm{T}} + Q_{i-1} \tag{4.170}$$

$$V_i^i = M_{i-1}^0 F_{i-1}^{\mathrm{T}} M_i^{i\ -1} \tag{4.171}$$

$$\underline{v}_i^i = \underline{\theta}_{i-1}^0 \tag{4.172}$$

$$W_i^i = M_{i-1}^0 \tag{4.173}$$

$$\kappa_i^i = \kappa_{i-1}^0 \tag{4.174}$$

for $i > 0$, and

$$\underline{\theta}_0^0 = \overline{\underline{\theta}}_0 \tag{4.175}$$

$$M_0^0 = M_0 \tag{4.176}$$

$$\kappa_0^0 = 1. \tag{4.177}$$

The normalization constant satisfies

$$\kappa_n^{-1} = (1 - \varepsilon)^{\omega} \kappa_n^0 + \varepsilon (1 - \varepsilon)^{\omega - 1} \sum_{i = n - \omega + 1}^{n} \kappa_n^i \int \mathrm{N}(\underline{z}_{i-1} - \underline{\xi}; \underline{v}_n^i, W_n^i) \, h(\underline{\xi}) \, d \underline{\xi}. \tag{4.178}$$

Remark Before proceeding with the proof of Theorem 4.2, some comments are in order:

(i) Equations (4.161)-(4.164) are a bank of Kalman Filters, each starting at a different point in time $i = 0, 1, 2, \cdots$. Because of the way in which they are initialized, the cases $i > 0$ correspond to Kalman Filters skipping the ith observation. The case $i = 0$ is based on all observations.

(ii) Equations (4.165)-(4.167) are a bank of optimal fixed-point smoothers (see for example Anderson and Moore, 1979, pp.170-175; also Gelb, 1974, pp.170-172 -- where, however, the error covariance matrix propagation equation is incorrect), each estimating the state at a different point in time $i = 0, 1, 2, \cdots$, based on all preceeding and subsequent observations.

(iii) Thus, each term in the summation on the right-hand side of (4.160) 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 (4.160) are given in Di Masi, Runggaldier, and Barazzi (1983).

(iv) From (4.157), it is possible to write

$$\underline{v}_n = (1 - \eta_n) \underline{v}_n^N + \eta_n \underline{v}_n^H \tag{4.179}$$

where η_n is a random variable independent of $\underline{\theta}_0$ and $\{\underline{w}_n\}$ obeying

$$\eta_n = \begin{cases} 0 & \text{w.p.} \quad (1 - \varepsilon) \\ 1 & \text{w.p.} \quad \varepsilon \end{cases} \tag{4.180}$$

and $\{\underline{v}_n^N\}$ and $\{\underline{v}_n^H\}$ are random variables independent of $\{\eta_n\}$, $\underline{\theta}_0$, and $\{\underline{w}_n\}$ with $L(\underline{v}_n^N) = N(0, R)$ (for some $R > 0$) and $L(\underline{v}_n^H) = H$. Then, neglecting for a moment the question of ω, it is possible to interpret equation (4.160) as follows:

$$\begin{aligned}
& p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}) \, p(\underline{z}_0, \cdots, \underline{z}_{n-1}) \\
&\quad = p(\eta_0 = 0, \cdots, \eta_{n-1} = 0) \, p(\underline{z}_0, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{n-1} = 0) \\
&\qquad\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0) \\
&\quad + \sum_{i=1}^{n} p(\eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, , \eta_{n-1} = 0) \\
&\qquad\qquad p(\underline{z}_0, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, , \eta_{n-1} = 0) \\
&\qquad\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, , \eta_{n-1} = 0) \\
&\quad + \text{higher-order terms.}
\end{aligned} \tag{4.181}$$

In other words, loosely defining a random variable distributed as H as an "outlier," the first term in (4.160) and (4.181) corresponds to the event that "there has been no outlier among the first n observations," each term in the summation to the event "there has been no outlier among the first n observations except for one, at time $i - 1$," and higher-order terms to the occurrence of two or more outliers. Moreover,

$$\begin{aligned}
& p(\underline{z}_0, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0) \\
&\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0) \\
&\quad = p(\underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0) \\
&\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \\
&\qquad\qquad\qquad \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0) \\
&\qquad p(\underline{z}_{i-1} \mid \underline{\theta}_n, \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1},
\end{aligned}$$

$$\eta_0=0, \; \cdots, \eta_{i-1}=1, \; \cdots, \eta_{n-1}=0 \;) \tag{4.182}$$

Note that the *only* non-normal term on the right-hand side of (4.182) is the last one, and it corresponds to the convolution in equations (4.160) and (4.181). 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 (4.182), on the other hand, is the distribution of a normal random variable (the state $\underline{\theta}_n$) conditioned on normal observations $\{ \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1} \}$. It therefore is a normal distribution, whose mean and variance are given by the Kalman Filter that skips the observation $\underline{z}_{i-1}$.

(v) Evidently, as $n \rightarrow \infty$, 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 (4.160) 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.

(vi) Finally, it is easy to show that

$$\begin{aligned}
&(1 - \varepsilon)^n \; \kappa_n \; \kappa_n^0 \\
&\quad = \frac{p(\eta_0=0, \; \cdots, \eta_{n-1}=0 \;) \, p(\underline{z}_0, \; \cdots, \underline{z}_{n-1} \mid \eta_0=0, \; \cdots, \eta_{n-1}=0 \;)}{p(\underline{z}_0, \; \cdots, \underline{z}_{n-1} \;)} && (4.183) \\
&\quad = p(\eta_0=0, \; \cdots, \eta_{n-1}=0 \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) && (4.184)
\end{aligned}$$

is the posterior probability, conditioned on all past observations $\{ \underline{z}_0, \cdots, \underline{z}_{n-1} \}$, that no outliers have occurred among the first n observations. Similarly, it is easy to show that

$$\begin{aligned}
&\varepsilon \, (1 - \varepsilon)^{n-1} \; \kappa_n \; \kappa_n^i \int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i , W_n^i \;) \, h(\xi) \, d\xi \\
&\quad = p(\eta_0=0, \; \cdots, \eta_{i-1}=1, \; \cdots, \eta_{n-1}=0 \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) && (4.185)
\end{aligned}$$

is the posterior probability that exactly one outlier occurred, at time $i-1$. Thus, equation (4.160) may be interpreted also as a weighted sum of conditional distributions, with weights equal to the posterior probability that each event has occurred.

Proof The proof of Theorem 4.2 proceeds by induction. Note first that

$$\begin{aligned}
&p(\underline{\theta}_{n+1} \mid \underline{z}_0, \; \cdots, \underline{z}_n \;) \, p(\underline{z}_n \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) \\
&\quad = p(\underline{\theta}_{n+1}, \underline{z}_n \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) && (4.186) \\
&\quad = \int p(\underline{\theta}_n, \underline{\theta}_{n+1}, \underline{z}_n \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) \, d\underline{\theta}_n && (4.187) \\
&\quad = \int p(\underline{\theta}_{n+1} \mid \underline{\theta}_n, \underline{z}_0, \; \cdots, \underline{z}_n \;) \\
&\qquad\qquad p(\underline{z}_n \mid \underline{\theta}_n, \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) \, p(\underline{\theta}_n \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1} \;) \, d\underline{\theta}_n && (4.188)
\end{aligned}$$

$$= \int p(\,\underline{\theta}_{n+1} \mid \underline{\theta}_n\,)\, p(\,\underline{z}_n \mid \underline{\theta}_n\,)\, p(\,\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}\,)\, d\underline{\theta}_n, \tag{4.189}$$

where (4.186) and (4.188) follow from the definition of the conditional probability, (4.187) from that of the marginal probability, and (4.189) from (4.1)-(4.2) and the independence of $\{\underline{w}_n\}$ and $\{\underline{v}_n\}$. In particular, for $n = 0$, equation (4.189) yields

$$p(\,\underline{\theta}_1 \mid \underline{z}_0\,)\, p(\,\underline{z}_0\,)$$

$$= \int \mathrm{N}(\,\underline{\theta}_1;\, F_0\underline{\theta}_0,\, Q_0\,)\, f(\,\underline{z}_0 - \underline{\theta}_0\,)\, \mathrm{N}(\,\underline{\theta}_0;\, \overline{\underline{\theta}}_0,\, M_0\,)\, d\underline{\theta}_0 \tag{4.190}$$

$$= \int \mathrm{N}(\,\underline{\theta}_1;\, F_0\underline{\theta}_0,\, Q_0\,) \Big[(1-\varepsilon)\mathrm{N}(\,\underline{z}_0 - \underline{\theta}_0;\, 0, R\,) + \varepsilon\, h(\underline{z}_0 - \underline{\theta}_0) \Big]$$

$$\mathrm{N}(\,\underline{\theta}_0;\, \overline{\underline{\theta}}_0,\, M_0\,)\, d\underline{\theta}_0 \tag{4.191}$$

where f denotes the Radon-Nikodym derivative of P (which exists since both $\mathrm{N}(\,0, R\,)$ and H are absolutely continuous with respect to the Lebesgue measure), (4.1)-(4.2) as well as the initial condition of (4.1) are used in (4.190), and (4.157) is used in (4.191). But for any n,

$$\int \mathrm{N}(\,\underline{\theta}_{n+1};\, F_n\,\underline{\theta}_n;\, Q_n\,)\, \mathrm{N}(\,\underline{z}_n - \underline{\theta}_n;\, 0, R\,)\, \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0\,)\, d\underline{\theta}_n$$

$$= \int \mathrm{N}(\,F_n\,\underline{\theta}_n;\, \underline{\theta}_{n+1},\, Q_n\,)\, \mathrm{N}(\,\underline{\theta}_n;\, \underline{z}_n,\, R\,)\, \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0\,)\, d\underline{\theta}_n \tag{4.192}$$

$$= \int \mathrm{N}(\,\underline{\theta}_n;\, \underline{z}_n + R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}(\,\underline{\theta}_{n+1} - F_n\,\underline{z}_n\,),$$

$$R - R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}F_n\,R\,)$$

$$\mathrm{N}(\,\underline{\theta}_{n+1};\, F_n\,\underline{z}_n,\, Q_n + F_n\,R\,F_n^{\mathrm{T}}\,)\, \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0\,)\, d\underline{\theta}_n \tag{4.193}$$

$$= \mathrm{N}(\,\underline{z}_n + R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}(\,\underline{\theta}_{n+1} - F_n\,\underline{z}_n\,);\, \underline{\theta}_n^0,$$

$$R - R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}F_n\,R + M_n^0\,)$$

$$\mathrm{N}(\,\underline{\theta}_{n+1};\, F_n\,\underline{z}_n,\, Q_n + F_n\,R\,F_n^{\mathrm{T}}\,) \quad . \tag{4.194}$$

$$= \mathrm{N}(\,R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}\underline{\theta}_{n+1};\, \underline{\theta}_n^0 - \underline{z}_n + R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}}\,)^{-1}F_n\,\underline{z}_n,$$

$$R - R\,F_n^{\mathrm{T}}(\,Q_n + F_n\,R\,F_n^{\mathrm{T}})^{-1}F_n\,R + M_n^0\,)$$

$$\mathrm{N}(\,\underline{\theta}_{n+1};\, F_n\,\underline{z}_n,\, Q_n + F_n\,R\,F_n^{\mathrm{T}}\,) \tag{4.195}$$

$$= \mathrm{N}(\,\underline{\theta}_{n+1};\, F_n\,\underline{z}_n + F_n\,R\,(M_n^0 + R\,)^{-1}(\,\underline{\theta}_n^0 - \underline{z}_n\,),$$

$$Q_n + F_n\,R\,F_n^{\mathrm{T}} - F_n\,R\,(M_n^0 + R\,)^{-1}R\,F_n^{\mathrm{T}}\,)$$

$$\mathrm{N}(\,\underline{z}_n;\, \underline{\theta}_n^0,\, M_n^0 + R\,), \tag{4.196}$$

where (4.192) and (4.195) are obtained by rearranging terms, (4.193), (4.194), and (4.196) follow from repeated applications of Lemma 4.2, and (4.194) from the fact that the distribution integrates to unity. Furthermore,

$$F_n\,\underline{z}_n + F_n\,R\,(M_n^0 + R\,)^{-1}(\,\underline{\theta}_n^0 - \underline{z}_n\,)$$

$$= F_n \Big[(M_n^0 + R\,)(M_n^0 + R\,)^{-1}\underline{z}_n - R\,(M_n^0 + R\,)^{-1}\underline{z}_n$$

$$+ \; R\,(M_n^0+R\,)^{-1}\underline{\theta}_n^0 \; + \; (M_n^0+R\,)(M_n^0+R\,)^{-1}\underline{\theta}_n^0$$

$$- \; (M_n^0+R\,)(M_n^0+R\,)^{-1}\underline{\theta}_n^0 \;\Big] \tag{4.197}$$

$$= F_n \left[M_n^0(M_n^0+R\,)^{-1}\underline{z}_n \; + \; \underline{\theta}_n^0 \; - \; M_n^0(M_n^0+R\,)^{-1}\underline{\theta}_n^0 \right] \tag{4.198}$$

$$= F_n\,\underline{\theta}_n^0 \; + \; F_n\,M_n^0\Gamma_n^{0\;-1}(\,\underline{z}_n - \underline{\theta}_n^0\,) \tag{4.199}$$

$$= \underline{\theta}_{n+1}^0, \tag{4.200}$$

from (4.161). Similarly,

$$Q_n \; + \; F_n\,R\,F_n^{\mathrm{T}} \; - \; F_n\,R\,(M_n^0+R\,)^{-1}R\,F_n^{\mathrm{T}}$$

$$= Q_n \; + \; F_n \left[R\,(M_n^0+R\,)^{-1}(M_n^0+R\,) \; - \; R\,(M_n^0+R\,)^{-1}R \;\right] F_n^{\mathrm{T}} \tag{4.201}$$

$$= Q_n \; + \; F_n\,R\,(M_n^0+R\,)^{-1}M_n^0\,F_n^{\mathrm{T}} \tag{4.202}$$

$$= Q_n \; + \; F_n\,(M_n^{0\;-1}+R^{-1})^{-1}F_n^{\mathrm{T}} \tag{4.203}$$

$$= Q_n \; + \; F_n \left[M_n^0 - M_n^0(M_n^0+R\,)M_n^0 \;\right] F_n^{\mathrm{T}} \tag{4.204}$$

$$= Q_n \; + \; F_n\,P_n^0\,F_n^{\mathrm{T}} \tag{4.205}$$

$$= M_{n+1}^0\,, \tag{4.206}$$

where (4.204) follows from the Sherman-Morrison-Woodbury theorem, (4.205) from (4.163)-(4.164), and (4.206) from (4.162). It therefore follows that

$$\int \mathrm{N}(\,\underline{\theta}_{n+1};\,F_n\,\underline{\theta}_n\,,\,Q_n\;)\,\mathrm{N}(\,\underline{z}_n - \underline{\theta}_n\,;\,0,\,R\;)\,\mathrm{N}(\,\underline{\theta}_n\,;\,\underline{\theta}_n^0,\,M_n^0\;)\,d\underline{\theta}_n$$

$$= \mathrm{N}(\,\underline{\theta}_{n+1};\,\underline{\theta}_{n+1}^0,\,M_{n+1}^0\;)\,\mathrm{N}(\,\underline{z}_n\,;\,\underline{\theta}_n^0,\,\Gamma_n^0\,), \tag{4.207}$$

from (4.164), (4.196), (4.200), and (4.206).

Going back to equation (4.191),

$$\int \mathrm{N}(\,\underline{\theta}_{n+1};\,F_n\,\underline{\theta}_n\,.\,Q_n\;)\,h(\,\underline{z}_n - \underline{\theta}_n\,)\,\mathrm{N}(\,\underline{\theta}_n\,:\,\underline{\theta}_n^0,\,M_n^0\;)\,d\underline{\theta}_n$$

$$= \int \mathrm{N}(\,\underline{\theta}_n\,;\,\underline{\theta}_n^0 + M_n^0\,F_n^{\mathrm{T}}(\,Q_n + F_n\,M_n^0\,F_n^{\mathrm{T}}\,)^{-1}(\underline{\theta}_{n+1} - F_n\,\underline{\theta}_n^0),$$

$$M_n^0 - M_n^0\,F_n^{\mathrm{T}}(\,Q_n + F_n\,M_n^0\,F_n^{\mathrm{T}}\,)^{-1}F_n\,M_n^0\,)$$

$$\mathrm{N}(\,\underline{\theta}_{n+1};\,F_n\,\underline{\theta}_n^0,\,Q_n \; + F_n\,M_n^0\,F_n^{\mathrm{T}}\;)\,h(\,\underline{z}_n - \underline{\theta}_n\,)\,d\underline{\theta}_n \tag{4.208}$$

$$= \mathrm{N}(\,\underline{\theta}_{n+1};\,F_n\,\underline{\theta}_n^0,\,Q_n \; + F_n\,M_n^0\,F_n^{\mathrm{T}}\;)$$

$$\int \mathrm{N}(\,\underline{z}_n \; - \underline{\xi};\,\underline{\theta}_n^0 + M_n^0\,F_n^{\mathrm{T}}(\,Q_n + F_n\,M_n^0\,F_n^{\mathrm{T}})^{-1}(\underline{\theta}_{n+1} - F_n\,\underline{\theta}_n^0),$$

$$M_n^0 - M_n^0\,F_n^{\mathrm{T}}(\,Q_n + F_n\,M_n^0\,F_n^{\mathrm{T}}\,)^{-1}F_n\,M_n^0\,)\,h(\,\underline{\xi}\,)\,d\underline{\xi} \tag{4.209}$$

$$= \mathrm{N}(\,\underline{\theta}_{n+1};\,\underline{\theta}_{n+1}^{n+1},\,M_{n+1}^{n+1}\;)$$

$$\int \mathrm{N}(\, \underline{z}_n - \underline{\xi};\ \underline{v}_{n+1}^{n+1} + V_{n+1}^{n+1}\,(\underline{\theta}_{n+1} - \underline{\theta}_{n+1}^{n+1}),$$
$$W_{n+1}^{n+1} + V_{n+1}^{n+1}\, M_{n+1}^{n+1}\, V_{n+1}^{n+1\ \mathrm{T}}\,)\; h(\underline{\xi})\; d\underline{\xi}, \tag{4.210}$$

where (4.208) follows from rearranging terms and using Lemma 4.2, (4.209) from making the substitution $\underline{\xi} = \underline{z}_n - \underline{\theta}_n$, and (4.210) from (4.169)-(4.173) with $i = n+1$.

Substituting equations (4.207) and (4.210) with $n = 0$ into (4.191), and using (4.175)-(4.176), yields

$$p(\,\underline{\theta}_1 \mid \underline{z}_0\,)\, p(\,\underline{z}_0\,)$$
$$= (1-\varepsilon)\mathrm{N}(\,\underline{z}_0;\, \overline{\underline{\theta}}_0,\, \Gamma_0^0\,)\, \mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^0,\, M_1^0\,)$$
$$+\ \varepsilon\mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^1,\, M_1^1\,)$$
$$\int \mathrm{N}(\,\underline{z}_0 - \underline{\xi};\ \underline{v}_1^1 + V_1^1\,(\underline{\theta}_1 - \underline{\theta}_1^1),\ W_1^1 + V_1^1\, M_1^1\, V_1^{1\ \mathrm{T}}\,)\; h(\underline{\xi})\; d\underline{\xi}. \tag{4.211}$$

But since

$$\int p(\,\underline{\theta}_n \mid \underline{z}_0,\ \cdots,\ \underline{z}_{n-1}\,)\, p(\,\underline{z}_0,\ \cdots,\ \underline{z}_{n-1}\,)\; d\underline{\theta}_n$$
$$= \int p(\,\underline{\theta}_n,\, \underline{z}_0,\ \cdots,\ \underline{z}_{n-1}\,)\; d\underline{\theta}_n \tag{4.212}$$
$$= p(\,\underline{z}_0,\ \cdots,\ \underline{z}_{n-1}\,), \tag{4.214}$$

respectively by the definitions of conditional and marginal probabilities, and since

$$\int \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^i,\, M_n^i\,)\, \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i - V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i),\ W_n^i - V_n^i\, M_n^i\, V_n^{i\ \mathrm{T}}\,)\; d\underline{\theta}_n$$
$$= \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i,\ W_n^i\,) \tag{4.214}$$

by Lemma 4.2, it follows that

$$p(\,\underline{z}_0\,) = \int (1-\varepsilon)\mathrm{N}(\,\underline{z}_0;\, \overline{\underline{\theta}}_0,\, \Gamma_0^0\,)\, \mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^0,\, M_1^0\,)$$
$$+\ \varepsilon\mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^1,\, M_1^1\,)$$
$$\int \mathrm{N}(\,\underline{z}_0 - \underline{\xi};\ \underline{v}_1^1 + V_1^1\,(\underline{\theta}_1 - \underline{\theta}_1^1),$$
$$W_1^1 + V_1^1\, M_1^1\, V_1^{1\ \mathrm{T}}\,)\; h(\underline{\xi})\; d\underline{\xi}\; d\underline{\theta}_1 \tag{4.215}$$
$$= (1-\varepsilon)\mathrm{N}(\,\underline{z}_0;\, \overline{\underline{\theta}}_0,\, \Gamma_0^0\,)\ +\ \varepsilon \int \mathrm{N}(\,\underline{z}_0 - \underline{\xi};\ \underline{v}_1^1,\ W_1^1\,)\; d\underline{\xi} \tag{4.216}$$
$$:= \kappa_0^{-1}, \tag{4.217}$$

where the interchange of the order of integration of $\underline{\theta}_1$ and $\underline{\xi}$ is justified by Fubini's theorem (since both the normal density and h are Lebesgue-integrable). Thus, combining (4.211) and (4.217), and using (4.168), (4.174), and (4.177), it follows that

$$p(\,\underline{\theta}_1 \mid \underline{z}_0\,) = (1-\varepsilon)\,\kappa_0\,\kappa_0^0\, \mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^0,\, M_1^0\,)\ +\ \varepsilon\,\kappa_0\,\kappa_0^1\, \mathrm{N}(\,\underline{\theta}_1;\, \underline{\theta}_1^1,\, M_1^1\,)$$
$$\int \mathrm{N}(\,\underline{z}_0 - \underline{\xi};\ \underline{v}_1^1 + V_1^1\,(\underline{\theta}_1 - \underline{\theta}_1^1),\ W_1^1 + V_1^1\, M_1^1\, V_1^{1\ \mathrm{T}}\,)\; h(\underline{\xi})\; d\underline{\xi}. \tag{4.218}$$

Assume now by the induction argument that

$$
\begin{aligned}
p(\,\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}\,)
&= (\,1-\varepsilon\,)^n \,\kappa_n\, \kappa_n^0\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0\,) \\
&\quad + \varepsilon(\,1-\varepsilon\,)^{n-1} \kappa_n \sum_{i=1}^{n} \kappa_n^i\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i\,) \\
&\qquad \int \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi} \\
&\quad + O_p(\,\varepsilon^2 (1-\varepsilon)^{n-2}\,)
\end{aligned}
\tag{4.219}
$$

for some n. From (4.189),

$$
\begin{aligned}
&p(\,\underline{\theta}_{n+1} \mid \underline{z}_0, \cdots, \underline{z}_n\,)\, p(\,\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}\,) \\
&= \int \mathrm{N}(\,\underline{\theta}_{n+1} ; F_n\, \underline{\theta}_n, Q_n\,) \Big[(1-\varepsilon)\mathrm{N}(\,\underline{z}_n - \underline{\theta}_n ; 0, R\,) + \varepsilon\, h(\underline{z}_n - \underline{\theta}_n\,) \Big] \\
&\qquad \Big[(\,1-\varepsilon\,)^n \,\kappa_n\, \kappa_n^0\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0\,) \\
&\qquad\quad + \varepsilon(\,1-\varepsilon\,)^{n-1} \kappa_n \sum_{i=1}^{n} \kappa_n^i\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i\,) \\
&\qquad\qquad \int \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi} \\
&\qquad\quad + O_p(\,\varepsilon^2 (1-\varepsilon)^{n-2}\,) \Big]\, d\underline{\theta}_n .
\end{aligned}
\tag{4.220}
$$

Now:

$$
\begin{aligned}
&\int \mathrm{N}(\,\underline{\theta}_{n+1} ; F_n\, \underline{\theta}_n, Q_n\,)\, \mathrm{N}(\,\underline{z}_n - \underline{\theta}_n ; 0, R\,)\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i\,) \\
&\qquad \int \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi}\ \ d\underline{\theta}_n \\
&= \iint \mathrm{N}(\,F_n\, \underline{\theta}_n ; \underline{\theta}_{n+1}, Q_n\,)\, \mathrm{N}(\,\underline{\theta}_n ; \underline{z}_n, R\,)\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i\,) \\
&\qquad \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi}\ \ d\underline{\theta}_n
\end{aligned}
\tag{4.221}
$$

$$
\begin{aligned}
&= \iint \mathrm{N}(\,\underline{\theta}_n ; \underline{z}_n + R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} (\underline{\theta}_{n+1} - F_n\, \underline{z}_n\,), \\
&\qquad\qquad R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,) \\
&\qquad \mathrm{N}(\,\underline{\theta}_{n+1} ; F_n\, \underline{z}_n, Q_n + F_n\, R\, F_n^{\mathrm{T}}\,)\, \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i\,) \\
&\qquad \mathrm{N}(\,\underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi}\ \ d\underline{\theta}_n
\end{aligned}
\tag{4.222}
$$

$$
\begin{aligned}
&= \iint \mathrm{N}(\,\underline{\theta}_n ; \underline{\theta}_n^i + M_n^i\,(M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1} \\
&\qquad\qquad (\underline{z}_n + R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} (\underline{\theta}_{n+1} - F_n\, \underline{z}_n\,) - \underline{\theta}_n^i\,), \\
&\qquad\qquad M_n^i - M_n^i\,(M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1} M_n^i\,) \\
&\qquad \mathrm{N}(\,R\, F_n^{\mathrm{T}} (\,Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} \underline{\theta}_{n+1} ; R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, \underline{z}_n - \underline{z}_n + \underline{\theta}_n^i,
\end{aligned}
$$

$$M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R \;)$$

$$\mathrm{N}(\, \underline{\theta}_{n+1};\, F_n\, \underline{z}_n,\, Q_n + F_n\, R\, F_n^{\mathrm{T}} \,)$$

$$\mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{v}_n^i + V_n^i\, (\underline{\theta}_n - \underline{\theta}_n^i),\, W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\; h(\underline{\xi})\, d\underline{\xi}\;\; d\underline{\theta}_n \tag{4.223}$$

$$= \mathrm{N}(\, R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} \underline{\theta}_{n+1};\, R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, \underline{z}_n - \underline{z}_n + \underline{\theta}_n^i,$$

$$M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R \;)$$

$$\mathrm{N}(\, \underline{\theta}_{n+1};\, F_n\, \underline{z}_n,\, Q_n + F_n\, R\, F_n^{\mathrm{T}} \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{v}_n^i + V_n^i\, (M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1}$$

$$(\underline{z}_n + R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} (\underline{\theta}_{n+1} - F_n\, \underline{z}_n\,) - \underline{\theta}_n^i\,)),$$

$$W_n^i - V_n^i M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1} M_n^i\, V_n^{i\,\mathrm{T}} \,)$$

$$h(\underline{\xi})\, d\underline{\xi} \tag{4.224}$$

$$= \mathrm{N}(\, \underline{\theta}_{n+1};\, F_n\, \underline{z}_n + F_n\, R\, (M_n^i + R\,)^{-1} (\underline{\theta}_n^i - \underline{z}_n\,),$$

$$Q_n + F_n\, R\, F_n^{\mathrm{T}} - F_n\, R\, (M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} \,)$$

$$\mathrm{N}(\, \underline{z}_n\, ;\, \underline{\theta}_n^i,\, M_n^i + R \;)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{v}_n^i + V_n^i\, (M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1}$$

$$(\underline{z}_n + R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} (\underline{\theta}_{n+1} - F_n\, \underline{z}_n\,) - \underline{\theta}_n^i\,)),$$

$$W_n^i - V_n^i M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1} M_n^i\, V_n^{i\,\mathrm{T}} \,)$$

$$h(\underline{\xi})\, d\underline{\xi} \tag{4.225}$$

$$= \mathrm{N}(\, \underline{\theta}_{n+1};\, \underline{\theta}_{n+1}^i,\, M_{n+1}^i \;)\; \mathrm{N}(\, \underline{z}_n\, ;\, \underline{\theta}_n^i,\, \Gamma_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{v}_n^i + V_n^i\, (M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1}$$

$$(\underline{z}_n + R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} (\underline{\theta}_{n+1} - F_n\, \underline{z}_n\,) - \underline{\theta}_n^i\,)),$$

$$W_n^i - V_n^i M_n^i\, (M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R\,)^{-1} M_n^i\, V_n^{i\,\mathrm{T}} \,)$$

$$h(\underline{\xi})\, d\underline{\xi}, \tag{4.226}$$

where (4.221) is obtained by rearranging terms, (4.222)-(4.225) follow from repeated applications of Lemma 4.2, (4.223) from the fact that the distribution integrates to unity (the interchange of the order of integration of $\underline{\theta}_n$ and $\underline{\xi}$ is justified by Fubini's theorem), and (4.226) from (4.200) and (4.206) with the superscript 0 replaced by i.

Now, the coefficient of $\underline{\theta}_{n+1}$ in the integrand of (4.226) is

$$V_n^i\, M_n^i\, \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1} F_n\, R \;\right]^{-1} R\, F_n^{\mathrm{T}} (Q_n + F_n\, R\, F_n^{\mathrm{T}})^{-1}$$

$$= V_n^i\, M_n^i\, \left[(M_n^i + R\,)^{-1} + (M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} \left[(Q_n + F_n\, R\, F_n^{\mathrm{T}}) \right.\right.$$

$$\begin{aligned} &- \; F_n R\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} \Big]^{-1} F_n R\,(M_n^i + R\,)^{-1} \Big] \\ &\qquad R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \end{aligned} \tag{4.227}$$

$$\begin{aligned} &= V_n^i M_n^i \Big[(M_n^i + R\,)^{-1} + (M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1} F_n R\,(M_n^i + R\,)^{-1} \Big] \\ &\qquad R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \end{aligned} \tag{4.228}$$

$$\begin{aligned} &= V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \\ &\quad + \; V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1} F_n R\,(M_n^i + R\,)^{-1} \\ &\qquad R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \end{aligned} \tag{4.229}$$

$$\begin{aligned} &= V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1} M_{n+1}^i\,(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \\ &\quad + \; V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1} F_n R\,(M_n^i + R\,)^{-1} \\ &\qquad R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \end{aligned} \tag{4.230}$$

$$= V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1}(Q_n + F_n R\, F_n^{\mathrm{T}})(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \tag{4.231}$$

$$= V_n^i M_n^i\,(M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\;-1} \tag{4.232}$$

$$= V_n^i\,(M_n^{i\;-1} + R^{-1})^{-1} F_n^{\mathrm{T}} M_{n+1}^{i\;-1} \tag{4.233}$$

$$= V_n^i \Big[M_n^i \; - \; M_n^i\,(M_n^i - R\,)^{-1} M_n^i \Big] F_n^{\mathrm{T}} M_{n+1}^{i\;-1} \tag{4.234}$$

$$= V_n^i P_n^i F_n^{\mathrm{T}} M_{n+1}^{i\;-1} \tag{4.235}$$

$$= V_{n+1}^i\,, \tag{4.236}$$

where (4.227) and (4.234) follow from the Sherman-Morrison-Woodbury theorem, (4.228) and (4.231) from (4.206) with superscript i, (4.235) from (4.163), and (4.236) from (4.165).

Using (4.236) and rewriting the mean of the normal distribution in the integrand of (4.226) as

$$\begin{aligned} &\underline{\mathbf{v}}_n^i \; + \; V_{n+1}^i\,(\underline{\theta}_{n+1} - \underline{\theta}_{n+1}^i) \; + \; V_{n+1}^i\,\underline{\theta}_{n+1}^i \\ &\qquad - \; V_n^i M_n^i \Big[M_n^i + R - R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \Big]^{-1} \underline{\theta}_n^i \\ &\qquad + \; V_n^i M_n^i \Big[M_n^i + R - R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \Big]^{-1} \\ &\qquad\qquad \Big[I - R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n \Big] \underline{z}_n \\ &\quad = \underline{\mathbf{v}}_n^i \; + \; V_{n+1}^i\,(\underline{\theta}_{n+1} - \underline{\theta}_{n+1}^i) \; + \; V_{n+1}^i \Big[F_n \underline{\theta}_n^i + F_n M_n^i \Gamma_n^{i\,-1}\,(\,\underline{z}_n - \underline{\theta}_n^i\,) \Big] \\ &\qquad - \; V_n^i M_n^i \Big[M_n^i + R - R\, F_n^{\mathrm{T}}(Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \Big]^{-1} \underline{\theta}_n^i \end{aligned}$$

$$+ \; V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[I - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n \right] \underline{z}_n , \tag{4.237}$$

it follows that the coefficient of $\underline{z}_n$ is

$$V_{n+1}^i F_n M_n^i \Gamma_n^{i-1} + V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[I - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n \right]$$

$$= V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n M_n^i (M_n^i + R)^{-1} \right.$$

$$\left. + \; I - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n \right] \tag{4.238}$$

$$= V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n \left[M_n^i (M_n^i + R)^{-1} \right. \right.$$

$$\left. \left. - \; (M_n^i + R)(M_n^i + R)^{-1} \right] + I \right] \tag{4.239}$$

$$= V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[- R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R\, (M_n^i + R)^{-1} \right.$$

$$\left. + \; (M_n^i + R)(M_n^i + R)^{-1} \right] \tag{4.240}$$

$$= V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right] (M_n^i + R)^{-1} \tag{4.241}$$

$$= V_n^i M_n^i (M_n^i + R)^{-1}, \tag{4.242}$$

where (4.238) follows from (4.236). Similarly, the coefficient of $\underline{\theta}_n^i$ is

$$V_n^i M_n^i \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1}$$

$$\left[R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} \left[F_n - F_n M_n^i (M_n^i + R)^{-1} \right] - I \right]$$

$$= - V_n^i M_n^i (M_n^i + R)^{-1}, \tag{4.243}$$

as in (4.239)-(4.242). It follows, therefore, that the mean of the normal distribution in the integrand of (4.226) is

$$\underline{v}_n^i + V_{n+1}^i (\underline{\theta}_{n+1} - \underline{\theta}_{n+1}^i) + V_n^i M_n^i (M_n^i + R)^{-1} (\underline{z}_n - \underline{\theta}_n^i)$$

$$= \underline{v}_{n+1}^i + V_{n+1}^i (\underline{\theta}_{n+1} - \underline{\theta}_{n+1}^i), \tag{4.244}$$

from (4.166).

Finally, the covariance of the normal distribution in the integrand of (4.226) is given by

$$\begin{aligned}
W_n^i \; - \; V_n^i M_n^i & \left[M_n^i + R - R\, F_n^{\mathrm{T}} (Q_n + F_n R\, F_n^{\mathrm{T}})^{-1} F_n R \right]^{-1} M_n^i V_n^{i\ \mathrm{T}} \\
&= W_n^i \; - \; V_n^i M_n^i \left[(M_n^i + R\,)^{-1} \right. \\
&\qquad + (M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} \left[Q_n + F_n R\, F_n^{\mathrm{T}} \right. \\
&\qquad\qquad \left. \left. F_n R\, (\, M_n^i + R\,)^{-1} R\, F_n^{\mathrm{T}} \right]^{-1} F_n R\, (M_n^i + R\,)^{-1} \right] M_n^i V_n^{i\ \mathrm{T}} && (4.245) \\
&= W_n^i - V_n^i M_n^i \left[\Gamma_n^{i\ -1} + \Gamma_n^{i\ -1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\ \ -1} F_n R\, \Gamma_n^{i\ -1} \right] M_n^i V_n^{i\ \mathrm{T}} && (4.246) \\
&= W_n^i \; - \; V_n^i M_n^i \Gamma_n^{i\ -1} M_n^i V_n^{i\ \mathrm{T}} \; - \; V_n^i M_n^i \Gamma_n^{i\ -1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\ \ -1} F_n R\, \Gamma_n^{i\ -1} M_n^i V_n^{i\ \mathrm{T}} && (4.247) \\
&= W_{n+1}^i \; - \; V_n^i M_n^i \Gamma_n^{i\ -1} R\, F_n^{\mathrm{T}} M_{n+1}^{i\ \ -1} M_{n+1}^i M_{n+1}^{i\ \ -1} F_n R\, \Gamma_n^{i\ -1} M_n^i V_n^{i\ \mathrm{T}} && (4.248) \\
&= W_{n+1}^i \; - \; V_{n+1}^i M_{n+1}^i V_{n+1}^{i\ \mathrm{T}}, && (4.249)
\end{aligned}$$

from (4.167) and (4.165). Thus, substituting (4.244) and (4.249) into (4.226), and this latter in turn into (4.220), along with (4.207) and (4.210), and finally noting that

$$\begin{aligned}
\kappa_{n+1} &= \frac{1}{p(\, \underline{z}_0, \;\cdots, \underline{z}_n\,)} && (4.250) \\
&= \frac{1}{p(\, \underline{z}_n \mid \underline{z}_0, \;\cdots, \underline{z}_{n-1}\,)\, p(\, \underline{z}_0, \;\cdots, \underline{z}_{n-1}\,)} && (4.251) \\
&= \frac{\kappa_n}{p(\, \underline{z}_n \mid \underline{z}_0, \;\cdots, \underline{z}_{n-1}\,)}, && (4.252)
\end{aligned}$$

establishes the validity of (4.219) for all n.

There remains to put (4.219) into the form of (4.161), and to show that the error term remains bounded as $n \to \infty$. This proof exploits the exponential asymptotic stability of the Kalman Filter, demonstrated in Theorem 4.1 and its corollary.

Consider first the case where only one outlier occurs during the first n time steps. If it occurs early enough, its effects will have become negligible by time n, and hence the corresponding term can be lumped up with the "no outlier during the first n time steps" term.

By Corollary 4.1,

$$\mathrm{N}(\, \underline{\theta}_n \,;\, \underline{\theta}_n^i, M_n^i \,) \; = \; \mathrm{N}(\, \underline{\theta}_n \,;\, \underline{\theta}_n^0, M_n^0 \,) \; + \; O_p(\, \delta^{n-i} \,). \qquad (4.253)$$

Furthermore, equations (4.165) and (4.163) imply that

$$\begin{aligned}
V_n^i &= V_{n-1}^i \,(I - M_{n-1}^i \,\Gamma_{n-1}^{i\ \ -1}) M_{n-1}^i \, F_{n-1}^{\mathrm{T}} \, M_n^{i\ -1} && (4.254) \\
&= V_{n-1}^i \, M_{n-1}^i \,(I - M_{n-1}^i \,\Gamma_{n-1}^{i\ \ -1})^{\mathrm{T}} F_{n-1}^{\mathrm{T}} \, M_n^{i\ -1} && (4.255) \\
&= V_{n-2}^i \, M_{n-2}^i \,(I - M_{n-2}^i \,\Gamma_{n-2}^{i\ \ -1})^{\mathrm{T}} F_{n-2}^{\mathrm{T}} \, M_{n-1}^{i\ \ -1} M_{n-1}^i \,(I - M_{n-1}^i \,\Gamma_{n-1}^{i\ \ -1})^{\mathrm{T}} F_{n-1}^{\mathrm{T}} \, M_n^{i\ -1} && (4.256)
\end{aligned}$$

$$= V_i^i M_i^i \left[\prod_{j=i}^{n-1} F_j (I - M_j^i \Gamma_j^{i\ -1}) \right]^{\mathrm{T}} M_n^{i\ -1} \tag{4.257}$$

$$= M_{i-1}^0 \, F_{i-1}^{\mathrm{T}} \, M_i^{i\ -1} M_i^i \left[\prod_{j=i}^{n-1} F_j (I - M_j^i \Gamma_j^{i\ -1}) \right]^{\mathrm{T}} M_n^{i\ -1} \tag{4.258}$$

$$= M_{i-1}^0 \, F_{i-1}^{\mathrm{T}} \left[\prod_{j=i}^{n-1} F_j (I - M_j^i \Gamma_j^{i\ -1}) \right]^{\mathrm{T}} M_n^{i\ -1} \tag{4.259}$$

$$= M_{i-1}^0 \left[\prod_{j=i}^{n-1} (I - M_j^i \Gamma_j^{i\ -1}) F_{j-1} \right]^{\mathrm{T}} F_{n-1}^{\mathrm{T}} \, M_n^{i\ -1} \tag{4.260}$$

$$= O(\delta^{n-i}), \tag{4.261}$$

where (4.255) holds by symmetry, (4.258) follows from (4.171), and (4.261) from Theorem 4.1 and the facts that $\{F_n\}$ is bounded above by hypothesis, and $\{M_n^0\}$ and $\{M_n^i\}$ are bounded above and below, by equations (4.57) and (4.90). (Note that both bounds carry over to $\{M_n^i\}$ -- which skips an observation update -- because $(I - K_{i-1} H_{i-1})$ is bounded both above and below.) Thus, it also holds that

$$\mathrm{N}(\underline{z}_{i-1} - \xi; \underline{\nu}_n^i + V_n^i (\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}})$$
$$= \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{\nu}_n^i, W_n^i) + O_p(\delta^{n-i}) \tag{4.262}$$

from (4.127), where use is again made of the fact that $\{M_n^i\}$ is bounded above, and also of the boundedness w.p.1 of $\underline{\theta}_n$ and $\underline{\theta}_n^i$, due to the bound (4.103) on the transition matrix. Hence, each term in the summation in (4.219) may be written as

$$\varepsilon(1-\varepsilon)^{n-1} \kappa_n \kappa_n^i \, \mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i)$$
$$\int \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{\nu}_n^i + V_n^i (\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}) \, h(\xi) \, d\xi$$
$$= \varepsilon(1-\varepsilon)^{n-1} \kappa_n \kappa_n^i \left[\mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0) + O_p(\delta^{n-i}) \right]$$
$$\int \left[\mathrm{N}(\underline{z}_{i-1} - \xi; \underline{\nu}_n^i, W_n^i) + O_p(\delta^{n-i}) \right] h(\xi) \, d\xi \tag{4.263}$$
$$= \varepsilon(1-\varepsilon)^{n-1} \kappa_n \kappa_n^i \left[\mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0) \right.$$
$$\left. \int \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{\nu}_n^i, W_n^i) \, h(\xi) \, d\xi \; + \; O_p(\delta^{n-i}) \right], \tag{4.264}$$

where (4.263) follows from (4.253) and (4.262), and (4.264) from the fact that h has bounded moments, by hypothesis. Moreover, it is clear from (4.168) and (4.174) that

$$\kappa_n^0 = p(\underline{z}_0, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{n-1} = 0) \tag{4.265}$$

$$= p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)$$
$$p(\underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{n-1} = 0). \tag{4.266}$$

But

$$p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)$$

$$= \mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i,\, W_n^i + R \,), \tag{4.267}$$

as a consequence of (4.2) and the fact that the probability is conditioned on the event "there were no outliers among the first n observations." (Note that $\underline{\nu}_n^i$ is the optimal estimator of $\underline{\theta}_{i-1}$ given the observations $\{\, \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1} \,\}$, and W_n^i is its covariance.) Thus,

$$\kappa_n^0 = \mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i,\, W_n^i + R \,)$$

$$p(\, \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1} \mid \eta_0 = 0, \cdots, \eta_{n-1} = 0 \,) \tag{4.268}$$

$$= \kappa_n^i\, \mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i,\, W_n^i + R \,), \tag{4.269}$$

where (4.268) follows from (4.266) and (4.267), and (4.269) from (4.168) and (4.174). Hence, substituting into (4.264) yields

$$\varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n\, \kappa_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^i,\, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{\nu}_n^i + V_n^i(\, \underline{\theta}_n - \underline{\theta}_n^i \,),\, W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}$$

$$= \varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n\, \kappa_n^0 \left[\rho_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0 \,) \; + \; O_p(\, \delta^{n-i} \,) \right], \tag{4.270}$$

where

$$\rho_n^i \;:=\; \frac{\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{\nu}_n^i,\, W_n^i \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}}{\mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i,\, W_n^i + R \,)} \tag{4.271}$$

is the likelihood ratio for the dual alternatives of whether or not $\underline{v}_{i-1}$ was an outlier.

For $n \geq \omega$, rewrite (4.219) as

$$p(\, \underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1} \,)$$

$$= (\, 1 - \varepsilon \,)^n\, \kappa_n\, \kappa_n^0\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0 \,)$$

$$+ \; \varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n \sum_{i=1}^{n-\omega} \kappa_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^i,\, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{\nu}_n^i + V_n^i(\, \underline{\theta}_n - \underline{\theta}_n^i \,),\, W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}$$

$$+ \; \varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^i,\, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{\nu}_n^i + V_n^i(\, \underline{\theta}_n - \underline{\theta}_n^i \,),\, W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}$$

$$+ \; O_p(\, \varepsilon^2 (1 - \varepsilon)^{n-2} \,) \tag{4.272}$$

$$= (\, 1 - \varepsilon \,)^n\, \kappa_n\, \kappa_n^0\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0 \,)$$

$$+ \; \varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n\, \kappa_n^0 \sum_{i=1}^{n-\omega} \left[\rho_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^0,\, M_n^0 \,) \; + \; O_p(\, \delta^{n-i} \,) \right]$$

$$+ \; \varepsilon(\, 1 - \varepsilon \,)^{n-1}\, \kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\, \mathrm{N}(\, \underline{\theta}_n;\, \underline{\theta}_n^i,\, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i + V_n^i\,(\,\underline{\theta}_n - \underline{\theta}_n^i\,),\ W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\ h(\,\underline{\xi}\,)\, d\underline{\xi}$$

$$+\ O_p(\, \varepsilon^2 (1-\varepsilon)^{n-2}\,) \tag{4.273}$$

$$= \left[(\,1-\varepsilon\,)^n \ +\ \varepsilon(\,1-\varepsilon\,)^{n-1} \sum_{i=1}^{n-\omega} \rho_n^i \right] \kappa_n\ \kappa_n^0\ \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^0, M_n^0\,)$$

$$+\ \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0 \sum_{i=1}^{n-\omega} O_p(\,\delta^{n-i}\,)$$

$$+\ \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\ \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^i, M_n^i\,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i + V_n^i\,(\,\underline{\theta}_n - \underline{\theta}_n^i\,),\ W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\ h(\,\underline{\xi}\,)\, d\underline{\xi}$$

$$+\ O_p(\, \varepsilon^2 (1-\varepsilon)^{n-2}\,), \tag{4.274}$$

where (4.273) follows from (4.270). But since

$$\varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0 \sum_{i=1}^{n-\omega} O_p(\,\delta^{n-i}\,)$$

$$= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0 \sum_{i=\omega}^{n-1} O_p(\,\delta^{i}\,) \tag{4.275}$$

$$= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0 \left[\sum_{i=0}^{n-1} O_p(\,\delta^{i}\,)\ -\ \sum_{i=0}^{\omega-1} O_p(\,\delta^{i}\,) \right] \tag{4.276}$$

$$= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0\ O_p \left[\frac{1-\delta^n}{1-\delta}\ -\ \frac{1-\delta^\omega}{1-\delta} \right] \tag{4.277}$$

$$= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\ \kappa_n^0\ O_p \left[\frac{\delta^\omega - \delta^n}{1-\delta} \right] \tag{4.278}$$

$$= O_p(\, \varepsilon^2 (\,1-\varepsilon\,)^{n-1}\,) \tag{4.279}$$

or less, from (4.158), and the fact that $\{\kappa_n\}$ and $\{\kappa_n^0\}$ are bounded above. It follows that (4.274) can be rewritten as

$$p(\,\underline{\theta}_n \mid \underline{z}_0,\ \cdots,\ \underline{z}_{n-1}\,)$$

$$= \left[(\,1-\varepsilon\,)^n \ +\ \varepsilon(\,1-\varepsilon\,)^{n-1} \sum_{i=1}^{n-\omega} \rho_n^i \right] \kappa_n\ \kappa_n^0\ \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^0, M_n^0\,)$$

$$+\ \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\ \mathrm{N}(\,\underline{\theta}_n;\, \underline{\theta}_n^i, M_n^i\,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i + V_n^i\,(\,\underline{\theta}_n - \underline{\theta}_n^i\,),\ W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}\,)\ h(\,\underline{\xi}\,)\, d\underline{\xi}$$

$$+ \; O_p(\, \varepsilon^2 (1-\varepsilon)^{n-1} \,). \tag{4.280}$$

Let μ_n^i be the measure induced on $\mathbf{R}^p$ by the system (4.1)-(4.2), conditioned on the observations $\{\, \underline{z}_0, \; \cdots, \underline{z}_{i-2}, \underline{z}_i, \; \cdots, \underline{z}_{n-1} \,\}$, and let $\bar{\mu}_n^i$ be the same measure further conditioned on the events $\{\, \eta_0 = 0, \; \cdots, \eta_{i-1} = 0 \,\}$. From Theorem 4.1 and its corollary,

$$\bar{\mu}_n^i \; = \; \mu_n^i \; + \; O_p(\, \delta^{n-i+1}). \tag{4.281}$$

It follows that

$$E_{\mu_n^i} [\, \rho_n^i \,] \; = \; E_{\bar{\mu}_n^i} [\, \rho_n^i \,] \; + \; O_p(\, \delta^{n-i+1}) \tag{4.282}$$

$$= \int \frac{\int \mathrm{N}(\, \underline{z}_{i-1} - \xi ;\, \underline{\nu}_n^i, W_n^i \,)\, h(\xi)\, d\xi}{\mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i, W_n^i + R \,)} \, d\bar{\mu}_n^i(\underline{z}_{i-1}) \; + \; O_p(\, \delta^{n-i+1}) \tag{4.283}$$

$$= \int \frac{\int \mathrm{N}(\, \underline{z}_{i-1} - \xi ;\, \underline{\nu}_n^i, W_n^i \,)\, h(\xi)\, d\xi}{\mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i, W_n^i + R \,)} \, \mathrm{N}(\, \underline{z}_{i-1};\, \underline{\nu}_n^i, W_n^i + R \,)\, d\, \underline{z}_{i-1}$$

$$+ \; O_p(\, \delta^{n-i+1}) \tag{4.284}$$

$$= \int \int \mathrm{N}(\, \underline{z}_{i-1} - \xi ;\, \underline{\nu}_n^i, W_n^i \,)\, h(\xi)\, d\xi\, d\, \underline{z}_{i-1} \; + \; O_p(\, \delta^{n-i+1}) \tag{4.285}$$

$$= 1 \; + \; O_p(\, \delta^{n-i+1}) \tag{4.286}$$

w.p.1. Thus, there is a $0 < \rho < 1$ such that

$$\frac{1}{n-\omega} \sum_{i=1}^{n-\omega} \rho_n^i \; = \; \left[1 \; + \; \frac{1}{n-\omega} \sum_{i=1}^{n-\omega} O_p(\, \delta^{n-i+1}) \right] \; + \; O_p(\, \rho^{n-\omega}), \tag{4.287}$$

by virtue of the Chernoff bound (see for instance Chernoff, 1952; 1972, pp.44-45). It follows therefore that

$$\sum_{i=1}^{n-\omega} \rho_n^i \; = \; n \; - \; \omega \; + \; \sum_{i=1}^{n-\omega} O_p(\, \delta^{n-i+1}) \; + \; O_p((\, n-\omega\,)\, \rho^{n-\omega}) \tag{4.288}$$

$$= \; n \; - \; \omega \; + \; O_p(\, \varepsilon\,) \; + \; O_p((\, n-\omega\,)\, \rho^{n-\omega}), \tag{4.289}$$

where (4.288) follows from (4.287), and (4.289) from (4.275)-(4.279). The $O_p((\, n-\omega\,)\, \rho^{n-\omega})$ clearly vanishes as $n \to \infty$. Substituting into the first term on the right-hand side of (4.280) yields

$$\left[(\, 1-\varepsilon\,)^n \; + \; \varepsilon(\, 1-\varepsilon\,)^{n-1} \sum_{i=1}^{n-\omega} \rho_n^i \right] \kappa_n \; \kappa_n^0 \, \mathrm{N}(\, \underline{\theta}_n ;\, \underline{\theta}_n^0, M_n^0 \,)$$

$$= \left[1 \; - \; n\, \varepsilon \; + \; \sum_{k=2}^{n} \begin{bmatrix} n \\ k \end{bmatrix} \varepsilon^k \; + \; \varepsilon(n-\omega) \; + \; (n-\omega) \sum_{k=1}^{n-1} \begin{bmatrix} n \\ k \end{bmatrix} \varepsilon^{k+1} \right]$$

$$\kappa_n \; \kappa_n^0 \, \mathrm{N}(\, \underline{\theta}_n ;\, \underline{\theta}_n^0, M_n^0 \,) \; + \; O_p(\, \varepsilon^2) \; + \; O_p((\, n-\omega\,)\, \rho^{n-\omega}) \tag{4.290}$$

$$= \left[1 - \omega\varepsilon + \sum_{k=2}^{n} \begin{bmatrix} n \\ k \end{bmatrix} \varepsilon^{k} + (n-\omega) \sum_{k=1}^{n-1} \begin{bmatrix} n \\ k \end{bmatrix} \varepsilon^{k+1} \right]$$

$$\kappa_n \, \kappa_n^0 \, \mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0) + O_p(\varepsilon^2) + O_p((n-\omega)\rho^{n-\omega}), \qquad (4.291)$$

from (4.289). It is worth noting that $1-\omega\varepsilon$ in (4.291) corresponds to the first two terms of the expansion of $(1-\varepsilon)^{\omega}$.

More generally, suppose a finite number m of outliers occurred during the first n time steps. The prior probability of such an event is $\varepsilon^m (1-\varepsilon)^{n-m}$. All the outliers may have occurred during the most recent ω time steps, resulting in

$$\begin{bmatrix} \omega \\ m \end{bmatrix} = \frac{\omega !}{m!\,(\omega - m)!} \qquad (4.292)$$

terms in the corresponding sum, which is consequently bounded. Alternatively, $m-1$ outliers may have occurred during the most recent ω time steps, and one during the earlier $n-\omega$ time steps. In the latter case, the effects of that early outlier will have attentuated to $O(\varepsilon)$, by (4.158), and the corresponding term will therefore be indistinguishable, to $O(\varepsilon^2)$, from the case where only $m-1$ outliers occurred. Clearly, there are

$$\begin{bmatrix} n-\omega \\ 1 \end{bmatrix} = n - \omega \qquad (4.293)$$

such terms. Analogous arguments can be made for $m-2, \cdots, 0$ outliers occurring during the last ω time steps.

Obviously, 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 $m-1$ outliers occurred, neither of which during the most recent ω time steps, and so on. In general, therefore, the "no outliers" term has the coefficient

$$(1-\varepsilon)^n + \varepsilon(1-\varepsilon)^{n-1} \begin{bmatrix} n-\omega \\ 1 \end{bmatrix} + \varepsilon^2 (1-\varepsilon)^{n-2} \begin{bmatrix} n-\omega \\ 2 \end{bmatrix} + \cdots$$

$$= \sum_{m=0}^{n-\omega} \varepsilon^m (1-\varepsilon)^{n-m} \begin{bmatrix} n-\omega \\ m \end{bmatrix} \qquad (4.294)$$

$$= (1-\varepsilon)^{\omega} \sum_{m=0}^{n-\omega} \varepsilon^m (1-\varepsilon)^{n-\omega-m} \begin{bmatrix} n-\omega \\ m \end{bmatrix} \qquad (4.295)$$

$$= (1-\varepsilon)^{\omega} \; (\varepsilon + 1 - \varepsilon)^{n-\omega} \qquad (4.296)$$

$$= (1-\varepsilon)^{\omega}, \qquad (4.297)$$

which agrees with (4.160). Similarly, the "one outlier" term corresponds to the coefficient

$$\varepsilon(1-\varepsilon)^{n-1} + \varepsilon^2(1-\varepsilon)^{n-2}\left[\begin{matrix} n-\omega \\ 1 \end{matrix}\right] + \varepsilon^3(1-\varepsilon)^{n-3}\left[\begin{matrix} n-\omega \\ 2 \end{matrix}\right] + \cdots$$

$$= \sum_{m=0}^{n-\omega} \varepsilon^{m+1}(1-\varepsilon)^{n-1-m}\left[\begin{matrix} n-\omega \\ m \end{matrix}\right] \tag{4.298}$$

$$= \varepsilon(1-\varepsilon)^{\omega-1} \sum_{m=0}^{n-\omega} \varepsilon^{m}(1-\varepsilon)^{n-\omega-m}\left[\begin{matrix} n-\omega \\ m \end{matrix}\right] \tag{4.299}$$

$$= \varepsilon(1-\varepsilon)^{\omega-1}\ (\varepsilon+1-\varepsilon)^{n-\omega} \tag{4.300}$$

$$= \varepsilon(1-\varepsilon)^{\omega-1}, \tag{4.301}$$

which also agrees with (4.160). Similar arguments may be made for higher numbers of outliers. It follows from (4.292) that the order of each term is

$$\varepsilon^m(1-\varepsilon)^{\omega-m}\ \frac{\omega(\omega-1)\cdots(\omega-m+1)}{m\,!} = O(\varepsilon^m\,\omega^m). \tag{4.302}$$

From (4.159), the most significant term is for the smallest possible m, i.e. for $m = 2$, concluding the proof. ■

Remark The analogue of equation (4.160) for the case $n < \omega$ is equation (4.219).

The following corollary is immediate.

Corollary 4.2 Let the conditions of Theorem 4.1 and Corollary 4.1 be satisfied for the system (4.1) and (4.9), and let δ be a real number for which (4.14) holds. Let ω be the smallest integer such that (4.158) is satisfied. If (4.159) holds and if the distribution H has bounded moments, then

$$\begin{aligned} p(\,\underline{\theta}_n \mid \underline{z}_0,\ \cdots,\underline{z}_{n-1}\,) \\ &= (\,1-\varepsilon\,)^{\omega}\,\kappa_n\ \kappa_n^0\,\mathrm{N}(\,\underline{\theta}_n\,;\,\underline{\theta}_n^0,\,M_n^0\,) \\ &\quad + \varepsilon(\,1-\varepsilon\,)^{\omega-1}\,\kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i\,\mathrm{N}(\,\underline{\theta}_n\,;\,\underline{\theta}_n^i,\,M_n^i\,) \\ &\qquad \int \mathrm{N}(\,\underline{z}_{i-1}-\xi;\ H_{i-1}\underline{v}_n^i + H_{i-1}V_n^i(\underline{\theta}_n-\underline{\theta}_n^i), \\ &\qquad\qquad H_{i-1}W_n^i H_{i-1}^{\mathrm{T}} - H_{i-1}V_n^i M_n^i V_n^{i\,\mathrm{T}} H_{i-1}^{\mathrm{T}}\,)\,h(\xi)\,d\xi \\ &\quad + O_p(\omega^2\varepsilon^2) \end{aligned} \tag{4.303}$$

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

$$\underline{\theta}_n^i = F_{n-1}\underline{\theta}_{n-1}^i + F_{n-1}M_{n-1}^i H_{n-1}^{\mathrm{T}}\Gamma_{n-1}^{i\;-1}(\underline{z}_{n-1} - H_{n-1}\underline{\theta}_{n-1}^i) \tag{4.304}$$

$$M_n^i = F_{n-1}P_{n-1}^i F_{n-1}^{\mathrm{T}} + Q_{n-1} \tag{4.305}$$

$$P_n^i = M_n^i - M_n^i H_n^{\mathrm{T}}\Gamma_n^{i\;-1}H_n M_n^i \tag{4.306}$$

$$\Gamma_n^i = H_n M_n^i H_n^{\mathrm{T}} + D_n R D_n^{\mathrm{T}} \tag{4.307}$$

$$V_n^i = V_{n-1}^i P_{n-1}^i F_{n-1}^{\mathrm{T}} M_n^{i\;-1} \tag{4.308}$$

$$\underline{v}_n^i = \underline{v}_{n-1}^i + V_{n-1}^i M_{n-1}^i H_{n-1}^{\mathrm{T}} \Gamma_{n-1}^{i\;-1}(\underline{z}_{n-1} - H_{n-1}\underline{\theta}_{n-1}^i) \tag{4.309}$$

$$W_n^i = W_{n-1}^i - V_{n-1}^i M_{n-1}^i H_{n-1}^{\mathrm{T}} \Gamma_{n-1}^{i\;-1} H_{n-1} M_{n-1}^i V_{n-1}^{i\;\mathrm{T}} \tag{4.310}$$

$$\kappa_n^i = \kappa_{n-1}^i \, \mathrm{N}(\underline{z}_{n-1};\, H_{n-1}\underline{\theta}_{n-1}^i,\, \Gamma_{n-1}^i) \tag{4.311}$$

subject to the initial conditions (4.169)-(4.177). The normalization constant satisfies

$$\begin{aligned}\kappa_n^{-1} &= (1-\varepsilon)^{\omega}\,\kappa_n^0 \\ &\quad + \varepsilon(1-\varepsilon)^{\omega-1}\sum_{i=n-\omega+1}^{n}\kappa_n^i \int \mathrm{N}(\underline{z}_{i-1}-\underline{\xi};\, H_{i-1}\underline{v}_n^i, \\ &\qquad H_{i-1}W_n^i H_{i-1}^{\mathrm{T}})\, h(\underline{\xi})\, d\underline{\xi}.\end{aligned} \tag{4.312}$$

Proof The proof is identical to that of Theorem 4.2, and is omitted. ■

4.2 A First-Order Approximation to the Conditional Mean Estimator

The approximate conditional prior probability distribution of the state $\underline{\theta}_n$ given the observations $\{ \underline{z}_0, \cdots, \underline{z}_{n-1} \}$ is now used in an extension of a theorem due to Masreliez. This results in a first-order approximation to the conditional mean (i.e. minimum-variance) estimator.

The following notation is used, respectively for the conditional mean and conditional variance of $\underline{\theta}_n$:

$$\underline{T}_n := E\,[\,\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_n\,] \tag{4.313}$$

$$\Sigma_n := E\,[\,(\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} \mid \underline{z}_0, \cdots, \underline{z}_n\,] \tag{4.314}$$

In addition, the functional

$$\underline{\psi}_n^0(\underline{z}_n) := -\frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)} \underline{\nabla}_{\underline{z}_n} p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0) \tag{4.315}$$

denotes the score function for the conditional probability of $\underline{z}_n$ -- i.e. the additive inverse of the gradient of its logarithm -- as defined earlier in equations (3.226)-(3.227), and similarly, for $i = 1, 2, \cdots$ and

$n \geq i$,

$$\underline{\psi}_n^i(\underline{z}_{i-1}) := -\frac{1}{p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0)}$$

$$\underline{\nabla}_{\underline{z}_{i-1}} p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n,$$

$$\eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0) \qquad (4.316)$$

Finally, for $i = 0, 1, 2, \cdots$ and $n \geq i$,

$$\Psi_n^i(\underline{z}) := \underline{\nabla}_{\underline{z}}\, \underline{\psi}_n^{i\,\mathrm{T}}(\underline{z}) \qquad (4.317)$$

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

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

$$\underline{T}_n = (1-\varepsilon)^{\omega} \kappa_{n+1} \pi_n^0 \underline{T}_n^0 + \varepsilon (1-\varepsilon)^{\omega-1} \kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \underline{T}_n^i + O_p(\omega^2 \varepsilon^2) \qquad (4.318)$$

for all $n \geq \omega$, where

$$\underline{T}_n^0 = \underline{\theta}_n^0 + M_n^0 \underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0) \qquad (4.319)$$

$$\underline{T}_n^i = \underline{\theta}_n^i + M_n^i \Gamma_n^{i\,-1}(\underline{z}_n - \underline{\theta}_n^i) + P_n^i V_n^{i\,\mathrm{T}} \underline{\psi}_n^i(\underline{z}_{i-1} - \underline{v}_{n+1}^i) \qquad (4.320)$$

$$\pi_n^0 = (1-\varepsilon) \kappa_{n+1}^0 + \varepsilon\, \kappa_n^0 \int \mathrm{N}(\underline{z}_n - \xi; \underline{\theta}_n^0, M_n^0)\, h(\xi)\, d\xi \qquad (4.321)$$

$$\pi_n^i = (1-\varepsilon) \kappa_{n+1}^i \int \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{v}_{n+1}^i, W_{n+1}^i)\, h(\xi)\, d\xi \qquad (4.322)$$

$$\underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0) = -\frac{\underline{\nabla}_{\underline{z}_n} \left[(1-\varepsilon) \mathrm{N}(\underline{z}_n; \underline{\theta}_n^0, \Gamma_n^0) + \varepsilon \int \mathrm{N}(\underline{z}_n - \xi; \underline{\theta}_n^0, M_n^0)\, h(\xi)\, d\xi \right]}{(1-\varepsilon) \mathrm{N}(\underline{z}_n; \underline{\theta}_n^0, \Gamma_n^0) + \varepsilon \int \mathrm{N}(\underline{z}_n - \xi; \underline{\theta}_n^0, M_n^0)\, h(\xi)\, d\xi} \qquad (4.323)$$

$$\underline{\psi}_n^i(\underline{z}_{i-1} - \underline{v}_{n+1}^i) = -\frac{\underline{\nabla}_{\underline{z}_{i-1}} \int \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{v}_{n+1}^i, W_{n+1}^i)\, h(\xi)\, d\xi}{\int \mathrm{N}(\underline{z}_{i-1} - \xi; \underline{v}_{n+1}^i, W_{n+1}^i)\, h(\xi)\, d\xi} \qquad (4.324)$$

with $\underline{\theta}_n^i$, M_n^i, P_n^i, Γ_n^i, V_n^i, $\underline{v}_n^i$, W_n^i, κ_n^i, and κ_n as defined in equations (4.161)-(4.168), subject to the initial conditions (4.169)-(4.177). Furthermore,

$$\Sigma_n = (1-\varepsilon)^{\omega} \kappa_{n+1} \pi_n^0 \Sigma_n^0 + \varepsilon (1-\varepsilon)^{\omega-1} \kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \Sigma_n^i + O_p(\omega^2 \varepsilon^2) \qquad (4.325)$$

for all $n \geq \omega$, where

$$\Sigma_n^0 = M_n^0 - M_n^0 \Psi_n^0(\underline{z}_n - \underline{\theta}_n^0) M_n^0 + (\underline{T}_n - \underline{T}_n^0)(\underline{T}_n - \underline{T}_n^0)^{\mathrm{T}} \qquad (4.326)$$

$$\Sigma_n^i = P_n^i - P_n^i V_n^{i\,\mathrm{T}} \Psi_n^i(\underline{z}_{i-1} - \underline{v}_{n+1}^i) V_n^i P_n^i + (\underline{T}_n - \underline{T}_n^i)(\underline{T}_n - \underline{T}_n^i)^{\mathrm{T}}, \qquad (4.327)$$

and Ψ_n^i is given by equation (4.317), with (4.323) and (4.324).

Proof Note first that

$$p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_n) = \frac{p(\underline{\theta}_n, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \tag{4.328}$$

$$= \frac{p(\underline{z}_n \mid \underline{\theta}_n, \underline{z}_0, \cdots, \underline{z}_{n-1})\, p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \tag{4.329}$$

$$= \frac{p(\underline{z}_n \mid \underline{\theta}_n)\, p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}, \tag{4.330}$$

where (4.328) and (4.329) follow from the definition of the conditional probability, and (4.330) from (4.2) and the fact that $\{\underline{v}_n\}$ are independent.

It therefore follows that

$$\underline{T}_n = \int \underline{\theta}_n\, p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_n)\, d\underline{\theta}_n \tag{4.331}$$

$$= \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int \underline{\theta}_n\, p(\underline{z}_n \mid \underline{\theta}_n)\, p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})\, d\underline{\theta}_n \tag{4.332}$$

$$\begin{aligned} = \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int \underline{\theta}_n\, f(\underline{z}_n - \underline{\theta}_n) \\ \Big[(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0) \\ + \varepsilon(1-\varepsilon)^{n-1}\, \kappa_n \sum_{i=1}^{n} \kappa_n^i\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^i, M_n^i) \\ \int \mathrm{N}(\underline{z}_{i-1} - \underline{\xi}; \underline{v}_n^i + V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}})\, h(\underline{\xi})\, d\underline{\xi} \\ + O_p(\varepsilon^2 (1-\varepsilon)^{n-2}) \Big]\, d\underline{\theta}_n, \end{aligned} \tag{4.333}$$

where (4.331) follows from (4.313), (4.332) from (4.330), and (4.333) from (4.219), (4.2), and the definition of f.

Consider the first term on the right-hand side of (4.333), i.e. the "no outliers among the first n observations" term, and rewrite as

$$\begin{aligned} \frac{(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int \underline{\theta}_n\, f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ = \frac{(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \\ \Big[M_n^0 \int M_n^{0\,-1} (\underline{\theta}_n - \underline{\theta}_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ + \underline{\theta}_n^0 \int f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \Big]. \end{aligned} \tag{4.334}$$

Now, by independence,

$$f(\underline{z}_n - \underline{\theta}_n) = p(\underline{z}_n \mid \underline{\theta}_n) \tag{4.335}$$

$$= p(\underline{z}_n \mid \underline{\theta}_n, \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0), \tag{4.336}$$

and moreover,

$$\mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0) = p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0). \tag{4.337}$$

It follows that

$$\int f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= \int p(\underline{z}_n \mid \underline{\theta}_n, \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0)$$

$$p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0)\, d\underline{\theta}_n \tag{4.338}$$

$$= \int p(\underline{\theta}_n, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0)\, d\underline{\theta}_n \tag{4.339}$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0), \tag{4.340}$$

from the definition of the marginal probability. Thus, using (4.184),

$$\frac{(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}\, \underline{\theta}_n^0 \int f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= \frac{p(\eta_0=0, \cdots, \eta_{n-1}=0 \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}\, \underline{\theta}_n^0$$

$$p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0) \tag{4.341}$$

$$= p(\eta_0=0, \cdots, \eta_{n-1}=0 \mid \underline{z}_0, \cdots, \underline{z}_n)\, \underline{\theta}_n^0, \tag{4.342}$$

from the definition of the conditional probability.

Note next that

$$M_n^{0\ -1}(\underline{\theta}_n - \underline{\theta}_n^0)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0) = -\underline{\nabla}_{\underline{\theta}_n} \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0), \tag{4.343}$$

so that

$$\int \underline{\nabla}_{\underline{\theta}_n} \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n = -\int M_n^{0\ -1}(\underline{\theta}_n - \underline{\theta}_n^0)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \tag{4.344}$$

$$= -M_n^{0\ -1}(\underline{\theta}_n^0 - \underline{\theta}_n^0) \tag{4.345}$$

$$= 0. \tag{4.346}$$

Hence,

$$\int M_n^{0\ -1}(\underline{\theta}_n - \underline{\theta}_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= -\int f(\underline{z}_n - \underline{\theta}_n)\, \underline{\nabla}_{\underline{\theta}_n} \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \tag{4.347}$$

$$= - f(\underline{z}_n - \underline{\theta}_n) \int \underline{\nabla}_{\underline{\theta}_n} \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$+ \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, \underline{\nabla}_{\underline{\theta}_n} f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \tag{4.348}$$

$$= \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, \underline{\nabla}_{\underline{\theta}_n} f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \tag{4.349}$$

$$= - \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, \underline{\nabla}_{\underline{z}_n} f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \tag{4.350}$$

$$= - \underline{\nabla}_{\underline{z}_n} \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \tag{4.351}$$

$$= - \underline{\nabla}_{\underline{z}_n} p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0), \tag{4.352}$$

where (4.347) follows from (4.343), (4.348) from integration by parts, (4.349) from (4.346), (4.351) is justified by the dominated convergence theorem (since both the normal density with $R > 0$ and h are bounded and differentiable a.e., and hence, so is f), and (4.352) follows from (4.340). Thus,

$$\frac{(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}\, M_n^0 \int M_n^{0\ -1} (\underline{\theta}_n - \underline{\theta}_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= \frac{p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}$$

$$\frac{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)}\, M_n^0$$

$$\left[- \underline{\nabla}_{\underline{z}_n} p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0) \right] \tag{4.353}$$

$$= p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_n)\, M_n^0$$

$$\left[- \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)} \right.$$

$$\left. \underline{\nabla}_{\underline{z}_n} p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0) \right] \tag{4.354}$$

$$= p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_n)\, M_n^0\, \underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0), \tag{4.355}$$

where (4.353) follows from (4.184) and (4.352), (4.354) from the definition of the conditional probability, and (4.355) from (4.315). Substituting (4.342) and (4.355) into (4.334) yields

$$\frac{(1-\varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int \underline{\theta}_n\, f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_n) \left[\underline{\theta}_n^0 + M_n^0\, \underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0) \right] \tag{4.356}$$

$$= p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_n)\, \underline{T}_n^0, \tag{4.357}$$

from (4.319).

But from the definition of the conditional probability,

$$p(\, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \,\cdots, \underline{z}_n \,)\, p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1} \,)$$

$$= p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \,)$$

$$p(\, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1} \,) \tag{4.358}$$

$$= p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \,)\ (1 - \varepsilon)^n\, \kappa_n\, \kappa_n^0 \tag{4.359}$$

from (4.184). Since $\{\underline{v}_n\}$ are independent,

$$p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \,)$$

$$= p(\, \underline{\theta}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}, \eta_0 = 0, \,\cdots, \eta_{n-1} = 0 \,) \ * \ p(\, \underline{v}_n \,) \tag{4.360}$$

$$= \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^0, M_n^0 \,) \ * \ \Big[(1 - \varepsilon)\, \mathrm{N}(\, \underline{v}_n, 0, R \,) \ + \ \varepsilon\, h(\, \underline{v}_n \,) \Big] \tag{4.361}$$

$$= (\, 1 - \varepsilon\,)\, \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n^0, \Gamma_n^0 \,) \ + \ \varepsilon \int \mathrm{N}(\, \underline{z}_n - \underline{\xi} ; \underline{\theta}_n^0, M_n^0 \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}, \tag{4.362}$$

where (4.361) follows from (4.157), and the first term on the right-hand side of (4.362) from the fact that the convolution of two normal distributions is also normal, with appropriate mean and variance. Comparing (4.362) with (4.315) establishes (4.323). Substituting into (4.359), and using (4.168) and (4.252), establishes that (4.334) can be rewritten as

$$\frac{(\, 1 - \varepsilon\,)^n\, \kappa_n\, \kappa_n^0}{p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1} \,)} \int \underline{\theta}_n\, f(\, \underline{z}_n - \underline{\theta}_n \,)\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^0, M_n^0 \,)\, d\underline{\theta}_n$$

$$= (\, 1 - \varepsilon\,)^n\, \kappa_{n+1}\, \pi_n^0\, \underline{T}_n^0, \tag{4.363}$$

from (4.357) and (4.321).

Consider now each term in the summation in (4.333). Although these terms are not normal, they involve convolutions of normal distributions. For this reason, manipulations similar to those above (equations (4.334) and (4.343)-(4.352)) are still possible. Each term in the summation in (4.333) may be rewritten as

$$\frac{\varepsilon(\, 1 - \varepsilon\,)^{n-1}\, \kappa_n\, \kappa_n^i}{p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1} \,)} \int \underline{\theta}_n\, f(\, \underline{z}_n - \underline{\theta}_n \,)\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\, (\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i\, M_n^i\, V_n^{i\,\mathrm{T}} \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}\ \ d\underline{\theta}_n$$

$$= \frac{\varepsilon(\, 1 - \varepsilon\,)^{n-1}\, \kappa_n\, \kappa_n^i}{p(\, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1} \,)}$$

$$\int \underline{\theta}_n \Big[(1 - \varepsilon)\, \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n, R \,) + O(\varepsilon) \Big]\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi} ; \underline{v}_n^i + V_n^i\, (\underline{\theta}_n - \underline{\theta}_n^i\,), W_n^i - V_n^i\, M_n^i\, V_n^{i\,\mathrm{T}} \,)\, h(\, \underline{\xi} \,)\, d\underline{\xi}\ \ d\underline{\theta}_n, \tag{4.364}$$

from (4.157). For economy of notation, define for given n and i the function

$$g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,) \ := \int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i + V_n^i (\underline{\theta}_n - \underline{\theta}_n^i),\ W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\, h(\underline{\xi})\, d\underline{\xi}. \qquad (4.365)$$

Then,

$$\begin{aligned}
&\int \underline{\theta}_n\ \mathrm{N}(\, \underline{z}_n; \underline{\theta}_n, R \,)\, \mathrm{N}(\, \underline{\theta}_n; \underline{\theta}_n^i, M_n^i \,) \\
&\qquad\qquad \int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\ \underline{v}_n^i + V_n^i (\underline{\theta}_n - \underline{\theta}_n^i),\ W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}} \,)\, h(\underline{\xi})\, d\underline{\xi}\ \ d\underline{\theta}_n \\
&\quad = \int \underline{\theta}_n\ \mathrm{N}(\, \underline{\theta}_n; \underline{\theta}_n^i + M_n^i \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,), P_n^i \,)\, g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,)\, d\underline{\theta}_n \\
&\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{N}(\, \underline{z}_n; \underline{\theta}_n^i, \Gamma_n^i \,) \qquad (4.366) \\
&\quad = \Big[\ P_n^i \int P_n^{i\ -1} \Big[\underline{\theta}_n - \underline{\theta}_n^i - M_n^i \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,) \Big] \\
&\qquad\qquad \mathrm{N}(\, \underline{\theta}_n; \underline{\theta}_n^i + M_n^i \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,), P_n^i \,)\, g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,)\, d\underline{\theta}_n \\
&\qquad + \Big[\underline{\theta}_n^i + M_n^i \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,) \Big] \\
&\qquad\qquad \int \mathrm{N}(\, \underline{\theta}_n; \underline{\theta}_n^i + M_n^i \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,), P_n^i \,) \\
&\qquad\qquad\qquad g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,)\, d\underline{\theta}_n\ \Big]\, \mathrm{N}(\, \underline{z}_n; \underline{\theta}_n^i, \Gamma_n^i \,), \qquad (4.367)
\end{aligned}$$

where (4.366) follows from Lemma 4.2, (4.163), and (4.365). Note that

$$\underline{\nabla}_{\underline{\theta}_n} g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,) \ = \ - \ V_n^{i\ \mathrm{T}}\, \underline{\nabla}_{\underline{z}_{i-1}} g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,). \qquad (4.368)$$

Moreover,

$$\begin{aligned}
&\mathrm{N}(\, \underline{z}_n; \underline{\theta}_n^i, \Gamma_n^i \,) \int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi} - \underline{v}_n^i,\ V_n^i M_n^i \Gamma_n^{i\ -1} (\underline{z}_n - \underline{\theta}_n^i), \\
&\qquad\qquad\qquad\qquad W_n^i - V_n^i M_n^i \Gamma_n^{i\ -1} M_n^i V_n^{i\,\mathrm{T}} \,)\ \ h(\underline{\xi})\, d\underline{\xi} \\
&\quad = p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0 \,) \\
&\qquad\quad p(\, \underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n, \\
&\qquad\qquad\qquad \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0 \,) \ + \ O(\varepsilon) \qquad (4.369) \\
&\quad = p(\, \underline{z}_{i-1}, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \\
&\qquad\qquad\qquad \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0 \,) \ + \ O(\varepsilon). \qquad (4.370)
\end{aligned}$$

Finally,

$$\begin{aligned}
&- \ \underline{\nabla}_{\underline{z}_{i-1}} \int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi} - \underline{v}_n^i,\ V_n^i M_n^i \Gamma_n^{i\ -1} (\underline{z}_n - \underline{\theta}_n^i), \\
&\qquad\qquad\qquad\qquad W_n^i - V_n^i M_n^i \Gamma_n^{i\ -1} M_n^i V_n^{i\,\mathrm{T}} \,)\ \ h(\underline{\xi})\, d\underline{\xi} \\
&\quad = - \ \frac{p(\, \underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0 \,)}{p(\, \underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_n, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0 \,)}
\end{aligned}$$

$$\begin{aligned}&\underline{\nabla}_{\underline{z}_{i-1}}\, p(\, \underline{z}_{i-1} \mid \underline{z}_0, \,\cdots, \underline{z}_{i-2}, \underline{z}_i\,, \,\cdots, \underline{z}_n\,, \\ &\qquad\qquad \eta_0=0, \,\cdots, \eta_{i-1}=1, \,\cdots, \eta_{n-1}=0\,) \end{aligned} \qquad (4.371)$$

$$\begin{aligned}&= p(\, \underline{z}_{i-1} \mid \underline{z}_0, \,\cdots, \underline{z}_{i-2}, \underline{z}_i\,, \,\cdots, \underline{z}_n\,, \\ &\qquad\qquad \eta_0=0, \,\cdots, \eta_{i-1}=1, \,\cdots, \eta_{n-1}=0\,)\;\; \underline{\psi}_n^i\,(\underline{z}_{i-1}-\underline{\nu}_{n+1}^i), \end{aligned} \qquad (4.372)$$

from (4.316) with (4.166). Comparing (4.367) with (4.334), and using (4.368), (4.370), and (4.372), establishes (following the reasoning of equations (4.335)-(4.357)) that (4.364) may be rewritten as

$$\begin{aligned}&\frac{\varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\,\kappa_n^i}{p(\,\underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}\,)} \int \underline{\theta}_n\, f(\,\underline{z}_n - \underline{\theta}_n\,)\,\mathrm{N}(\,\underline{\theta}_n\,;\, \underline{\theta}_n^i, M_n^i\,) \\ &\qquad\qquad \int \mathrm{N}(\,\underline{z}_{i-1}-\underline{\xi};\, \underline{\nu}_n^i + V_n^i\,(\underline{\theta}_n - \underline{\theta}_n^i\,),\, W_n^i - V_n^i\,M_n^i\,V_n^{i\,\mathrm{T}}\,)\, h(\,\underline{\xi}\,)\, d\underline{\xi}\;\; d\underline{\theta}_n \\ &= \frac{\varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_n\,\kappa_n^i}{p(\,\underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{n-1}\,)} \\ &\qquad\qquad \Big[\; (\,1-\varepsilon\,)\, p(\,\underline{z}_{i-1}, \underline{z}_n \mid \underline{z}_0, \,\cdots, \underline{z}_{i-2}, \underline{z}_i\,, \,\cdots, \underline{z}_{n-1}, \\ &\qquad\qquad\qquad\qquad \eta_0=0, \,\cdots, \eta_{i-1}=1, \,\cdots, \eta_{n-1}=0\,) \\ &\qquad\qquad \Big[\; \underline{\theta}_n^i \;+\; M_n^i\,\Gamma_n^{i\;\,-1}\,(\underline{z}_n - \underline{\theta}_n^i\,) \;+\; P_n^i\,V_n^{i\,\mathrm{T}}\,\underline{\psi}_n^i\,(\underline{z}_{i-1}-\underline{\nu}_{n+1}^i\,)\,\Big] \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + O(\varepsilon)\,\Big] \end{aligned} \qquad (4.373)$$

$$\begin{aligned}&= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_{n+1}\,\pi_n^i\,\Big[\underline{\theta}_n^i \;+\; M_n^i\,\Gamma_n^{i\;\,-1}\,(\underline{z}_n - \underline{\theta}_n^i\,) \\ &\qquad\qquad + \; P_n^i\,V_n^{i\,\mathrm{T}}\,\underline{\psi}_n^i\,(\underline{z}_{i-1}-\underline{\nu}_{n+1}^i\,)\,\Big] \;+\; O(\,\varepsilon^2(\,1-\varepsilon\,)^{n-1}\,) \end{aligned} \qquad (4.374)$$

$$= \varepsilon(\,1-\varepsilon\,)^{n-1}\,\kappa_{n+1}\,\pi_n^i\,T_n^i \;+\; O(\,\varepsilon^2(\,1-\varepsilon\,)^{n-1}\,), \qquad (4.375)$$

where (4.374) follows from (4.370), (4.322), (4.252), and (4.168), and (4.375) from (4.320). Combining (4.363) and (4.376) yields

$$\underline{T}_n \;=\; (\,1-\varepsilon\,)^n\;\kappa_{n+1}\,\pi_n^0\,\underline{T}_n^0 \;+\; \varepsilon\,(\,1-\varepsilon\,)^{n-1}\,\kappa_{n+1}\sum_{i=1}^{n}\pi_n^i\,\underline{T}_n^i \;+\; O(\,\varepsilon^2(\,1-\varepsilon\,)^{n-1}\,), \qquad (4.376)$$

and the equivalence of (4.376) and (4.318) is an immediate consequence of Theorem 4.2.

Moving on to the estimation error covariance,

$$\Sigma_n = \int (\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_n)\, d\underline{\theta}_n \tag{4.377}$$

$$= \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int (\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} p(\underline{z}_n \mid \underline{\theta}_n) \, p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})\, d\underline{\theta}_n \tag{4.378}$$

$$\begin{aligned} = \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int & (\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \\ & \Big[(1-\varepsilon)^n \kappa_n \kappa_n^0 \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0) \\ & + \varepsilon(1-\varepsilon)^{n-1} \kappa_n \sum_{i=1}^{n} \kappa_n^i \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^i, M_n^i) \\ & \quad \int \mathrm{N}(\underline{z}_{i-1} - \underline{\xi}; \underline{\nu}_n^i + V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}})\, h(\underline{\xi})\, d\underline{\xi} \\ & + O_p(\varepsilon^2(1-\varepsilon)^{n-2}) \Big]\, d\underline{\theta}_n , \end{aligned} \tag{4.379}$$

where (4.377) follows from (4.314), (4.378) from (4.330), and (4.379) from (4.219), (4.2), and the definition of f.

Consider the first term on the right-hand side of (4.379), and rewrite as

$$\begin{aligned} & \frac{(1-\varepsilon)^n \kappa_n \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \int (\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ & = \frac{(1-\varepsilon)^n \kappa_n \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} \\ & \qquad \int (\underline{\theta}_n - \underline{\theta}_n^0 + \underline{\theta}_n^0 - \underline{T}_n)(\underline{\theta}_n - \underline{\theta}_n^0 + \underline{\theta}_n^0 - \underline{T}_n)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \\ & \qquad \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \end{aligned} \tag{4.380}$$

$$\begin{aligned} = \frac{(1-\varepsilon)^n \kappa_n \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})} & \Big[\int (\underline{\theta}_n - \underline{\theta}_n^0)(\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ & + \int (\underline{\theta}_n^0 - \underline{T}_n)(\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ & + \int (\underline{\theta}_n - \underline{\theta}_n^0)(\underline{\theta}_n^0 - \underline{T}_n)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \\ & + \int (\underline{\theta}_n^0 - \underline{T}_n)(\underline{\theta}_n^0 - \underline{T}_n)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n \Big]. \end{aligned} \tag{4.381}$$

Neglecting for now the coefficient, the first term on the right-hand side of (4.381) may be rewritten as

$$M_n^0 \int M_n^{0\;-1} (\underline{\theta}_n - \underline{\theta}_n^0)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n$$

$$= -\, M_n^0 \int \underline{\nabla}_{\underline{\theta}_n} \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \qquad (4.382)$$

$$= M_n^0 \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, \underline{\nabla}_{\underline{\theta}_n} \Big[(\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n) \Big]\, d\underline{\theta}_n \qquad (4.383)$$

$$= M_n^0 \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0) \Big[f(\underline{z}_n - \underline{\theta}_n)\, I$$

$$+ \, [\, \underline{\nabla}_{\underline{\theta}_n} f(\underline{z}_n - \underline{\theta}_n)\,]\, (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} \Big]\, d\underline{\theta}_n \qquad (4.384)$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0$$

$$+ \, M_n^0 \int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, [\, \underline{\nabla}_{\underline{\theta}_n} f(\underline{z}_n - \underline{\theta}_n)\,]\, (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}}\, d\underline{\theta}_n \qquad (4.385)$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0$$

$$- \, M_n^0\, \underline{\nabla}_{\underline{z}_n} \int (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}}\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \qquad (4.386)$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0$$

$$- \, M_n^0\, \underline{\nabla}_{\underline{z}_n} \Big[M_n^0 \int M_n^{0\;-1} (\underline{\theta}_n - \underline{\theta}_n^0)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)$$

$$f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \Big]^{\mathrm{T}} \qquad (4.387)$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0$$

$$+ \, M_n^0\, \underline{\nabla}_{\underline{z}_n}^2 \Big[\int \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, f(\underline{z}_n - \underline{\theta}_n)\, d\underline{\theta}_n \Big]\, M_n^0 \qquad (4.388)$$

$$= p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0$$

$$+ \, M_n^0\, \underline{\nabla}_{\underline{z}_n}^2\, p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)\, M_n^0, \qquad (4.389)$$

where (4.382) follows from (4.343), (4.383) is analogous to (4.349), (4.385) and (4.387) follow from (4.340), (4.386) is justified by the dominated convergence theorem, and (4.388) follows the same reasoning as (4.382)-(4.386), as well as the symmetry of the covariance matrix. Thus, the first term in (4.381) becomes

$$\frac{(1 - \varepsilon)^n\, \kappa_n\, \kappa_n^0}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})}$$

$$\int (\underline{\theta}_n - \underline{\theta}_n^0)\, (\underline{\theta}_n - \underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n - \underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n; \underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= p(\eta_0 = 0, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \cdots, \underline{z}_n) \Big[M_n^0$$

$$+ \, M_n^0\, \frac{1}{p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{n-1} = 0)}$$

$$\underline{\nabla}^2_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\, M_n^0 \,\Big], \qquad (4.390)$$

as in (4.354).

Consider next the remaining terms in (4.381). Note that since $\underline{\theta}_n^0$ and $\underline{T}_n$ are measurable with respect to the conditional probability measure in the integrals,

$$\int (\underline{\theta}_n^0-\underline{T}_n)(\underline{\theta}_n-\underline{\theta}_n^0)^{\mathrm{T}} f(\underline{z}_n-\underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n;\underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\,(\underline{\theta}_n^0-\underline{T}_n)(\underline{T}_n^0-\underline{\theta}_n^0)^{\mathrm{T}} \qquad (4.391)$$

$$\int (\underline{\theta}_n-\underline{\theta}_n^0)(\underline{\theta}_n^0-\underline{T}_n)^{\mathrm{T}} f(\underline{z}_n-\underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n;\underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\,(\underline{T}_n^0-\underline{\theta}_n^0)(\underline{\theta}_n^0-\underline{T}_n)^{\mathrm{T}} \qquad (4.392)$$

$$\int (\underline{\theta}_n^0-\underline{T}_n)(\underline{\theta}_n^0-\underline{T}_n)^{\mathrm{T}} f(\underline{z}_n-\underline{\theta}_n)\, \mathrm{N}(\underline{\theta}_n;\underline{\theta}_n^0, M_n^0)\, d\underline{\theta}_n$$

$$= p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\,(\underline{T}_n-\underline{\theta}_n^0)(\underline{T}_n-\underline{\theta}_n^0)^{\mathrm{T}}. \qquad (4.393)$$

Moreover, completing the square yields

$$(\underline{\theta}_n^0-\underline{T}_n)(\underline{T}_n^0-\underline{\theta}_n^0)^{\mathrm{T}} + (\underline{T}_n^0-\underline{\theta}_n^0)(\underline{\theta}_n^0-\underline{T}_n)^{\mathrm{T}} + (\underline{T}_n-\underline{\theta}_n^0)(\underline{T}_n-\underline{\theta}_n^0)^{\mathrm{T}}$$

$$= (\underline{T}_n-\underline{\theta}_n^0+\underline{\theta}_n^0-\underline{T}_n^0)(\underline{T}_n-\underline{\theta}_n^0+\underline{\theta}_n^0-\underline{T}_n^0)^{\mathrm{T}} - (\underline{\theta}_n^0-\underline{T}_n^0)(\underline{\theta}_n^0-\underline{T}_n^0)^{\mathrm{T}} \qquad (4.394)$$

$$= (\underline{T}_n-\underline{T}_n^0)(\underline{T}_n-\underline{T}_n^0)^{\mathrm{T}} - (\underline{\theta}_n^0-\underline{T}_n^0)(\underline{\theta}_n^0-\underline{T}_n^0)^{\mathrm{T}} \qquad (4.395)$$

$$= (\underline{T}_n-\underline{T}_n^0)(\underline{T}_n-\underline{T}_n^0)^{\mathrm{T}}$$

$$- M_n^0 \frac{1}{(\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\,)^2}$$

$$\underline{\nabla}_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)$$

$$\underline{\nabla}^{\mathrm{T}}_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\, M_n^0, \qquad (4.396)$$

from (4.355)-(4.356). But since

$$\underline{\nabla}_{\underline{z}_n} \frac{1}{p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)}$$

$$\underline{\nabla}^{\mathrm{T}}_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)$$

$$= \frac{1}{p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)}$$

$$\underline{\nabla}^2_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)$$

$$- \frac{1}{(\, p(\, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{n-1}=0\,)\,)^2}$$

$$\underline{\nabla}_{\underline{z}_n}\, p(\, \underline{z}_n \mid \underline{z}_0, \, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \, \cdots, \eta_{n-1} = 0 \,)$$

$$\underline{\nabla}_{\underline{z}_n}^{\mathrm{T}}\, p(\, \underline{z}_n \mid \underline{z}_0, \, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \, \cdots, \eta_{n-1} = 0 \,), \tag{4.397}$$

it follows, by substituting (4.390), (4.396), and (4.397) into (4.381), that

$$\frac{(\, 1 - \varepsilon \,)^n \; \kappa_n \; \kappa_n^0}{p(\, \underline{z}_n \mid \underline{z}_0, \, \cdots, \underline{z}_{n-1} \,)}$$

$$\int (\underline{\theta}_n - \underline{T}_n \,)\, (\underline{\theta}_n - \underline{T}_n \,)^{\mathrm{T}} f(\, \underline{z}_n - \underline{\theta}_n \,)\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^0, M_n^0 \,)\, d\underline{\theta}_n$$

$$= p(\, \eta_0 = 0, \, \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \, \cdots, \underline{z}_n \,)$$

$$\left[M_n^0 \; - \; M_n^0\, \Psi_n^0(\, \underline{z}_n - \underline{\theta}_n^0 \,)\, M_n^0 \; + \; (\underline{T}_n - \underline{T}_n^0)\, (\underline{T}_n - \underline{T}_n^0)^{\mathrm{T}} \right] \tag{4.398}$$

$$= (\, 1 - \varepsilon)^n \;\; \kappa_{n+1}\, \pi_n^0 \;\; \Sigma_n^0, \tag{4.399}$$

where (4.398) follows from (4.317) and (4.399) from (4.326) as well as (4.359), (4.362), (4.252), (4.168), and (4.321).

Consider now each term in the summation in (4.379):

$$\frac{\varepsilon(\, 1 - \varepsilon \,)^{n-1} \; \kappa_n \; \kappa_n^i}{p(\, \underline{z}_n \mid \underline{z}_0, \, \cdots, \underline{z}_{n-1} \,)} \int (\underline{\theta}_n - \underline{T}_n \,)\, (\underline{\theta}_n - \underline{T}_n \,)^{\mathrm{T}} f(\, \underline{z}_n - \underline{\theta}_n \,)\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i, M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \underline{\xi};\, \underline{v}_n^i + V_n^i\, (\underline{\theta}_n - \underline{\theta}_n^i \,),\, W_n^i - V_n^i\, M_n^i\, V_n^{i\;\mathrm{T}} \,)\, h(\underline{\xi})\, d\underline{\xi}\; d\underline{\theta}_n$$

$$= \frac{\varepsilon(\, 1 - \varepsilon \,)^{n-1} \; \kappa_n \; \kappa_n^i}{p(\, \underline{z}_n \mid \underline{z}_0, \, \cdots, \underline{z}_{n-1} \,)}$$

$$\int (\underline{\theta}_n - \underline{T}_n \,)\, (\underline{\theta}_n - \underline{T}_n \,)^{\mathrm{T}} \left[(1 - \varepsilon)\, \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n , R \,) + O(\, \varepsilon \,) \right]$$

$$\mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i, M_n^i \,)\, g(\, \underline{z}_{i-1} - V_n^i\, \underline{\theta}_n \,)\, d\underline{\theta}_n \tag{4.400}$$

from (4.157) and (4.365), where

$$\int (\underline{\theta}_n - \underline{T}_n \,)\, (\underline{\theta}_n - \underline{T}_n \,)^{\mathrm{T}}\, \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n , R \,)$$

$$\mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i, M_n^i \,)\, g(\, \underline{z}_{i-1} - V_n^i\, \underline{\theta}_n \,)\, d\underline{\theta}_n$$

$$= \int (\underline{\theta}_n - \underline{T}_n \,)\, (\underline{\theta}_n - \underline{T}_n \,)^{\mathrm{T}}\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i + M_n^i\, \Gamma_n^{i\;\;-1}(\, \underline{z}_n - \underline{\theta}_n^i \,),\, P_n^i \,)$$

$$g(\, \underline{z}_{i-1} - V_n^i\, \underline{\theta}_n \,)\, d\underline{\theta}_n \;\; \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n^i, \Gamma_n^i \,) \tag{4.401}$$

$$= \int \left[\underline{\theta}_n - \underline{\theta}_n^i - M_n^i\, \Gamma_n^{i\;\;-1}(\, \underline{z}_n - \underline{\theta}_n^i \,) \right] \left[\underline{\theta}_n - \underline{\theta}_n^i - M_n^i\, \Gamma_n^{i\;\;-1}(\, \underline{z}_n - \underline{\theta}_n^i \,) \right]^{\mathrm{T}}$$

$$\mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i + M_n^i \, \Gamma_n^{i\ -1} (\, \underline{z}_n - \underline{\theta}_n^i \,),\, P_n^i \,)$$

$$g(\, \underline{z}_{i-1} - V_n^i \underline{\theta}_n \,)\, d\underline{\theta}_n \;\; \mathrm{N}(\, \underline{z}_n ; \underline{\theta}_n^i , \Gamma_n^i \,), \tag{4.402}$$

where (4.401) follows from Lemma 4.2. Comparing (4.402) with (4.380), using (4.368), (4.370), and (4.372), and following the same reasoning as before, establishes that

$$\frac{\varepsilon(\, 1-\varepsilon\,)^{n-1}\, \kappa_n \; \kappa_n^i}{p(\, \underline{z}_n \mid \underline{z}_0, \; \cdots, \underline{z}_{n-1}\,)} \int (\underline{\theta}_n - \underline{T}_n\,)(\underline{\theta}_n - \underline{T}_n\,)^{\mathrm{T}} f(\, \underline{z}_n - \underline{\theta}_n \,)\, \mathrm{N}(\, \underline{\theta}_n ; \underline{\theta}_n^i , M_n^i \,)$$

$$\int \mathrm{N}(\, \underline{z}_{i-1} - \xi ; \underline{v}_n^i + V_n^i (\underline{\theta}_n - \underline{\theta}_n^i\,),\, W_n^i - V_n^i M_n^i \, V_n^{i\,\mathrm{T}} \,)\, h(\xi)\, d\xi\, d\underline{\theta}_n$$

$$= \varepsilon(\, 1-\varepsilon\,)^{n-1}\, \kappa_{n+1}\, \pi_n^i \, \Sigma_n^i \; + \; O(\, \varepsilon^2 (1-\varepsilon)^{n-1}\,), \tag{4.403}$$

so that

$$\Sigma_n \; = \; (1-\varepsilon)^n \;\kappa_{n+1}\, \pi_n^0 \, \Sigma_n^0 \; + \; \varepsilon\,(1-\varepsilon)^{n-1}\, \kappa_{n+1} \sum_{i=1}^{n} \pi_n^i \; \Sigma_n^i \quad + \quad O(\, \varepsilon^2 (1-\varepsilon)^{n-1}\,). \tag{4.404}$$

Once again, the equivalence of (4.404) and (4.325) follows from Theorem 4.2, completing the proof. (This theorem generalizes the result in Masreliez, 1975.) ■

Remark The approximate conditional-mean estimator of Theorem 4.3 has the following properties:

(i) The analogue of equations (4.318) and (4.325) for the case $n < \omega$ are equations (4.376) and (4.404).

(ii) Both Theorem 4.2 and Theorem 4.3 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 (4.320) -- which shows that $\underline{T}_n^i$ is linear in $\underline{z}_n$ -- 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 to outliers occurring in batches. This issue is further discussed later in this section.

(iii) It is easy to see that

$$(\, 1-\varepsilon\,)^n \;\; \kappa_{n+1}\, \pi_n^0 \; = \; p(\, \eta_0 = 0, \; \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \; \cdots, \underline{z}_n \,) \tag{4.405}$$

and

$$\varepsilon\,(1-\varepsilon\,)^{n-1} \;\; \kappa_{n+1}\, \pi_n^i \; = \; p(\, \eta_0 = 0, \; \cdots, \eta_{i-1} = 1, \; \cdots, \eta_{n-1} = 0 \mid \underline{z}_0, \; \cdots, \underline{z}_n \,), \tag{4.406}$$

i.e. 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.

(iv) Unlike the Kalman Filter, the estimation error covariance Σ_n (i.e. the conditional covariance of the state $\underline{\theta}_n$) *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 pre-computed and stored, as is sometimes done with the Kalman Filter. This would drastically reduce the on-line computational burden.

(v) The estimate of Theorem 4.3, 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 4.3 includes banks of parallel filters and smoothers that are entirely independent of each other. This suggests that the estimate derived here is well suited to parallel computation.

(vi) The error covariance Σ_n includes a weighted sum of quadratic terms of the form $(\underline{T}_n - \underline{T}_n^i)(\underline{T}_n - \underline{T}_n^i)^{\mathrm{T}}$. 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.

(vii) The "robust Kalman Filter" of Masreliez and Martin (1974, 1977) is approximately equivalent to the zeroeth-order term in equation (4.318), i.e. to $\underline{T}_n^0$ as given in (4.319). 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 $\underline{T}_n^i$ with the robust smoother of Martin (1979).

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

$$\underline{\psi}_n^0(\underline{z}_n - \underline{\theta}_n^0) = -\frac{\underline{\nabla}_{\underline{z}_n} \mathrm{N}(\underline{z}_n; \underline{\theta}_n^0, \Gamma_n^0)}{\mathrm{N}(\underline{z}_n; \underline{\theta}_n^0, \Gamma_n^0)} \tag{4.407}$$

$$= \Gamma_n^{0\ -1}(\underline{z}_n - \underline{\theta}_n^0), \tag{4.408}$$

so that $\underline{T}_n$ reduces to the Kalman Filter.

The following corollary is immediate.

Corollary 4.3 Let the conditions of Theorem 4.1, Corollary 4.1, and Theorem 4.2 be satisfied for the system (4.1) and (4.9). If h is bounded and differentiable a.e., then

$$\underline{T}_n = (1-\varepsilon)^{\omega} \kappa_{n+1} \pi_n^0 \underline{T}_n^0 + \varepsilon (1-\varepsilon)^{\omega-1} \kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \underline{T}_n^i + O_p(\omega^2 \varepsilon^2) \tag{4.409}$$

for all $n \geq \omega$, where

$$\underline{T}_n^0 = \underline{\theta}_n^0 + M_n^0 H_n^{\mathrm{T}} \underline{\psi}_n^0(\underline{z}_n - H_n \underline{\theta}_n^0) \tag{4.410}$$

$$\underline{T}_n^i = \underline{\theta}_n^i + M_n^i H_n^{\mathrm{T}} \Gamma_n^{i\ -1}(\underline{z}_n - H_n \underline{\theta}_n^i) + P_n^i V_n^{i\ \mathrm{T}} H_{i-1}^{\mathrm{T}} \underline{\psi}_n^i(\underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i) \tag{4.411}$$

$$\pi_n^0 = (1-\varepsilon) \kappa_{n+1}^0 + \varepsilon \kappa_n^0 \int \mathrm{N}(\underline{z}_n - \underline{\xi}; H_n \underline{\theta}_n^0, M_n^0) h(\underline{\xi}) d\underline{\xi} \tag{4.412}$$

$$\pi_n^i = (1-\varepsilon) \kappa_{n+1}^i \int \mathrm{N}(\underline{z}_{i-1} - \underline{\xi}; H_{i-1} \underline{v}_{n+1}^i, W_{n+1}^i) h(\underline{\xi}) d\underline{\xi} \tag{4.413}$$

$$\begin{aligned}\underline{\psi}_n^0(\underline{z}_n - H_n\,\underline{\theta}_n^0) = -\Big[&(1-\varepsilon)\,\mathrm{N}(\,\underline{z}_n;\,H_n\,\underline{\theta}_n^0,\,\Gamma_n^0\,)\\ &+\ \varepsilon\int \mathrm{N}(\,\underline{z}_n-\underline{\xi};\,H_n\,\underline{\theta}_n^0,\,M_n^0\,)\,h(\underline{\xi})\,d\underline{\xi}\,\Big]^{-1}\\ &\underline{\nabla}_{\underline{z}_n}\Big[(1-\varepsilon)\,\mathrm{N}(\,\underline{z}_n;\,H_n\,\underline{\theta}_n^0,\,\Gamma_n^0\,)\\ &+\ \varepsilon\int \mathrm{N}(\,\underline{z}_n-\underline{\xi};\,H_n\,\underline{\theta}_n^0,\,M_n^0\,)\,h(\underline{\xi})\,d\underline{\xi}\,\Big]\end{aligned}\tag{4.414}$$

$$\underline{\psi}_n^i\,(\underline{z}_{i-1}-H_{i-1}\underline{v}_{n+1}^i) = -\,\frac{\underline{\nabla}_{\underline{z}_{i-1}}\int \mathrm{N}(\,\underline{z}_{i-1}-\underline{\xi};\,H_{i-1}\underline{v}_{n+1}^i,\,W_{n+1}^i\,)\ \ h(\underline{\xi})\,d\underline{\xi}}{\int \mathrm{N}(\,\underline{z}_{i-1}-\underline{\xi};\,H_{i-1}\underline{v}_{n+1}^i,\,W_{n+1}^i\,)\ \ h(\underline{\xi})\,d\underline{\xi}}\tag{4.415}$$

with $\underline{\theta}_n^i$, M_n^i, P_n^i, Γ_n^i, V_n^i, $\underline{v}_n^i$, W_n^i, κ_n^i, and κ_n as defined in equations (4.304)-(4.312), subject to the initial conditions (4.169)-(4.177). Furthermore,

$$\Sigma_n \;=\; (1-\varepsilon)^{\omega}\,\kappa_{n+1}\,\pi_n^0\,\Sigma_n^0 \;+\; \varepsilon\,(1-\varepsilon)^{\omega-1}\,\kappa_{n+1}\sum_{i=n-\omega+1}^{n}\pi_n^i\,\Sigma_n^i \;+\; O_p(\omega^2\varepsilon^2)\tag{4.416}$$

for all $n \geq \omega$, where

$$\Sigma_n^0 \;=\; M_n^0 \;-\; M_n^0 H_n^{\mathrm{T}}\,\Psi_n^0(\,\underline{z}_n - H_n\,\underline{\theta}_n^0\,)\,H_n\,M_n^0 \;+\; (\underline{T}_n - \underline{T}_n^0)\,(\underline{T}_n - \underline{T}_n^0)^{\mathrm{T}}\tag{4.417}$$

$$\Sigma_n^i \;=\; P_n^i \;-\; P_n^i\,V_n^{i\ \mathrm{T}} H_{i-1}^{\mathrm{T}}\ \Psi_n^i(\,\underline{z}_{i-1} - H_{i-1}\underline{v}_{n+1}^i\,)\,H_{i-1}V_n^i\,P_n^i \;+\; (\underline{T}_n - \underline{T}_n^i)\,(\underline{T}_n - \underline{T}_n^i)^{\mathrm{T}},\tag{4.418}$$

and Ψ_n^i is given by equation (4.317), with (4.414) and (4.415).

Proof The proof is identical to that of Theorem 4.3, and is omitted. ■

The matter of the linearity of $\underline{T}_n^i$ in $\underline{z}_n$ is an important limitation of the estimator presented here. One way of dealing with this problem is to retain the function f rather than making the approximation of equation (4.364). Thus, using (4.365), each term in the summation in (4.333) yields

$$\begin{aligned}&\int \underline{\theta}_n\ \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n\\ &\quad= M_n^i\int M_n^{i\ -1}(\,\underline{\theta}_n-\underline{\theta}_n^i\,)\,\mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n\\ &\qquad+\ \underline{\theta}_n^i\int \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n,\end{aligned}\tag{4.419}$$

and proceeding as before (equations (4.343)-(4.352)),

$$\int M_n^{i\ -1}(\,\underline{\theta}_n-\underline{\theta}_n^i\,)\,\mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n$$

$$= -\int \underline{\nabla}_{\underline{\theta}_n}\mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n\tag{4.420}$$

$$= \int \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,\underline{\nabla}_{\underline{\theta}_n}\Big[f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\Big]\,d\underline{\theta}_n\tag{4.421}$$

$$\begin{aligned}&= \int \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,\Big[\underline{\nabla}_{\underline{\theta}_n}f(\,\underline{z}_n-\underline{\theta}_n\,)\Big]\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n\\ &\qquad+\int \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,\Big[\underline{\nabla}_{\underline{\theta}_n}g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\Big]\,d\underline{\theta}_n\end{aligned}\tag{4.422}$$

$$= -\ \underline{\nabla}_{\underline{z}_n}\int \mathrm{N}(\,\underline{\theta}_n;\,\underline{\theta}_n^i,\,M_n^i\,)\,f(\,\underline{z}_n-\underline{\theta}_n\,)\,g(\,\underline{z}_{i-1}-V_n^i\,\underline{\theta}_n\,)\,d\underline{\theta}_n$$

$$- \, V_n^{i\ \mathrm{T}} \, \underline{\nabla}_{\underline{z}_{i-1}} \int \mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i) \, f(\underline{z}_n - \underline{\theta}_n) \, g(\underline{z}_{i-1} - V_n^i \underline{\theta}_n) \, d\underline{\theta}_n , \quad (4.423)$$

where (4.421) follows from integration by parts, (4.423) holds by the dominated convergence theorem, and use is made of equation (4.368). Since

$$\int \mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i) \, f(\underline{z}_n - \underline{\theta}_n) \, g(\underline{z}_{i-1} - V_n^i \underline{\theta}_n) \, d\underline{\theta}_n$$

$$= p(\underline{z}_{i-1}, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1},$$

$$\eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0), \qquad (4.424)$$

it follows that $\underline{T}_n^i$ may be expressed as a function of influence-bounding functions that are the scores of a *joint* probability distribution. Similar arguments also apply to the derivation of Σ_n^i. An obvious difficulty with this approach is that, since the function g is itself a convolution, equation (4.424) represents a double convolution. This may be difficult to obtain in practice, except in cases where h has a special and convenient form. If, however, it is assumed that $\underline{z}_n$ and $\underline{z}_{i-1}$ are nearly conditionally independent, i.e. that

$$p(\underline{z}_{i-1}, \underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0)$$

$$= p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0)$$

$$p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0 = 0, \cdots, \eta_{i-1} = 1, \cdots, \eta_{n-1} = 0),$$

$$+ \, \Delta, \qquad (4.425)$$

where Δ is sufficiently small, then it is easy to see that equation (4.320) becomes

$$\underline{T}_n^i = \underline{\theta}_n^i + M_n^i \, \tilde{\underline{\psi}}_n^i (\underline{z}_n - \underline{\theta}_n^i) + M_n^i V_n^{i\ \mathrm{T}} \, \underline{\psi}_n^i (\underline{z}_{i-1} - \underline{v}_n^i), \qquad (4.426)$$

with appropriately defined influence-bounding functions. An important difficulty remains, however: while $\underline{T}_n^i$ is no longer a linear function of $\underline{z}_n$, it is easy to see (equation (4.161)) that $\underline{\theta}_{n+1}^i$ still is. Thus, while the influence of the outlier may be bounded at the current time, it is not bounded for future time steps. This limitation is due to the fact that the approximations are of first order. Under the assumption that at most one outlier occurs within ω time intervals, the posterior probabilities multiplying each term in (4.160) and (4.318) take care of bounding the influence of the outlier; thus, there is no further need for any non-linearity in (4.161). When that assumption is violated, however, this mechanism fails, and the influence of multiple outliers cannot be controlled. Using a second-order approximation would eliminate the non-robustness of the estimator against pairs of outliers within less than ω time intervals, but not, of course, against three or more outliers within the same period. Higher-order approximations are briefly mentioned in Section 6.2.

4.3 Further Approximations to the Conditional Mean Estimator

The estimator of Theorem 4.3 and its corollary makes explicit use of the exponential stability of the Kalman Filter. It retains a finite number ω of terms only, so that the complexity of the estimator does not increase without bound as $n \rightarrow \infty$. However, as the proof of Theorem 4.1 makes clear, the parameter ω is a function of various upper and lower bounds, and is therefore necessarily conservative. It may often be sufficient to retain a much smaller number of terms to preserve a comparable degree of accuracy. Furthermore, ω is not exactly trivial to obtain, making ways of eliminating this parameter quite desirable.

This section briefly discusses a number of further approximations, designed to simplify the estimator without sacrificing precision. The first two are based upon tests to decide which of the possible outlier configurations (i.e. each set $\{ \eta_0, \cdots, \eta_{n-1} \}$) are significant. The third is geared towards making a hard decision as to which single configuration best represents the observation history, and retaining it alone. As pointed out in Section 1.2, such hard decisions sometimes result in better performance at the nominal (i.e. normal) model.

Approximation 4.1 For the sake of discussion, suppose the current time is $n > \omega$, and let $\tilde{I}_n$ denote the set of integers $\{ n-\omega+1, \cdots, n \}$. For each $i \in \tilde{I}_n$, it is desired to make a decision as to whether or not to retain the corresponding term in the conditional prior distribution given by (4.160).

Clearly, if no outlier has occurred, or if one has occurred long enough ago that its effects on the ith term have sufficiently attenuated, then the ith term is indistinguishable from the 0th term and can be aggregated with it. Consider the alternative hypotheses

$$\begin{aligned} &\mathbf{H}_0 : \ \underline{\theta}_n^i \text{ is normally distributed} \\ &\mathbf{H}_1 : \ \underline{\theta}_n^i \text{ is not normally distributed} \end{aligned} \tag{4.427}$$

and the test statistic

$$S = (\underline{\theta}_n^i - \underline{\theta}_n^0)^{\mathrm{T}} M_n^{0 \ -1} (\underline{\theta}_n^i - \underline{\theta}_n^0) \tag{4.428}$$

Under $\mathbf{H}_0$, S is χ^2-distributed with q degrees of freedom: if the null hypothesis cannot be rejected, then the ith term can be dropped (i.e. consolidated with the 0th term).

Note that this test is designed to be conservative: if an outlier has occurred recently, then neither $\underline{\theta}_n^0$ nor $\underline{\theta}_n^i$ will be normally distributed. Thus, even though they might both be invalid models, they are each retained individually. Note also that the norming matrix in (4.428) is the inverse of M_n^0, not that of $\frac{1}{2}(M_n^0 + M_n^i)$ as might be expected. This too is to ensure that the test is conservative: the nominal covariance matrix M_n^0 is based on the hypothesis that no outliers are present, and is thus the minimum obtainable error covariance. Hence, the statistic S will be very sensitive to differences between the two estimates $\underline{\theta}_n^0$ and $\underline{\theta}_n^i$, i.e. the test will be powerful.

Defining I_n to be the set of all i for which $\mathbf{H}_0$ can be rejected, the conditional prior takes the form

$$p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1})$$

$$= (1-\varepsilon)^{\omega_n} \,\bar{\kappa}_n \,\kappa_n^0 \,\mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0)$$

$$+ \;\varepsilon(1-\varepsilon)^{\omega_n-1} \,\bar{\kappa}_n \sum_{i \in I_n} \kappa_n^i \,\mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i)$$

$$\int \mathrm{N}(\underline{z}_{i-1}-\xi ; \underline{v}_n^i + V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}) \, h(\xi) \, d\xi$$

$$+ \; O_p(\omega_n^2 \varepsilon^2), \tag{4.429}$$

where ω_n is the number of elements in the set I_n, and the normalization $\bar{\kappa}_n$ is defined appropriately. Similarly, the estimator is given by

$$\underline{T}_n = (1-\varepsilon)^{\omega_n} \,\bar{\kappa}_{n+1} \,\pi_n^0 \,\underline{T}_n^0 + \varepsilon(1-\varepsilon)^{\omega_n-1} \,\bar{\kappa}_{n+1} \sum_{i \in I_n} \pi_n^i \,\underline{T}_n^i \; + \; O_p(\omega_n^2 \varepsilon^2), \tag{4.430}$$

and its error covariance by

$$\Sigma_n = (1-\varepsilon)^{\omega_n} \,\bar{\kappa}_{n+1} \,\pi_n^0 \,\Sigma_n^0 + \varepsilon(1-\varepsilon)^{\omega_n-1} \,\bar{\kappa}_{n+1} \sum_{i \in I_n} \pi_n^i \,\Sigma_n^i \; + \; O_p(\omega_n^2 \varepsilon^2). \tag{4.431}$$

To ensure that the error term is still $O_p(\varepsilon^2)$, it is necessary to choose the level of the hypothesis test accordingly. Since M_n^0 is bounded below, by equation (4.90), requiring that

$$S = O(\varepsilon^2) \tag{4.432}$$

achieves the desired accuracy, by virtue of equation (4.127).

The same algorithm is implemented at the next time step $n+1$, starting with the set $\tilde{I}_{n+1} := \{ n+1 \} \bigcup I_n \quad \{ n-\omega+1 \}$.

Approximation 4.2 The conditional prior means $\underline{\theta}_n^i$ are easy to compute, making Approximation 4.1 easy to implement. Under some conditions, the posterior probabilities of each outlier configuration may also be easy to obtain. In those cases, a more direct approximation is possible.

Equations (4.184)-(4.185) and (4.405)-(4.406) show that the coefficients of each term in the expressions for the conditional prior and the conditional mean (respectively) are equal to the posterior probabilities that each outlier configuration has occurred. It is intuitively clear that those terms corresponding to the most improbable models may be dropped, resulting in simpler expressions and reduced computational burden. Since M_n^0 is bounded below, and h is bounded above. each term is itself bounded, and the coefficients can therefore be used for this purpose.

Always retaining the nominal (0th) term, a criterion for dropping terms from the expresion for the conditional prior is

$$\varepsilon(1-\varepsilon)^{\omega-1} \,\kappa_n \,\kappa_n^i \int \mathrm{N}(\underline{z}_{i-1}-\xi ; \underline{v}_n^i, W_n^i) \, h(\xi) \, d\xi \;\leq\; O_p(\varepsilon^2 (1-\varepsilon)^{\omega-2}), \tag{4.433}$$

or equivalently

$$\varepsilon(1-\varepsilon) \,\kappa_n \,\kappa_n^i \int \mathrm{N}(\underline{z}_{i-1}-\xi ; \underline{v}_n^i, W_n^i) \, h(\xi) \, d\xi \;\leq\; O_p(\varepsilon^2). \tag{4.434}$$

Similarly, a criterion for dropping terms from the expresion for the conditional mean is

$$\varepsilon\,(\,1-\varepsilon\,)\,\kappa_{n+1}\,\pi_n^i \;\leq\; O_p(\,\varepsilon^2\,). \tag{4.435}$$

As before, ω_n is the number of terms retained, and the approximate distribution, mean, and variance are given by equations (4.429)-(4.431), respectively, with I_n redefined as the set of i that do not satisfy (4.434)-(4.435).

Approximation 4.3 Finally, a third approximation is based upon choosing only one term at any given time, i.e. making a hard decision as to which model best represents reality. Although this approach is somewhat *ad hoc* and lacks strong theoretical justification, it is nevertheless attractive for the following reasons:

(i) The principal difficulty in implementing the estimator of Theorem 4.3 and its corollary is the need to perform, in real time, several convolutions at each time step. These convolutions are needed to compute both the weights of the parallel estimates and the overall normalization coefficient. Retaining only one term reduces the number of convolutions that need to be calculated to at most one per time step, and only following the detection of an outlier.

(ii) It was mentioned in Section 4.2 that the estimator of Theorem 4.3 is non-robust when two or more outliers occur within less than ω time intervals. In this approximation, a test is performed to detect outliers at each time step, and appropriate action is taken when one is detected, regardless of how recently a previous outlier may have occurred. This results in an estimator that is more resistant to the effects of a burst of outliers than that of Theorem 4.3.

(iii) As stated earlier in reference to the estimators of Guttman and Peña (1984, 1985) and Ershov and Lipster (1978), using mixture distributions as priors in a Bayesian setting can result in drastically reduced performance at the nominal model, unless competing models have negligible overlap. Thus, unless the outlier distribution h is such that the posterior probabilities for each outlier configuration are always either near zero or near unity, some "smearing" is likely to occur, resulting in suboptimal performance when no outliers are present. That effect is eliminated when a hard decision is made and only one term is retained: in that case, whenever no outlier is detected, the optimal (Kalman Filter) term remains in use. Furthermore, the main argument against such a hard decision -- the question as to what to do in case of uncertainty -- is not relevant here: if an observation is not an "obvious" outlier, then a conservative approach would dictate that it be treated as an ordinary observation.

For an observation $\underline{z}_n$, consider the alternative hypotheses

$$\begin{aligned} &\mathbf{H}_0:\ \underline{z}_n \text{ is normally distributed} \\ &\mathbf{H}_1:\ \underline{z}_n \text{ is not normally distributed} \end{aligned} \tag{4.436}$$

and the test statistic

$$S \;=\; (\,\underline{z}_n - H_n\,\underline{\theta}_n^j\,)^{\mathrm{T}}\,\Gamma_n^{j\ -1}\,(\,\underline{z}_n - H_n\,\underline{\theta}_n^j\,) \tag{4.437}$$

Initially, j equals zero. As long as no outliers occur, $\underline{\theta}_n^0$ is clearly normally distributed (it is the mean of a Gaussian process conditioned on Gaussian observations). Thus, under $\mathbf{H}_0$, S is χ^2-distributed with p degrees of freedom. As long as the null hypothesis cannot be rejected, the 0th term $\mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^0, M_n^0)$ (i.e. a standard Kalman Filter) can be used to approximate the conditional prior. Suppose the null hypothesis is rejected at some pre-selected significance level α for the observation $\underline{z}_{i-1}$. Then, set $j = i$. Since $\underline{\theta}_n^i$ is independent of $\underline{z}_{i-1}$ (equations (4.161) and (4.169)), it is still normally distributed, and the statistic defined in (4.437) can still be used in a χ^2 test.

Note furthermore that, by Bayes' rule,

$$
\begin{aligned}
& p(\underline{\theta}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \\
&= \frac{1}{p(\underline{z}_{i-1} \mid \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0)} \\
&\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \\
&\qquad p(\underline{z}_{i-1} \mid \underline{\theta}_n, \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \\
&\qquad\qquad \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \qquad (4.438) \\
&= \Bigl[\int p(\underline{z}_{i-1}, \underline{\theta}_n \mid \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \\
&\qquad\qquad \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \, d\underline{\theta}_n \Bigr]^{-1} \\
&\qquad p(\underline{\theta}_n \mid \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \\
&\qquad p(\underline{z}_{i-1} \mid \underline{\theta}_n, \underline{z}_0, \cdots \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}, \\
&\qquad\qquad \eta_0=0, \cdots, \eta_{i-1}=1, \cdots, \eta_{n-1}=0) \qquad (4.439) \\
&= \frac{1}{\int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i, W_n^i) \, h(\xi) \, d\xi} \, \mathrm{N}(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i) \\
&\qquad \int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i - V_n^i (\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}) \, h(\xi) \, d\xi, \qquad (4.440)
\end{aligned}
$$

from (4.214) and Fubini's theorem. Thus, following the detection of an outlier, the conditional prior density is given (approximately) by (4.440), until the effects of the outlier have sufficiently decayed. The point of sufficient attenuation can be determined by ensuring that

$$
1 - \varepsilon \le \frac{\int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i - V_n^i (\underline{\theta}_n - \underline{\theta}_n^i), W_n^i - V_n^i M_n^i V_n^{i\,\mathrm{T}}) \, h(\xi) \, d\xi,}{\int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i, W_n^i) \, h(\xi) \, d\xi} \le 1 + \varepsilon \qquad (4.441)
$$

for all $\underline{\theta}_n$, or, from equations (4.254)-(4.262), by verifying that

$$
\| V_n^i \| = O(\varepsilon), \qquad (4.442)
$$

which is simpler. Once this point is reached, the conditional prior may be approximated simply by the ith Kalman Filter, i.e. $N(\underline{\theta}_n ; \underline{\theta}_n^i, M_n^i)$.

In the unlikely event that another outlier is detected prior to this point, say z_{k-1}, it would appear that the above process can be repeated starting at the new outlier, i.e. replacing i in (4.432) by k, and using a modification of the initial conditions (4.169)-(4.174) so that the ith filter, not the nominal (0th) one, is used to initialize the kth term.

The equivalent algorithm for the conditional mean estimator and its variance should be obvious from the above discussion. The estimator is $\underline{T}_n^0$, until an outlier is detected. Following detection, $\underline{T}_n^i$ is used until the effects of the outlier are judged to have decayed sufficiently, or until another outlier is detected. A slight modification would consist in performing the outlier test *prior to*, not following, each update. In that case, a Kalman Filter would be used until an outlier is detected, $\underline{T}_n^0$ at the time of detection, and $\underline{T}_n^i$ following it. If Huber's influence-bounding function ψ_ε is used, as discussed further in Section 4.4, and an observation is considered an outlier when it lies in the region of truncation, the two methods are identical.

4.4 Choosing the Noise Distribution

As discussed in previous sections, 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, i.e. non-robust.

In Sections 2 and 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, i.e. optimal in the minimax sense.

In Section 4, 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 random, and the problem consists in optimally tracking it, rather than achieving minimum asymptotic estimation error. Thus, approximations to a conditional mean estimator are sought, since such estimators are known to achieve minimum error variance at each point in time. Throughout the discussion in Sections 4.1-4.3, however, the "outlier" noise distribution H is treated as known. In other words, the results of Sections 4.1-4.3 are better characterized as *non-Gaussian* filters than *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.

It is clear from equations (4.318)-(4.320) and (4.325)-(4.327) 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 (e.g. see equation (4.362)). The main difference between Huber's formulation and this one is thus that the former involves the neighborhood P_ε, whereas the corresponding neighborhood in the latter case is

$$P_{\Phi_1,\Phi_2,\varepsilon} := \{ (1-\varepsilon)\Phi_1 + \varepsilon\Phi_2 * H : \ H \in \mathbf{S} \}, \tag{4.443}$$

where Φ_1 and Φ_2 are given zero-mean normal distributions. To appreciate the distinction, note that when $\Phi_1 = \Phi_2$, Huber's case involves *replacing* outliers, and (4.443) *additive* ones.

The problem of minimizing the Fisher information for the location parameter of neighborhoods of the form (4.443) 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. 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 (Brown, 1971) that the 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).

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 end points. 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, are able to derive exactly the distribution minimizing the Fisher information in this case.

A conclusion strongly implied by this discussion is that the least favorable distribution in the neighborhood $P_{\Phi_1,\Phi_2,\varepsilon}$ is of a highly complex shape and extremely difficult to derive, and, moreover, that since the very choice of neighborhood is to a large extent arbitrary, the effort necessary is perhaps unwarranted. An approximation (also suggested by Marazzi, 198?) consists of the following: since $P_{\Phi_1,\Phi_2,\varepsilon} \subset P_\varepsilon$, the least favorable distribution in P_ε clearly has Fisher information no greater than that in $P_{\Phi_1,\Phi_2,\varepsilon}$. Indeed, the least favorable distribution in P_ε (derived by Huber and given in Theorem 2.5) can easily be shown not to be a member of $P_{\Phi_1,\Phi_2,\varepsilon}$, by noting that the support of the minimizing H distribution is not $\mathbf{R}$, so that it cannot be the result of a convolution with a normal distribution Φ_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 $P_{\Phi_1,\Phi_2,\varepsilon}$. Consequently, a *conservative* approach to approximating a minimax solution is simply to use the least favorable distribution in P_ε; this has also the additional advantage of simplicity.

Approximation 4.4 Note first that the conditional distribution of the innovation term is given by

$$p(\, \underline{z}_n - \underline{\theta}_n^0 \mid \eta_0 = 0, \;\cdots, \eta_{n-1} = 0 \,)$$

$$= (1-\varepsilon)\, \mathrm{N}(\, \underline{z}_n - \underline{\theta}_n^0;\, 0,\, \Gamma_n^0 \,) \;+\; \varepsilon \int \mathrm{N}(\, \underline{z}_n - \underline{\theta}_n^0 - \xi;\, 0,\, M_n^0 \,)\, h(\xi)\, d\xi. \tag{4.444}$$

Thus, defining the normalized innovation as

$$\underline{e}_n \;:=\; \Gamma_n^{0\;-\frac{1}{2}} (\underline{z}_n - \underline{\theta}_n^0), \tag{4.445}$$

it follows that

$$p(\, \underline{e}_n \mid \eta_0 = 0, \;\cdots, \eta_{n-1} = 0 \,)$$

$$= (1-\varepsilon)\, \mathrm{N}(\, \underline{e}_n;\, 0,\, I \,)$$

$$+ \;\varepsilon \int \mathrm{N}(\, \underline{e}_n - \xi;\, 0,\, \Gamma_n^{0\;-\frac{1}{2}} M_n^0\, \Gamma_n^{0\;-\frac{1}{2}} \,) \mid \Gamma_n^0 \mid^{-\frac{1}{2}} h(\Gamma_n^{0\;-\frac{1}{2}} \xi)\, d\xi. \tag{4.446}$$

Suppose that h is such that the above distribution may be approximated by

$$p(\, \underline{e}_n \mid \eta_0 = 0, \;\cdots, \eta_{n-1} = 0 \,)$$

$$= (1-\varepsilon)\, \mathrm{N}(\, \underline{e}_n;\, 0,\, I \,) \;+\; \varepsilon\, h^*(\, \underline{e}_n \,) \;+\; \Delta, \tag{4.447}$$

where h^* is the Huber distribution of equation (2.163), and Δ is a remainder term. As discussed earlier, there is no h for which (4.447) holds with $\Delta = 0$, but there may be some for which Δ is small. A similar argument can be made to show that

$$p(\, W_{n+1}^{i\;-\frac{1}{2}} (\, \underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i \,) \,) \mid \eta_0 = 0, \;\cdots, \eta_{i-1} = 1, \;\cdots, \eta_{n-1} = 0 \,)$$

$$= h^*(\, \underline{e}_n \,) \;+\; \Delta'. \tag{4.448}$$

Thus, the estimator of Corollary 4.3 reduces to (4.409) with conditional estimators given by

$$\underline{T}_n^0 \;=\; \underline{\theta}_n^0 \;+\; M_n^0 H_n^{\mathrm{T}}\, \Gamma_n^{0\;-\frac{1}{2}}\, \underline{\psi}_\varepsilon(\, \Gamma_n^{0\;-\frac{1}{2}} (\underline{z}_n - H_n\, \underline{\theta}_n^0) \,) \tag{4.449}$$

and

$$\underline{T}_n^i \;=\; \underline{\theta}_n^i \;+\; M_n^i H_n^{\mathrm{T}} \Gamma_n^{i\;-1} (\, \underline{z}_n - H_n\, \underline{\theta}_n^i \,)$$

$$+ \;P_n^i\, V_n^{i\;\mathrm{T}} H_{i-1}^{\mathrm{T}}\, W_{n+1}^{i\;-\frac{1}{2}}\, \underline{\psi}_1(\, W_{n+1}^{i\;-\frac{1}{2}} (\underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i \,) \,). \tag{4.450}$$

Note in passing that ψ_1 is ψ_ε at the limit $\varepsilon = 1$, and the vector version is the same, componentwise. The coefficients π_n^0 and π_n^i, and the conditional covariances Σ_n^0 and Σ_n^i, are defined similarly:

$$\pi_n^0 \;=\; \kappa_n^0\, \underline{\psi}_\varepsilon(\, \Gamma_n^{i\;-\frac{1}{2}} (\underline{z}_n - H_n\, \underline{\theta}_n^0) \,) \tag{4.451}$$

$$\pi_n^i \;=\; (1-\varepsilon)\, \kappa_{n+1}^i\; \underline{\psi}_1(\, W_{n+1}^{i\;-\frac{1}{2}} (\underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i) \,) \tag{4.452}$$

$$\Sigma_n^0 \;=\; M_n^0 \;-\; M_n^0 H_n^{\mathrm{T}} \Gamma_n^{i\;-1}\, \Psi_\varepsilon(\, \Gamma_n^{i\;-\frac{1}{2}} (\underline{z}_n - H_n\, \underline{\theta}_n^0) \,)\, \Gamma_n^{i\;-1} H_n\, M_n^0$$

$$+ \;(\underline{T}_n - \underline{T}_n^0)\, (\underline{T}_n - \underline{T}_n^0)^{\mathrm{T}} \tag{4.453}$$

$$\Sigma_n^i \;=\; P_n^i \;-\; P_n^i V_n^{i\;\mathrm{T}} H_{i-1}^{\mathrm{T}}\, W_{n+1}^{i\;-\frac{1}{2}}\, \Psi_1(\, W_{n+1}^{i\;-\frac{1}{2}} (\underline{z}_{i-1} - H_{i-1} \underline{v}_{n+1}^i) \,)\, W_{n+1}^{i\;-\frac{1}{2}} H_{i-1} V_n^i\, P_n^i$$

$$+ \;(\underline{T}_n - \underline{T}_n^i)\, (\underline{T}_n - \underline{T}_n^i)^{\mathrm{T}}, \tag{4.454}$$

with Ψ_ε and Ψ_1 defined analogously, by equation (4.317). This approximation can, of course, be combined with either of Approximations 4.1-4.3.

Deriving a least favorable distribution for the neighborhood $P_{\Phi_1,\Phi_2,\varepsilon}$ seems to be destined to remain an open problem for a while longer. In the interim, it would appear that Approximation 4.4 provides a simple and intuitively appealing framework for the robust recursive estimation of the state of a stochastic dynamic system in the presence of heavy-tailed observation noise.

5. Numerical Examples

Section 4 describes the derivation of a new robust recursive estimator of the state of a discrete-time stochastic linear dynamic system in the presence of ε-contaminated Gaussian observation noise, as a first-order approximation to the conditional mean given all past observations. As discussed in Section 1.2, there are a number of other robust recursive estimators in the literature, but many are based on heuristic arguments and *ad hoc* assumptions, making a theoretical comparison all but impossible.

This section presents the results of some Monte Carlo simulation experiments, comparing the performance of several estimators for a number of different observation noise distributions. The purpose of these simulations is emphatically *not* to determine the *best* method: most estimators could be "tuned" to specific applications, and it is possible that better performance could be obtained given enough preparatory work. Rather, this section describes a comparison of several estimators, in their *published* forms, with the first-order conditional mean estimator derived here, in order to give a general idea of their respective strengths and weaknesses.

For simplicity, only the scalar time-invariant case is considered, with $F < 1$ and $H = 1$. In other words, the dynamic system is given by

$$\theta_{n+1} = F\,\theta_n + w_n \tag{5.1}$$

and

$$z_n = \theta_n + v_n, \tag{5.2}$$

where θ_0 and $\{w_n\}$ are independent random variables with distributions $\mathbf{N}(\,\theta_0;\,\overline{\theta}_0,\,M_0\,)$ and $\mathbf{N}(\,w_n;\,0,\,Q\,)$, respectively, and $\{v_n\}$ are independent identically distributed random variables with various distribution.

Section 5.1 discusses the observation noise distributions, and Section 5.2 describes the estimators to be compared. Performance criteria are discussed in Section 5.3, and experiment results follow.

5.1 Observation Noise Distributions

A good robust estimator has at least the following properties: it is resistant to outliers, and it looses minimal efficiency at the nominal model. To verify these properties, several observation noise distributions were used in the simulation experiments, ranging from very light- to very heavy-tailed ones. The choice of distributions follows the well-known Princeton robustness study (Andrews *et al.*, 1972, pp.67-68). The following distributions were used:

(i) *The Gaussian Distribution*. To verify the performance of each estimator at the nominal model, i.e. when no outliers are present, the normal distribution is used in the first set of experiments:

$$\mathbf{L}(\,v_n\,) = \mathbf{N}(\,v_n;\,0,\,R\,). \tag{5.3}$$

As discussed elsewhere, there is a tradeoff between efficiency at the nominal model and resistance to outliers, and it is worth comparing the performance of each estimator with and without observation outliers.

(ii) *The Scale-Contaminated Gaussian Distribution.* The most commonly used form in modeling outliers for detection and robustness studies is the two-component Gaussian mixture, where both distributions are zero-mean, but one has a greater variance than the other (see for example Titterington, Smith and Makov, 1985, pp.22-25):

$$\mathbf{L}(\ v_n\ ) \ =\ (\ 1-\varepsilon\ )\,\mathrm{N}(\ v_n\,;\ 0,\ R\ )\ +\ \varepsilon\,\mathrm{N}(\ v_n\,;\ 0,\ R_{out}\ ), \tag{5.4}$$

with $R_{out} > R$ and $0 < \varepsilon < 1$. Although the tails of the normal distribution are relatively light, this model is the basis of a number of robust estimators in the literature.

(iii) *The Laplace Distribution.* Heavier tails than the Gaussian mixture are provided by the Laplace (or double-exponential) distribution, which is used as a contaminant to the Gaussian distribution:

$$\mathbf{L}(\ v_n\ ) \ =\ (\ 1-\varepsilon\ )\,\mathrm{N}(\ v_n\,;\ 0,\ R\ )\ +\ \varepsilon\,\frac{1}{\sqrt{2R_{out}}}\,e^{-\sqrt{\frac{2}{R_{out}}}\,|\ v_n\ |}\ . \tag{5.5}$$

Here, the Laplace distribution is zero-mean and has variance equal to R_{out}. It is worth noting that, as shown earlier, Huber found the least favorable member of the ε-contaminated normal family to have exponential tails (in the no process noise case).

(iv) *Tukey's "Slash" Distribution.* This distribution, for which an analytical expression is not available, is defined as follows (Andrews *et al.*, 1972, p.68): Let

$$\mathbf{L}(\ x_n\ ) \ =\ \mathrm{N}(\ x_n\,;\ 0,\ 1\ ) \tag{5.6}$$

and

$$\mathbf{L}(\ y_n\ ) \ =\ \mathrm{U}(\ y_n\,;\ 0,\ 1\ ), \tag{5.7}$$

where $\mathrm{U}(\ y\,;\ 0,\ 1\ )$ denotes a uniform distribution over the interval $[\ 0,\ 1\ ]$. Then, the distribution of the random variable

$$v_n \ :=\ \frac{x_n}{y_n} \tag{5.8}$$

is named Tukey's "Slash" distribution. It is easy to see that it has extremely heavy tails, and can therefore be used to test the performance of robust estimators in the presence of very large outliers. It is used as a contaminant to a Gaussian distribution, as in Equations (5.4) and (5.5).

(v) *The Cauchy Distribution.* Another model, also for heavy-tailed noise, is the Cauchy distribution. It is also used as a contaminant:

$$\mathbf{L}(\ v_n\ ) \ =\ (\ 1-\varepsilon\ )\,\mathrm{N}(\ v_n\,;\ 0,\ R\ )\ +\ \varepsilon\,\frac{1}{\pi}\,\frac{1}{1+v_n^2}, \tag{5.9}$$

The Cauchy distribution above is zero-mean and has infinite variance. This distribution too is frequently used in robustness studies.

(vi) *Fixed-Amplitude Outliers.* To test the performance of robust estimators as a function of the magnitude of the outliers, the following distribution is also used:

$$L(v_n) = (1 - \varepsilon) N(v_n ; 0, R) + \varepsilon \delta(v_n - \sqrt{R_{out}}), \tag{5.10}$$

where $\delta(v_n - \sqrt{R_{out}})$ denotes the Dirac delta function.

5.2 Recursive Estimators

The following recursive estimators are used in the present study:

(i) *The Kalman Filter.* It is well known that the Kalman Filter is optimal both in the sense of minimizing the mean squared error (regardless of any distributional assumptions), and, if the noise is Gaussian, in the Bayesian sense (regardless, this time, of the cost function). Thus, it can be used as a benchmark in the nominal case. The performance of the Kalman Filter does, however, severely degrade in the presence of outliers. The appropriate equations appear in (1.3)-(1.9), with $F_n = F$, $Q_n = Q$, and $H_n = D_n = 1$.

(ii) *The Guttman-Peña Estimator.* As discussed in Section 1.2, Guttman and Peña (1984, 1985) propose a Bayesian framework for adjusting the Kalman gain *a posteriori*, according to the respective probabilities that an outlier has or has not occurred. In principle, this approach could be used for any two (i.e. underlying and outlier) noise distributions. Indeed, it performs best when the noise distributions have relatively disjoint supports. However, Guttman and Peña only give the scale-contaminated normal case, and do not treat other kinds of observation noise. The equations for this estimator are identical to those of the Kalman Filter, except that (1.5) is replaced by

$$\Gamma_n = M_n + R (z_n), \tag{5.11}$$

where the *posterior* observation noise covariance matrix $R (z_n)$ is given by

$$R (z_n) = \frac{(1 - \tilde{\varepsilon}) N(v_n ; 0, R) R + \tilde{\varepsilon} N(v_n ; 0, \tilde{R}_{out}) \tilde{R}_{out}}{(1 - \tilde{\varepsilon}) N(v_n ; 0, R) + \tilde{\varepsilon} N(v_n ; 0, \tilde{R}_{out})}, \tag{5.12}$$

where $\tilde{R}_{out}$ is the *modeled* outlier variance and $\tilde{\varepsilon}$ is the *modeled* fraction of contamination. The extension of these results to other noise distributions is not always trivial: for instance, an analytical expression is not available for Tukey's "Slash" distribution, while the Cauchy and Laplace distributions are not mixtures at all. Nevertheless, since outliers do occur rarely, it is possible that they can still be modeled adequately as a Gaussian mixture.

(iii) *The Ershov-Lipster Estimator.* As discussed before, the estimator of Ershov and Lipster (1978) is similar to that of Guttman and Peña (1984, 1985), with the exception that equation (5.12) is replaced by a hard decision, i.e.

$$R (z_n) = \begin{cases} R & \text{if } z_n \text{ is an outlier} \\ \tilde{R}_{out} & \text{otherwise} \end{cases} \tag{5.13}$$

The decision as to whether or not an outlier has occurred may be made in several ways. Here, since the nominal (underlying) model is assumed to be normal, a χ^2 test is performed at the significance level $\alpha = 0.05$ on the statistic

$$X = \frac{\gamma_n^2}{\Gamma_n}, \tag{5.14}$$

i.e. the normalized squared innovation.

It is worth noting that both the Guttman-Peña and the Ershov-Lipster estimators could be expressed in the form of (1.12) with influence-bounding functions ψ that are *not* flat for very large innovations. This suggests that, while they may be very efficient near the nominal model, their performance declines significantly for very heavy-tailed noise. This problem could be circumvented by deriving estimators based on the approaches of Guttman and Peña or Ershov and Lipster, but on distributional assumptions other than scale-contaminated normal noise. That, however, is not done here.

(iv) *The Masreliez-Martin Estimator.* Essentially, the estimator derived by Masreliez and Martin (1974, 1977) is equivalent to the 0th-order term of that given in (4.318). It has the distinct advantage of being robust in the presence of patchy outliers. However, since it is a lower-order approximation to the conditional mean estimator than (4.318), its overall estimation error variance can be expected to be higher. The equations for this estimator are similar to those of the Kalman Filter, with (1.3) and (1.8) replaced by

$$T_{n+1} = F\,T_n + \frac{M_{n+1}}{\sqrt{\Gamma_{n+1}}}\,\psi\left[\frac{\gamma_{n+1}}{\sqrt{\Gamma_{n+1}}}\right] \tag{5.15}$$

and

$$\Sigma_{n+1} = \dot{M}_{n+1} - \frac{M_{n+1}^2}{\Gamma_{n+1}}\,\Psi\left[\frac{\gamma_{n+1}}{\sqrt{\Gamma_{n+1}}}\right], \tag{5.16}$$

where ψ is given by (2.186) (based on the modeled fraction of contamination $\tilde{\varepsilon}$), and Ψ is as defined in (4.317).

(v) *The First-Order Approximation to the Conditional Mean.* This is the estimator of Theorem 4.3. The values of δ and ω can easily be approximated for the time-invariant case by fitting a straight line of the form $\beta_0 + \beta_1 n$ to the ordered pairs

$$\left[\, n,\ \log \prod_{i=1}^{n} F \left[1 - \frac{M_{n+1}^0}{\Gamma_{n+1}^0} \right] \right] \tag{5.17}$$

and noting that

$$\delta = e^{\beta_1} \tag{5.18}$$

approximately. The window size ω is then the smallest integer such that (4.158) holds. As discussed in Section 4.4, the influence bounding function ψ is chosen to be that given by (2.186) (based on the modeled fraction of contamination $\tilde{\varepsilon}$), and Ψ is as defined in (4.317).

5.3 Performance Measures

The choice of criteria by which to measure the performance of robust estimators presents certain difficulties. In particular, it is clear that a global performance measure such as the mean squared error only gives a partial picture of reality: for instance, one estimator may do very well at the nominal model but badly at an outlier, while another may do poorly at the nominal model but well at an outlier, and yet the two could have the same mean squared error.

Another important measure of fit is the whiteness (or near-whiteness) of the residual sequence. The residual of an estimator that tracks the state very well under nominal conditions may exhibit large and systematic excursions from zero immediately following an outlier; conversely, an estimator that is insensitive to observations may be resistent to outliers, but its residual sequence may be significantly non-white at the nominal model.

This suggests that separate criteria must be used for determining the performance of each estimator for observation noise with and without outliers. The following performance measures are calculated:

(i) *The Mean Squared Error.* This is computed in order to determine the performance of each estimator under nominal conditions. Given K simulation runs, each N time steps long, the mean-squared error is given by

$$\text{MSE} = \frac{1}{KN} \sum_{i=1}^{K} \sum_{n=1}^{N} (\theta_{n,i} - T_{n,i})^2. \tag{5.19}$$

This measure is only truly meaningful in the nominal (no outliers) case.

(ii) *Error at Outliers.* To measure the behavior of each estimator specifically at outliers, the following experiment is performed: instead of outliers occurring randomly, as described in Section 5.1, they are forced to occur at a given time ($n = 20$). Then, the mean squared error at each time $n = 21, 22, \cdots$ is computed by averaging over all K runs. Thus, for each n,

$$\text{MSE}_n = \frac{1}{K} \sum_{i=1}^{K} (\theta_{n,i} - T_{n,i})^2. \tag{5.20}$$

This allows an assessment both of the resistance of the estimator to an outlier when it occurs, and of the persistence of the effects of an outlier due to the dynamics of the estimator.

(iii) *Serial Correlation Following Outliers.* As above, outliers are forced to occur at a specific time, and the serial correlation (autocorrelation) of the residual sequence $\{\gamma_n\}$ is computed for $n = 21, 22, \cdots$ by averaging over all K runs. Thus,

$$\text{SC}_{n,n-1} = \frac{\sum_{i=1}^{K} \gamma_{n,i}\, \gamma_{n-1,i}}{\sqrt{\sum_{i=1}^{K} \gamma_{n,i}^2}\ \sqrt{\sum_{i=1}^{K} \gamma_{n-1,i}^2}} \tag{5.21}$$

for each n.

(iv) *Normalized Estimation Error Covariance.* Each estimator provides an expression for the estimation error covariance. However, these expressions are derived based on certain distributional assumptions, and how close they are to the true covariances is not immediately clear. For this reason, the normalized error covariance is computed for each time n as

$$\text{NMSE}_n = \frac{1}{K} \sum_{i=1}^{K} \frac{(\theta_{n,i} - T_{n,i})^2}{\Sigma_{n,i}} \tag{5.22}$$

where Σ represents the theoretical variances of equations (1.8), (5.16), and (4.325). The more accurate the covariance estimate, the closer NMSE will be to unity. This criterion has the added advantage that it allows a comparison of the first two theoretical and empirical moments, giving an idea of the accuracy of a normal approximation to the distribution of the estimator.

5.4 Simulation Results

This section summarizes the results of 63 simulation experiments, comparing the performance of each estimator described in Section 5.2 under each noise distribution of Section 5.1. Each experiment consists of $K = 200$ runs of $N = 50$ time steps each, with initial conditions $\bar{\theta}_0 = 0$ and $M_0 = 1$. Random number generators from the IMSL package were used, and the model parameters were as follows: $Q = 1, R = 1, F = 0.1$ (unless otherwise noted), modeled outlier standard deviation equal to 2, 2.5, and 3 times the nominal standard deviation (i.e. $\tilde{R}_{out} = 4$, 6.25, and 9), and finally $\tilde{\varepsilon} = 0.01$, 0.05, and 0.10.

It is worth noting that the recursive computation of the coefficients κ_n^i in Theorem 4.3 presented some numerical difficulties: as they tended to vanish with respect to machine precision, periodic rescaling was necessary. Similarly, the probabilities for alternative hypotheses (outlier v.s. not outlier) in both Theorem 4.3 and the Guttman-Peña estimator tended to vanish with respect to machine precision when outliers were very large, and these cases therefore had to be treated specially.

Finally, a few words are in order about the presentation of results. Clearly, not every estimator is parametrized by modeled outlier variance and modeled fraction of contamination. In particular, the Kalman Filter depends on neither, the Guttman-Peña estimator depends on both, the Ershov-Lipster estimator depends only on $\tilde{R}_{out}$, and the Masreliez-Martin and First-Order estimators depend only on $\tilde{\varepsilon}$. To make the results easier to compare, however, the tables are organized so that an entry appears for each estimator and each pair $\{ \tilde{R}_{out}, \tilde{\varepsilon} \}$.

The Nominal Case (Pure Gaussian noise). To measure the loss of efficiency relative to the optimal estimator (the Kalman Filter) under nominal conditions (no outliers), two sets of simulations were run, for $F = 0.1$ and $F = 0.5$, respectively. Note that although the true fraction of contamination is $\varepsilon = 0$, different values are used in the estimators for the modeled parameters $\tilde{R}_{out}$ and $\tilde{\varepsilon}$. The overall mean squared estimation errors are given in Tables 5.1-2. It is easy to see that the Guttman-Peña estimator is very close to the optimal performance of the Kalman Filter for small $\tilde{R}_{out}$ and $\tilde{\varepsilon}$, as expected; however, its MSE increases with both $\tilde{R}_{out}$ and $\tilde{\varepsilon}$. The Masreliez-Martin estimator has a slightly higher MSE than the First-Order estimator, and the difference between the two increases with $\tilde{\varepsilon}$. It is also noteworthy that the MSE increases in all cases with the value of F, due to the "memory" inherent in slower dynamics. The mean squared estimation error at each n for each estimator is plotted

Table 5.1 Mean Squared Estimation Error (Gaussian, $F = 0.1$)

Kalman Filter

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.4993	0.4993	0.4993
$\tilde{R}_{out} = 6.25$	0.4993	0.4993	0.4993
$\tilde{R}_{out} = 9$	0.4993	0.4993	0.4993

Guttman-Peña

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5010	0.5094	0.5217
$\tilde{R}_{out} = 6.25$	0.5040	0.5241	0.5484
$\tilde{R}_{out} = 9$	0.5082	0.5407	0.5757

Ershov-Lipster

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5482	0.5482	0.5482
$\tilde{R}_{out} = 6.25$	0.5708	0.5708	0.5708
$\tilde{R}_{out} = 9$	0.5867	0.5867	0.5867

Masreliez-Martin

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5067	0.5298	0.5552
$\tilde{R}_{out} = 6.25$	0.5067	0.5298	0.5552
$\tilde{R}_{out} = 9$	0.5067	0.5298	0.5552

First-Order

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5066	0.5276	0.5491
$\tilde{R}_{out} = 6.25$	0.5066	0.5276	0.5491
$\tilde{R}_{out} = 9$	0.5066	0.5276	0.5491

Table 5.2 Mean Squared Estimation Error (Gaussian, $F = 0.5$)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5302	0.5302	0.5302
$\tilde{R}_{out} = 6.25$	0.5302	0.5302	0.5302
$\tilde{R}_{out} = 9$	0.5302	0.5302	0.5302
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5319	0.5409	0.5543
$\tilde{R}_{out} = 6.25$	0.5354	0.5577	0.5855
$\tilde{R}_{out} = 9$	0.5403	0.5774	0.6193
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5879	0.5879	0.5879
$\tilde{R}_{out} = 6.25$	0.6186	0.6186	0.6186
$\tilde{R}_{out} = 9$	0.6404	0.6404	0.6404
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5390	0.5670	0.5985
$\tilde{R}_{out} = 6.25$	0.5390	0.5670	0.5985
$\tilde{R}_{out} = 9$	0.5390	0.5670	0.5985
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5381	0.5599	0.5830
$\tilde{R}_{out} = 6.25$	0.5381	0.5599	0.5830
$\tilde{R}_{out} = 9$	0.5381	0.5599	0.5830

in Figures 5.1-8. These plots clearly illustrate the fact that the loss of efficiency of the First-Order estimator at the nominal model is minimal for small $\tilde{\varepsilon}$ (Figures 5.1, 5.3, 5.5, and 5.7) and that it favorably compares with the other estimators for large $\tilde{R}_{out}$ (Figures 5.2, 5.4, 5.6, and 5.8). In particular, Figures 5.4 and 5.8 show that its behavior is the closest of all robust estimators to the optimal (the Kalman Filter). The relationship between the parameters $\tilde{R}_{out}$ and $\tilde{\varepsilon}$ and the robustness of these estimators will become clear when simulations with heavy-tailed observation noise distributions are reviewed.

Another measure of the performance of the estimators is the whiteness of the residual sequence. The lag-one serial correlations of the residuals for each estimator are given in Tables 5.3-4, and confirm the findings outlined above: the Guttman-Peña estimator behaves nearly optimally for small $\tilde{R}_{out}$ and $\tilde{\varepsilon}$, while the First-Order estimator (and also the Ershov-Lipster estimator) perform very well for large $\tilde{R}_{out}$ and $\tilde{\varepsilon}$. Lag-one serial correlations for each n are plotted in Figures 5.9-12.

Finally, the mean squared estimation error normalized by the estimated covariance is computed in an effort to determine the accuracy of the second moment estimate. Clearly, perfect covariance estimates would yield mean squared errors near unity, and deviations in either direction indicate under- or over-estimation of the estimation error covariance. The results are presented in Tables 5.5-6, and plotted in Figures 5.13-16. As before, the First-Order estimator performs best for large $\tilde{R}_{out}$ and $\tilde{\varepsilon}$, while the Guttman-Peña estimator performs best for small $\tilde{R}_{out}$ and $\tilde{\varepsilon}$.

In most of the experiments discussed so far, the Ershov-Lipster estimator did not perform as well as the others; it must be remembered, however, that different outlier tests and different significance levels might yield better performance. In addition, the Masreliez-Martin estimator did not perform as well as the First-Order estimator at the nominal model. While this behavior would be expected if the assumed distributional model was identical to the true observation noise distribution, it need not hold when the two are different, as some examples in the sequel demonstrate.

Scale-Contaminated Gaussian Noise. To assess the performance of the various estimators in the presence of outliers distributed according to a Gaussian distribution with larger variance than the nominal model, simulation experiments were performed with nominal Gaussian observation noise except at $n = 20$, where the noise was Gaussian with variance $R_{out} = \tilde{R}_{out}$. The mean squared estimation error at times $n = 18, 19, \cdots, 28$ are plotted in Figures 5.17-20, while the MSE at $n = 20$ appears in Table 5.7. As expected, the Kalman Filter has the best performance except when affected by the outlier. While the Masreliez-Martin and First-Order estimators are virtually indistinguishable for small $\tilde{\varepsilon}$, the latter performs better in the aftermath of an outlier for large $\tilde{\varepsilon}$. This is a consequence of the "smoother" correction terms in Theorem 4.3. The Guttman-Peña and Ershov-Lipster estimators perform comparably to the Masreliez-Martin and First-Order estimators at the exact time of the outlier, but their performance is considerably worse right after the outlier in the case of large $\tilde{R}_{out}$ and $\tilde{\varepsilon}$.

The lag-one serial correlations in this case do not show a great difference among the robust estimators. The normalized mean squared estimation errors also show comparable performance among the estimators, which all tend to somewhat underestimate the covariance at the time of the outlier, but recover within a couple of time steps. One example appears in Figure 5.21.

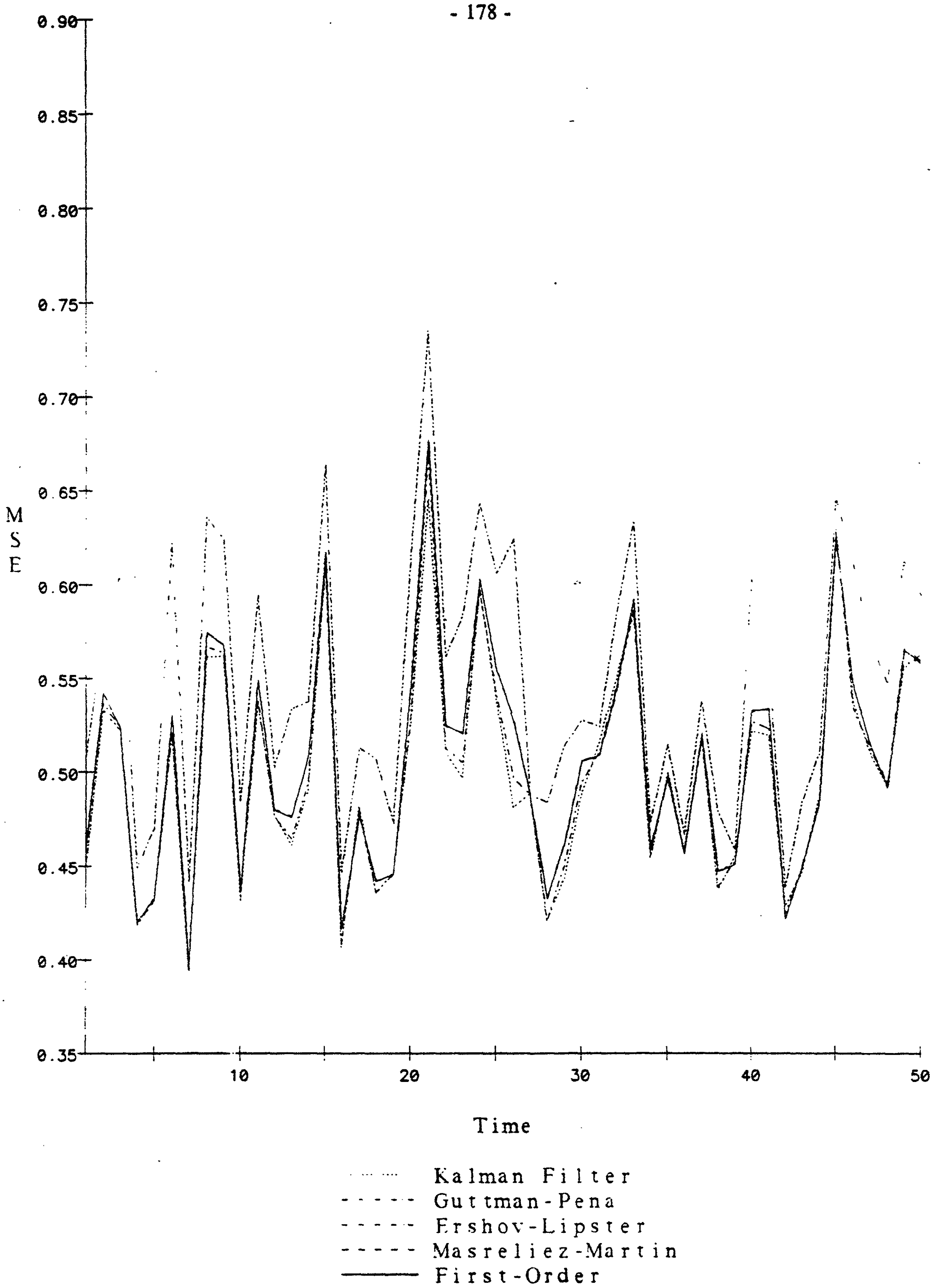

Figure 5.1 Mean Squared Estimation Error ($\bar{R}_{out}$ = 4, $\bar{\varepsilon}$ = 0.01, F = 0.1, Gaussian)

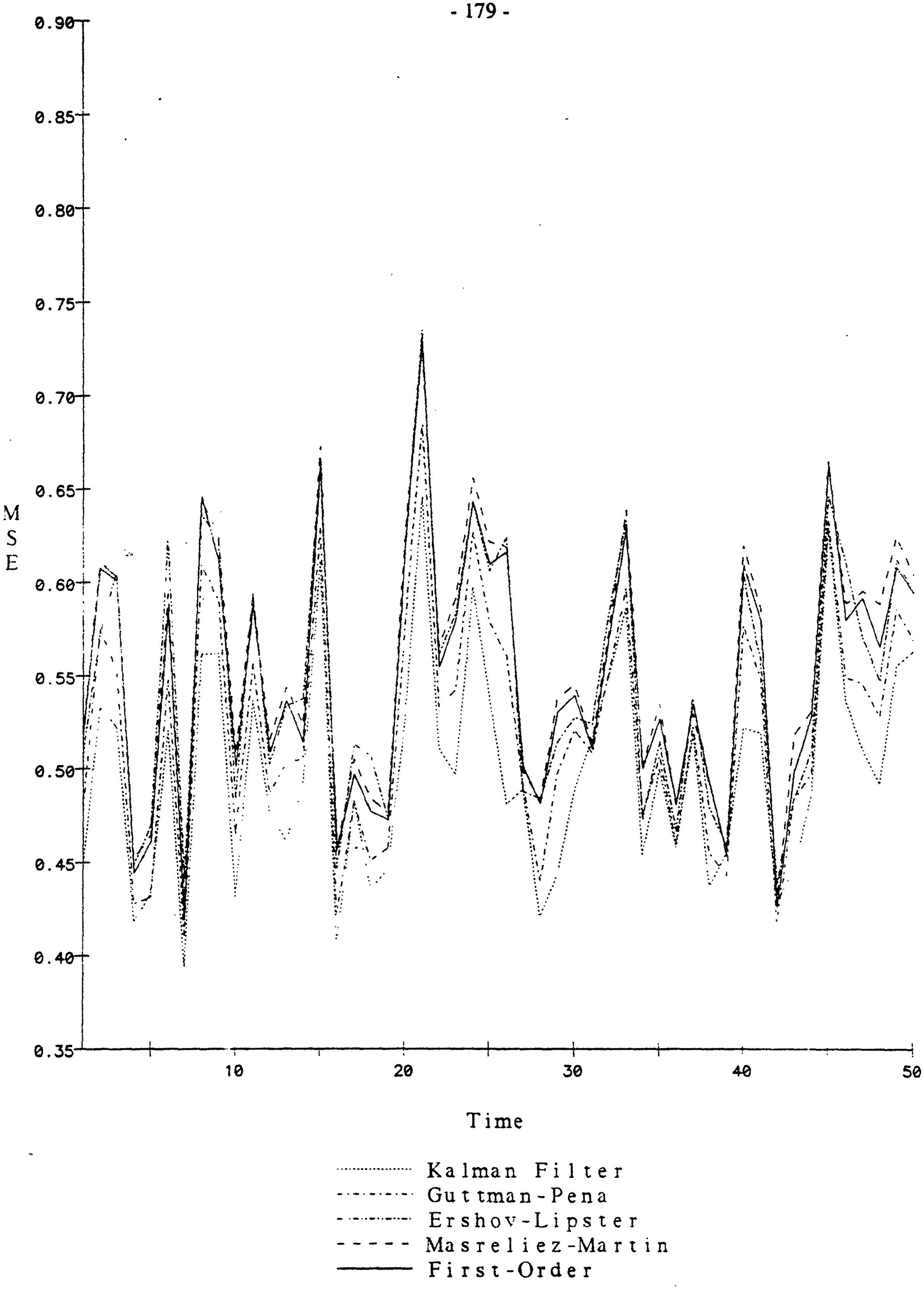

Figure 5.2 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.10, F = 0.1, Gaussian)

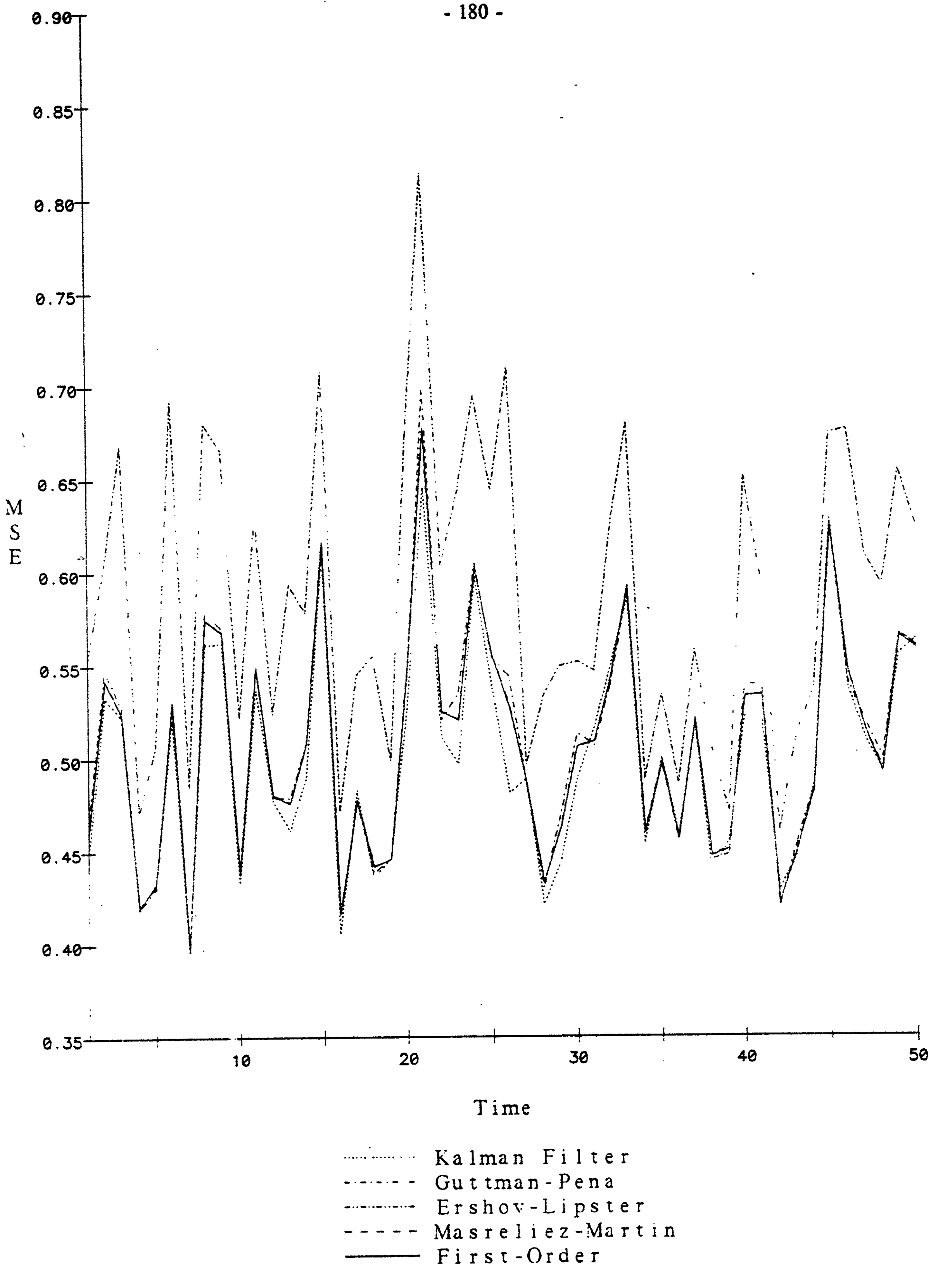

Figure 5.3 Mean Squared Estimation Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.01, F = 0.1, Gaussian)

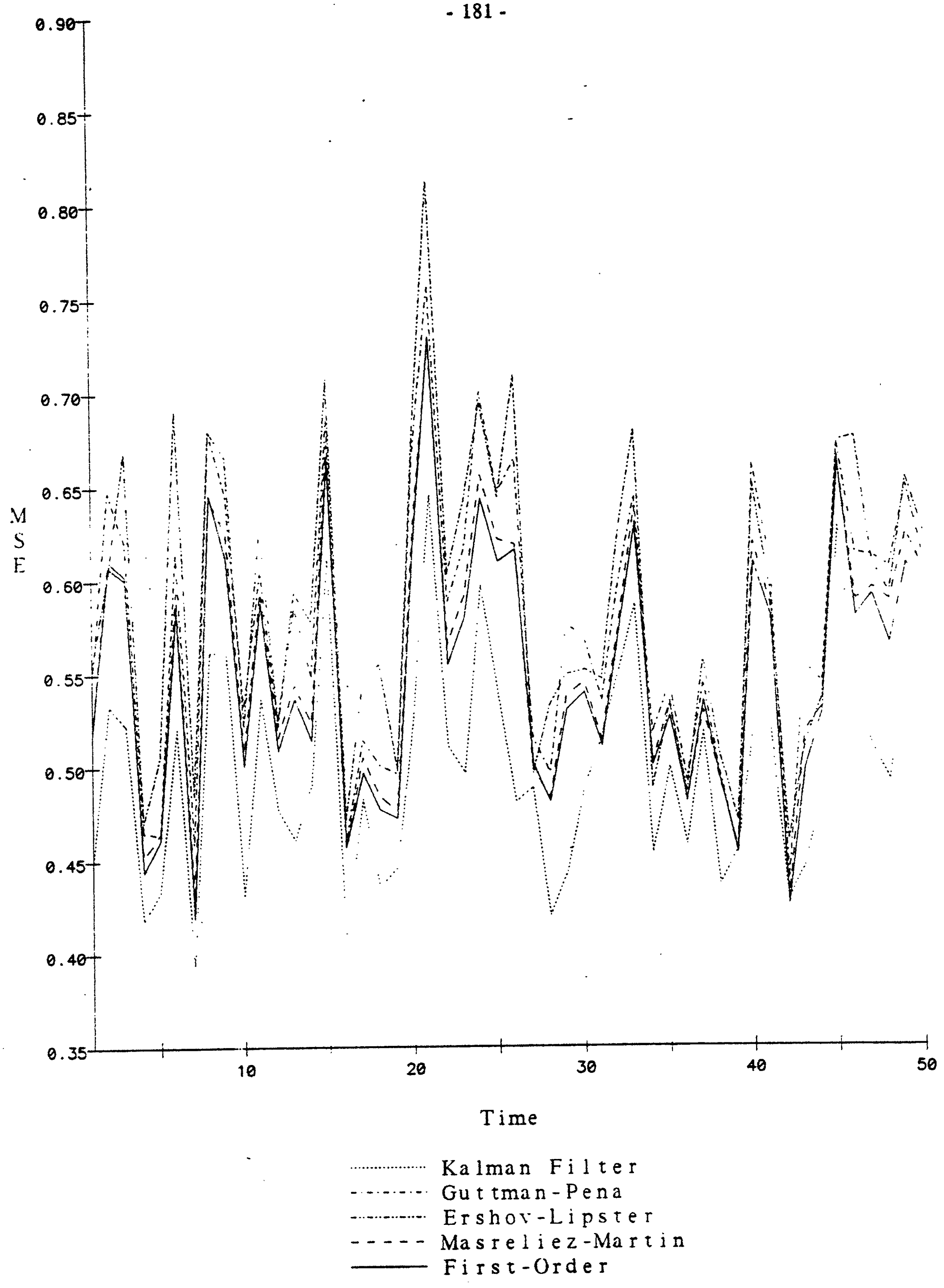

Figure 5.4 Mean Squared Estimation Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, F = 0.1, Gaussian)

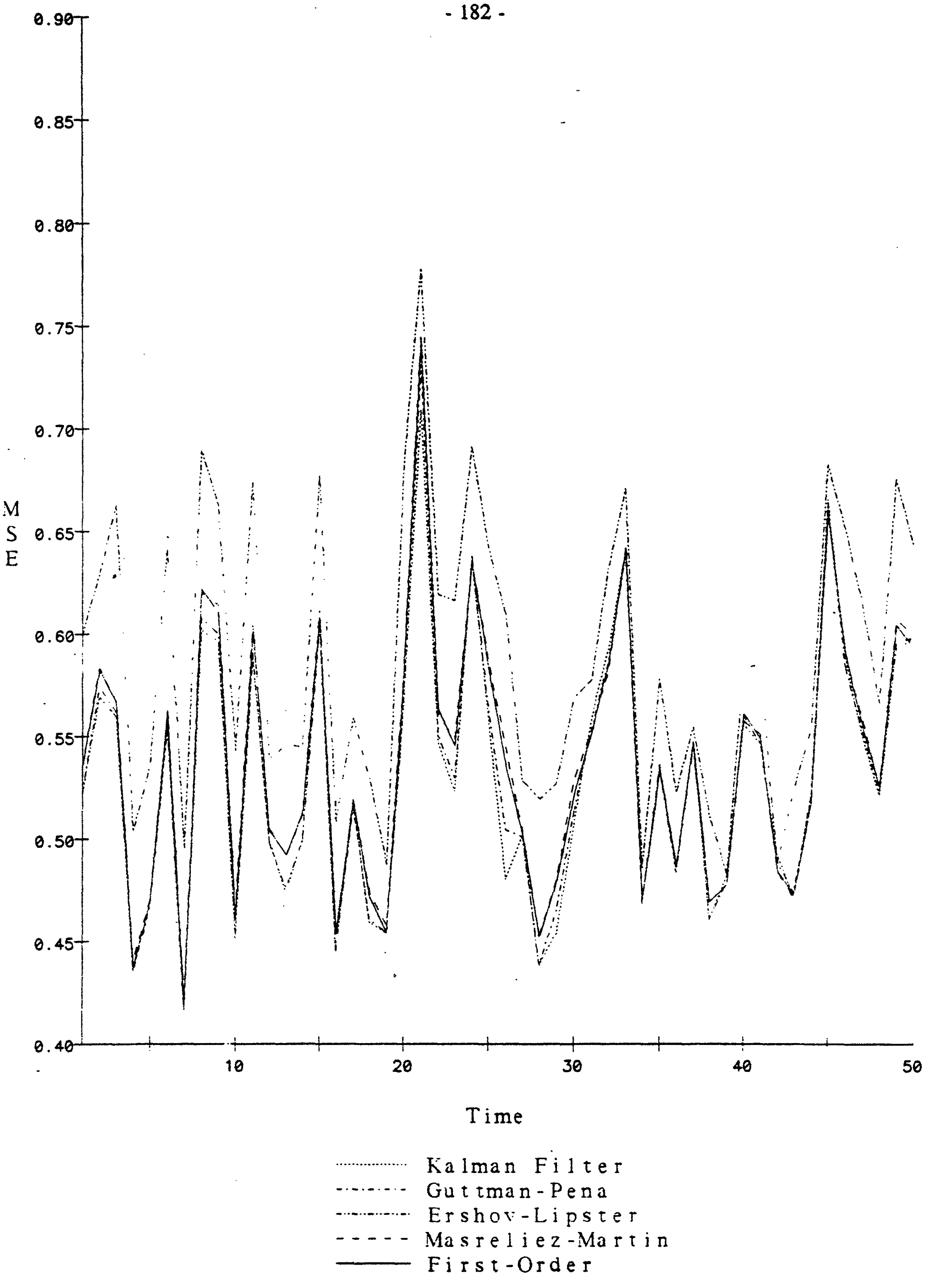

Figure 5.5 Mean Squared Estimation Error ($\bar{R}_{out} = 4$, $\bar{\varepsilon} = 0.01$, $F = 0.5$, Gaussian)

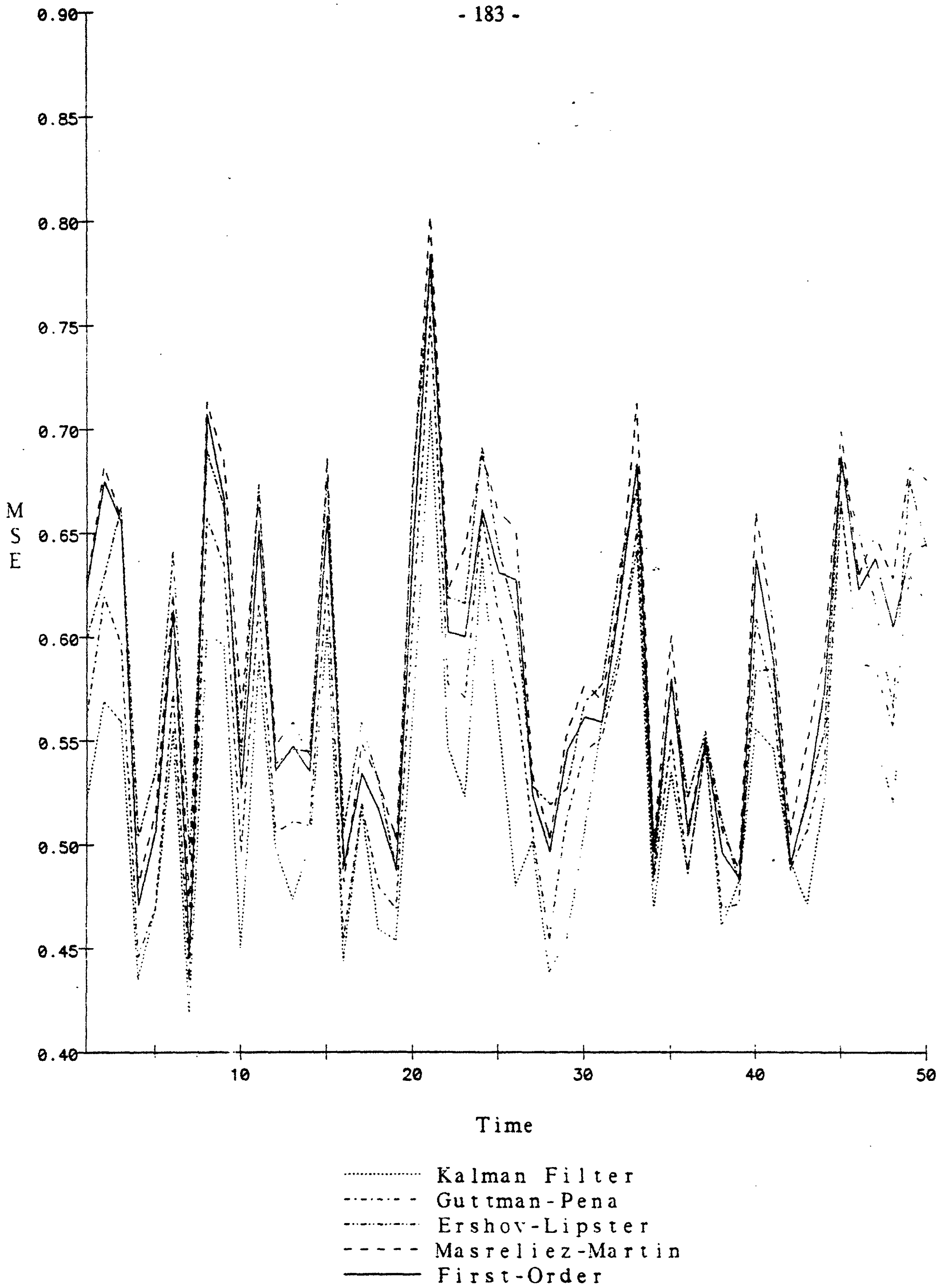

Figure 5.6 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\bar{\varepsilon}$ = 0.10, F = 0.5, Gaussian)

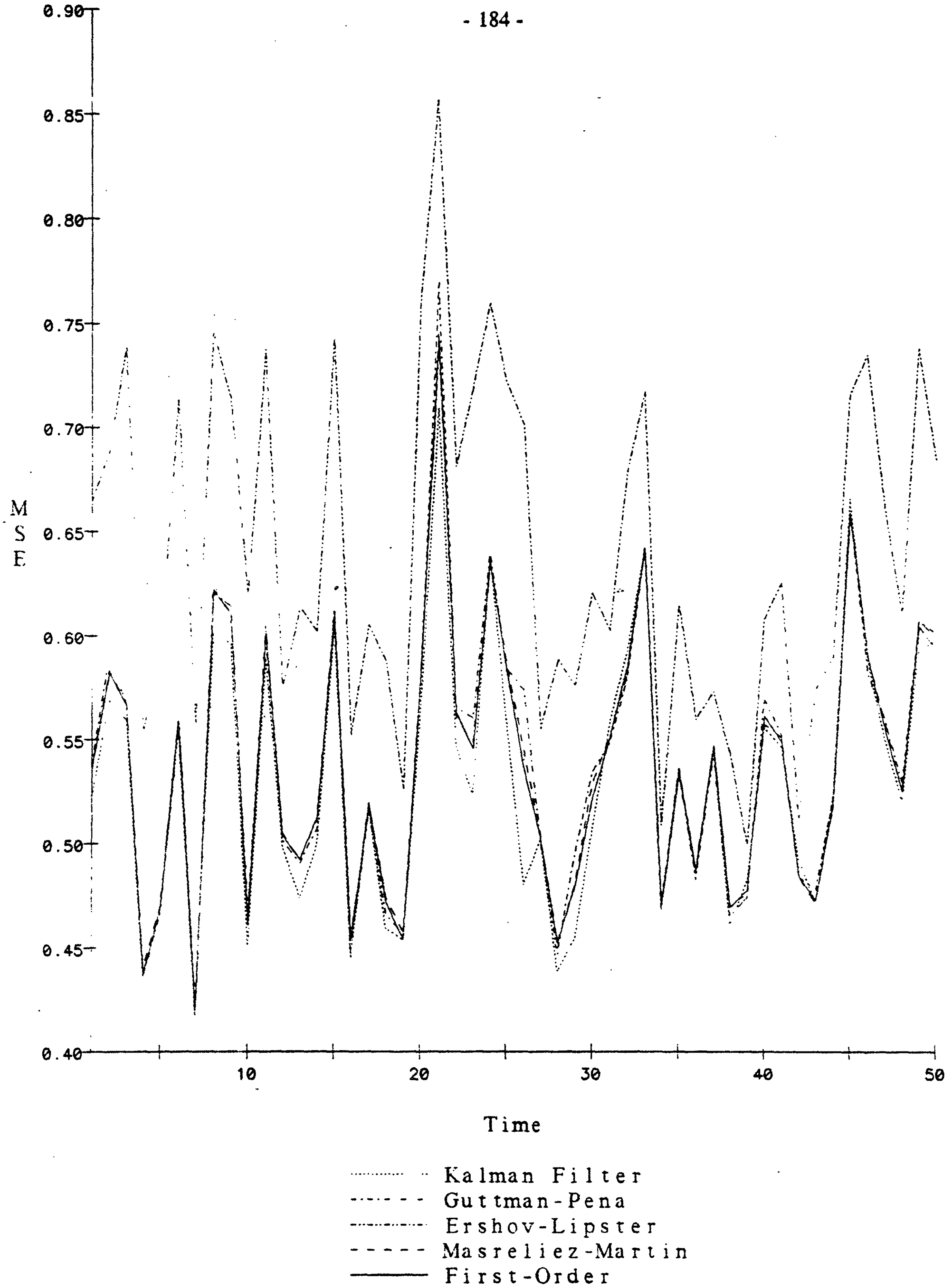

Figure 5.7 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.01, F = 0.5, Gaussian)

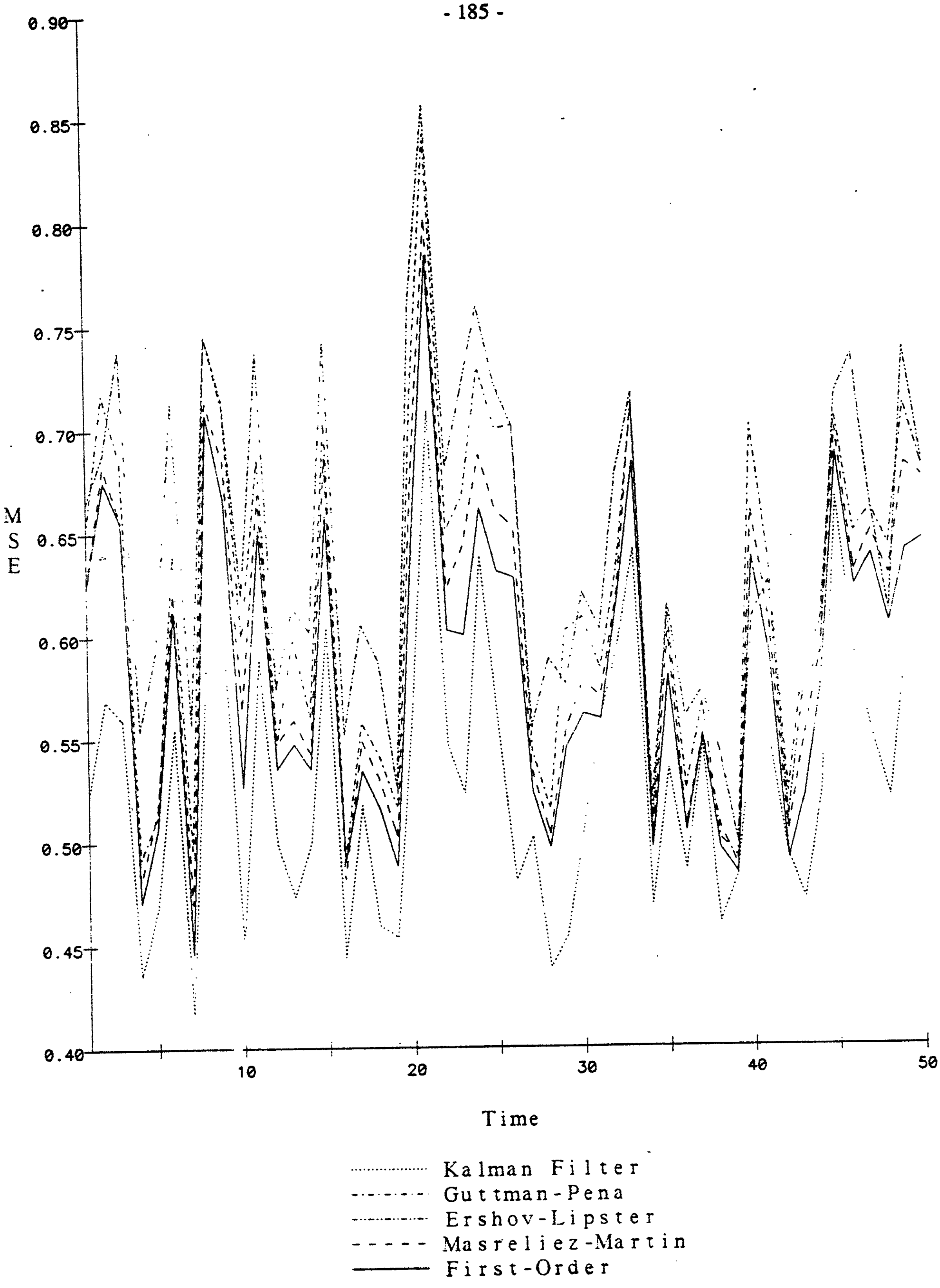

Figure 5.8 Mean Squared Estimation Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, F = 0.5, Gaussian)

Table 5.3 Lag-One Serial Correlation (Gaussian, F = 0.1)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0057	0.0057	0.0057
$\tilde{R}_{out} = 6.25$	0.0057	0.0057	0.0057
$\tilde{R}_{out} = 9$	0.0057	0.0057	0.0057
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0072	0.0112	0.0148
$\tilde{R}_{out} = 6.25$	0.0084	0.0145	0.0193
$\tilde{R}_{out} = 9$	0.0095	0.0173	0.0229
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0134	0.0134	0.0134
$\tilde{R}_{out} = 6.25$	0.0150	0.0150	0.0150
$\tilde{R}_{out} = 9$	0.0160	0.0160	0.0160
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0082	0.0134	0.0179
$\tilde{R}_{out} = 6.25$	0.0082	0.0134	0.0179
$\tilde{R}_{out} = 9$	0.0082	0.0134	0.0179
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0081	0.0130	0.0169
$\tilde{R}_{out} = 6.25$	0.0081	0.0130	0.0169
$\tilde{R}_{out} = 9$	0.0081	0.0130	0.0169

Table 5.4 Lag-One Serial Correlation (Gaussian, $F = 0.5$)

Kalman Filter

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0078	0.0078	0.0078
$\tilde{R}_{out} = 6.25$	0.0078	0.0078	0.0078
$\tilde{R}_{out} = 9$	0.0078	0.0078	0.0078

Guttman-Peña

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0147	0.0344	0.0521
$\tilde{R}_{out} = 6.25$	0.0207	0.0512	0.0756
$\tilde{R}_{out} = 9$	0.0267	0.0660	0.0950

Ershov-Lipster

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0501	0.0501	0.0501
$\tilde{R}_{out} = 6.25$	0.0601	0.0601	0.0601
$\tilde{R}_{out} = 9$	0.0662	0.0662	0.0662

Masreliez-Martin

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0202	0.0468	0.0704
$\tilde{R}_{out} = 6.25$	0.0202	0.0468	0.0704
$\tilde{R}_{out} = 9$	0.0202	0.0468	0.0704

First-Order

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.0199	0.0422	0.0612
$\tilde{R}_{out} = 6.25$	0.0199	0.0422	0.0612
$\tilde{R}_{out} = 9$	0.0199	0.0422	0.0612

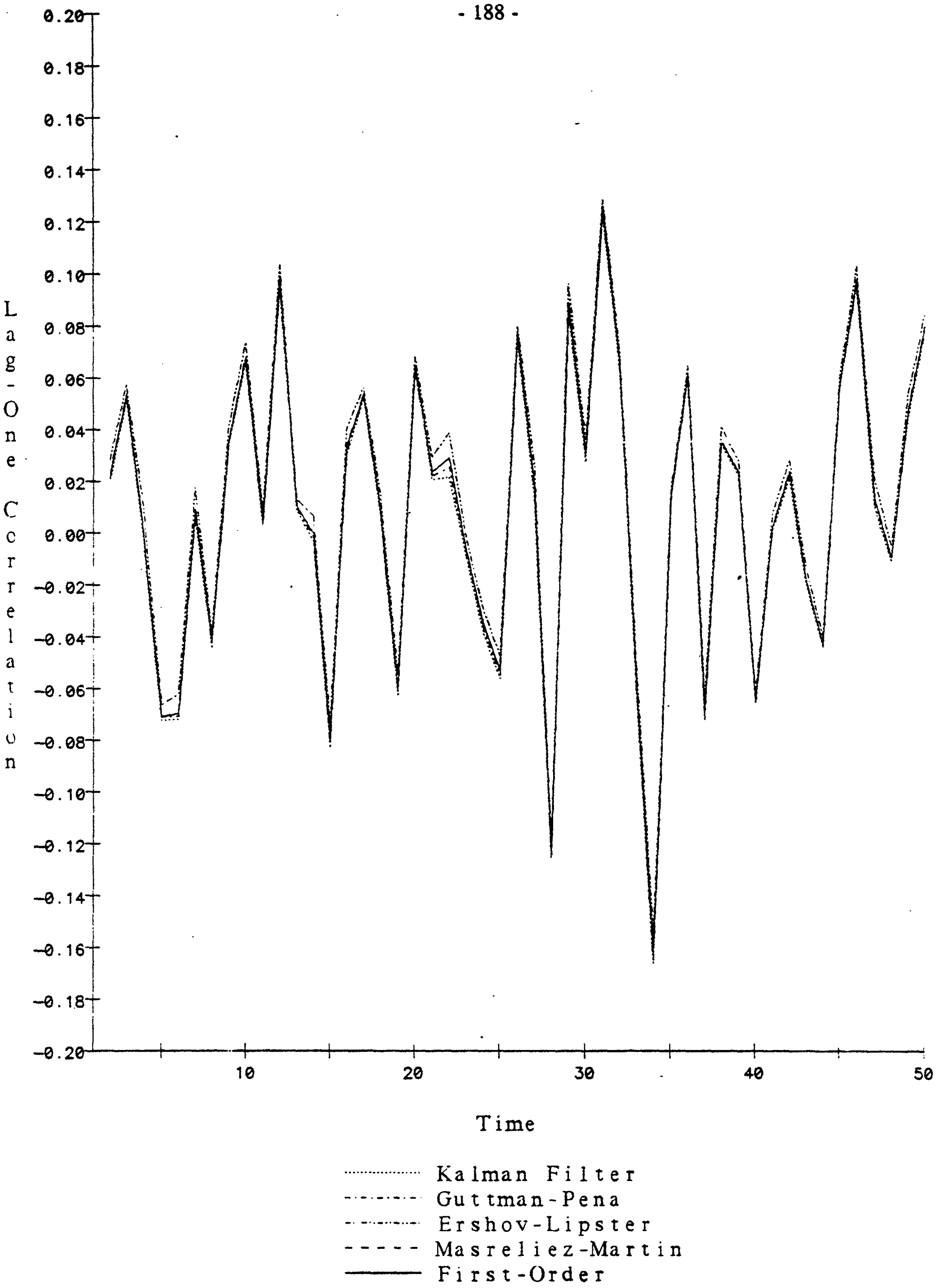

Figure 5.9 Lag-One Serial Correlation ($\tilde{R}_{out}$ = 4, $\bar{\varepsilon}$ = 0.01, F = 0.1, Gaussian)

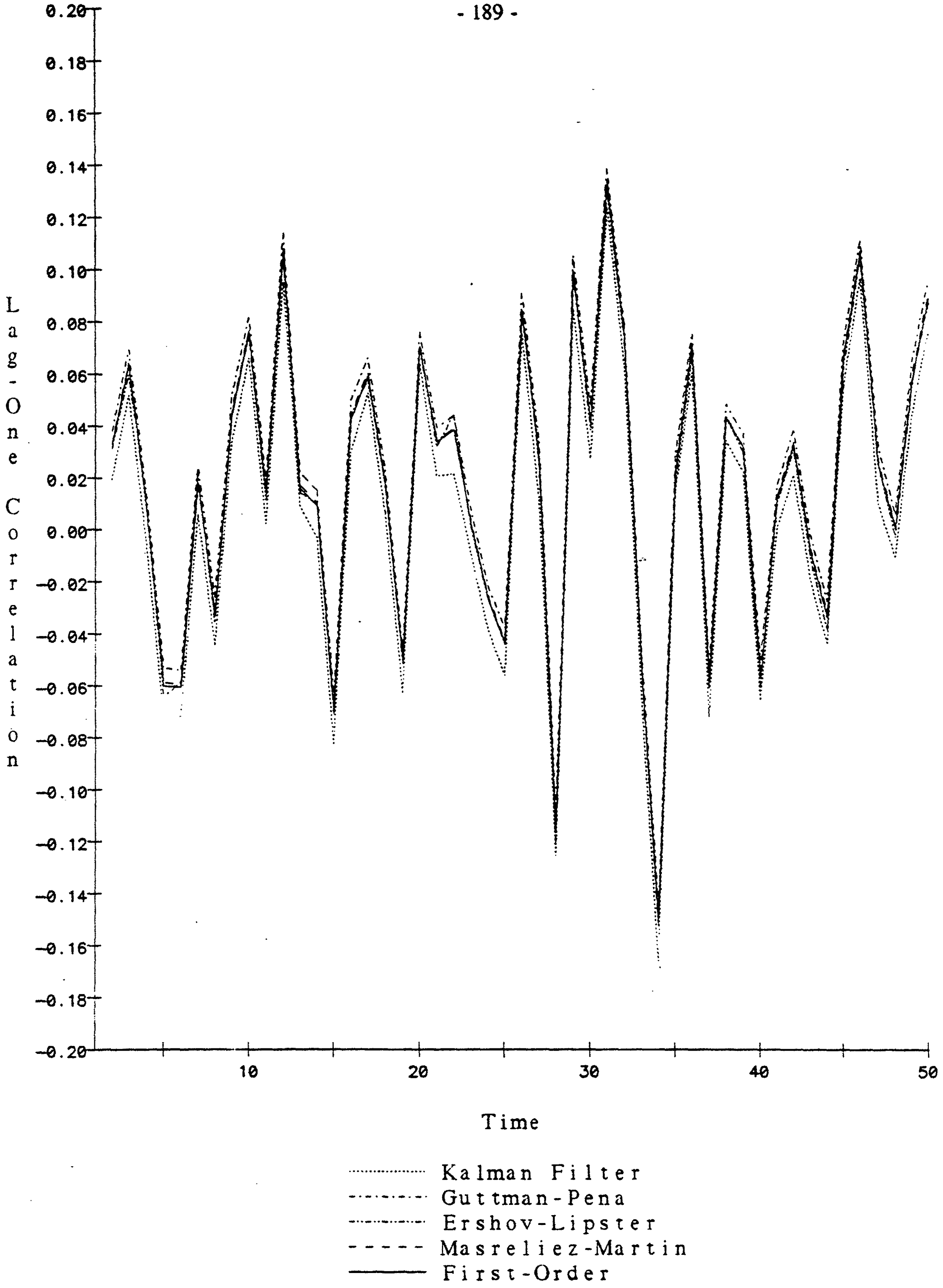

Figure 5.10 Lag-One Serial Correlation ($\bar{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.10, F = 0.1, Gaussian)

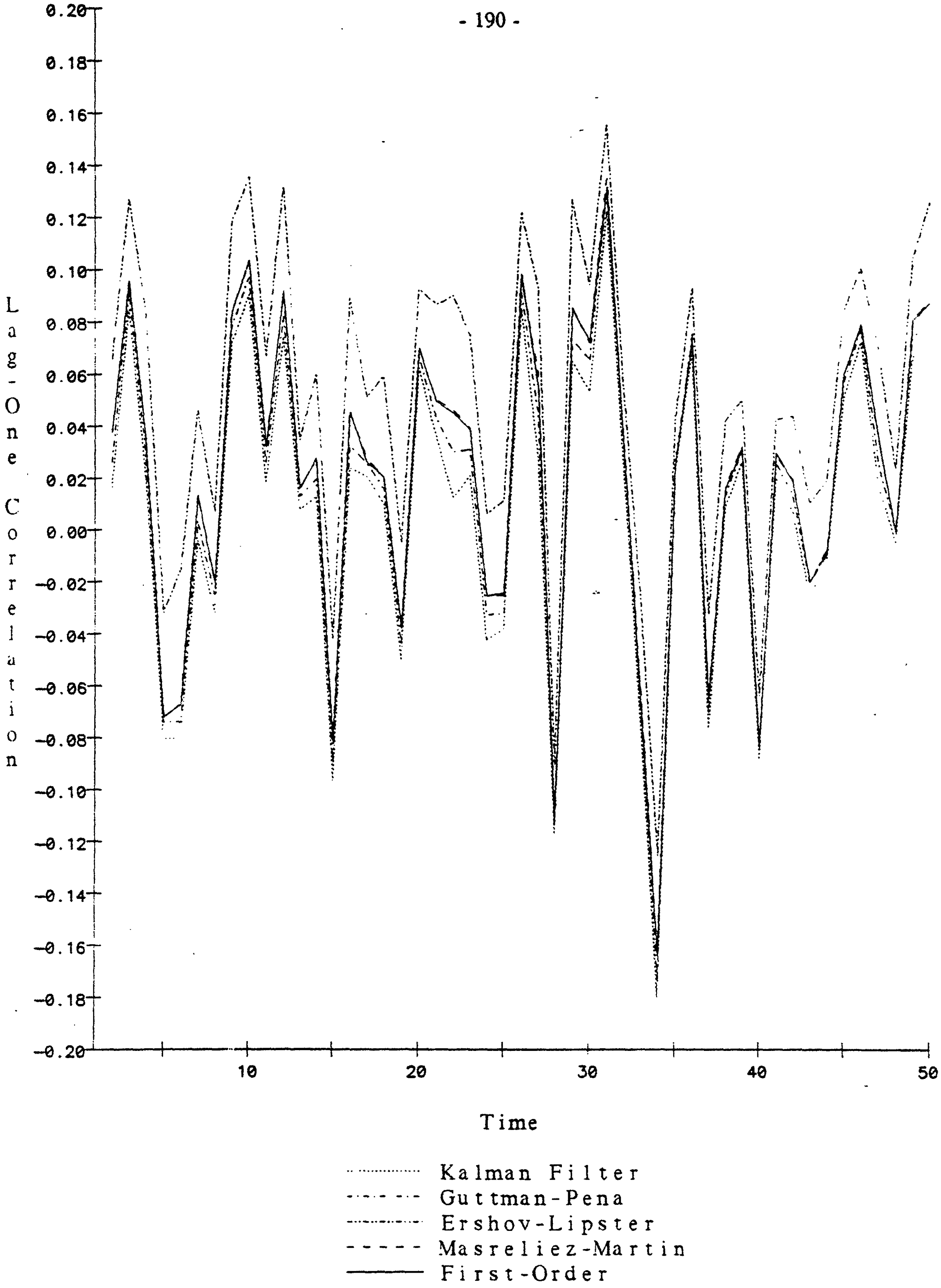

Figure 5.11 Lag-One Serial Correlation ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.01, F = 0.5, Gaussian)

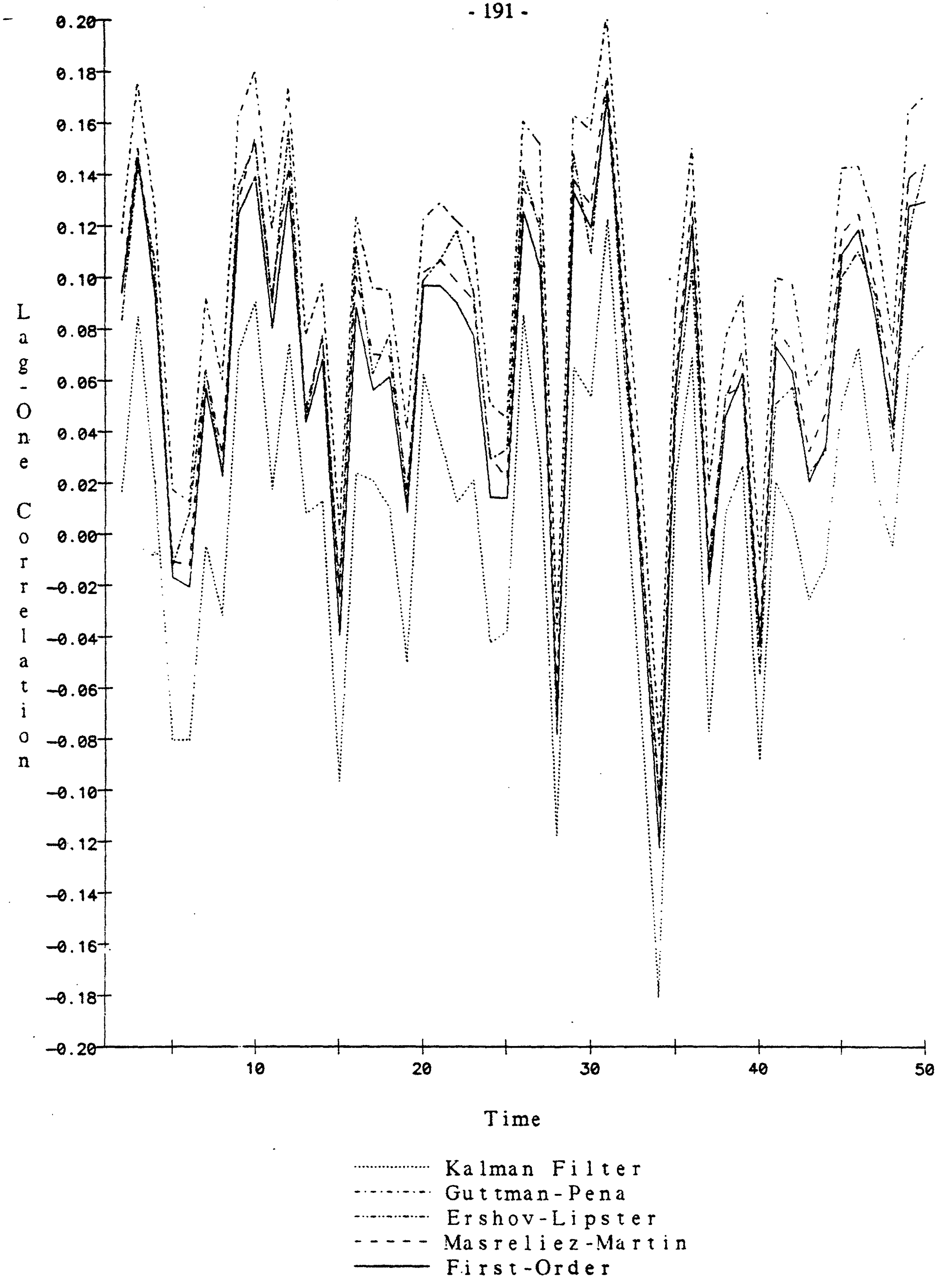

Figure 5.12 Lag-One Serial Correlation ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.10, F = 0.5, Gaussian)

Table 5.5 Normalized Mean Squared Estimation Error (Gaussian, F = 0.1)

Kalman Filter

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9961	0.9961	0.9961
$\tilde{R}_{out} = 6.25$	0.9961	0.9961	0.9961
$\tilde{R}_{out} = 9$	0.9961	0.9961	0.9961

Guttman-Peña

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9846	0.9517	0.9249
$\tilde{R}_{out} = 6.25$	0.9808	0.9443	0.9190
$\tilde{R}_{out} = 9$	0.9787	0.9419	0.9196

Ershov-Lipster

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.0390	1.0390	1.0390
$\tilde{R}_{out} = 6.25$	1.0586	1.0586	1.0586
$\tilde{R}_{out} = 9$	1.0715	1.0715	1.0715

Masreliez-Martin

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9789	0.9411	0.9204
$\tilde{R}_{out} = 6.25$	0.9789	0.9411	0.9204
$\tilde{R}_{out} = 9$	0.9789	0.9411	0.9204

First-Order

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9792	0.9436	0.9240
$\tilde{R}_{out} = 6.25$	0.9792	0.9436	0.9240
$\tilde{R}_{out} = 9$	0.9792	0.9436	0.9240

Table 5.6 Normalized Mean Squared Estimation Error (Gaussian, $F = 0.5$)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9972	0.9972	0.9972
$\tilde{R}_{out} = 6.25$	0.9972	0.9972	0.9972
$\tilde{R}_{out} = 9$	0.9972	0.9972	0.9972
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9839	0.9446	0.9114
$\tilde{R}_{out} = 6.25$	0.9793	0.9338	0.9007
$\tilde{R}_{out} = 9$	0.9763	0.9292	0.8990
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.0417	1.0417	1.0417
$\tilde{R}_{out} = 6.25$	1.0665	1.0665	1.0665
$\tilde{R}_{out} = 9$	1.0829	1.0829	1.0829
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9748	0.9319	0.9014
$\tilde{R}_{out} = 6.25$	0.9748	0.9319	0.9014
$\tilde{R}_{out} = 9$	0.9748	0.9319	0.9014
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9772	0.9401	0.9163
$\tilde{R}_{out} = 6.25$	0.9772	0.9401	0.9163
$\tilde{R}_{out} = 9$	0.9772	0.9401	0.9163

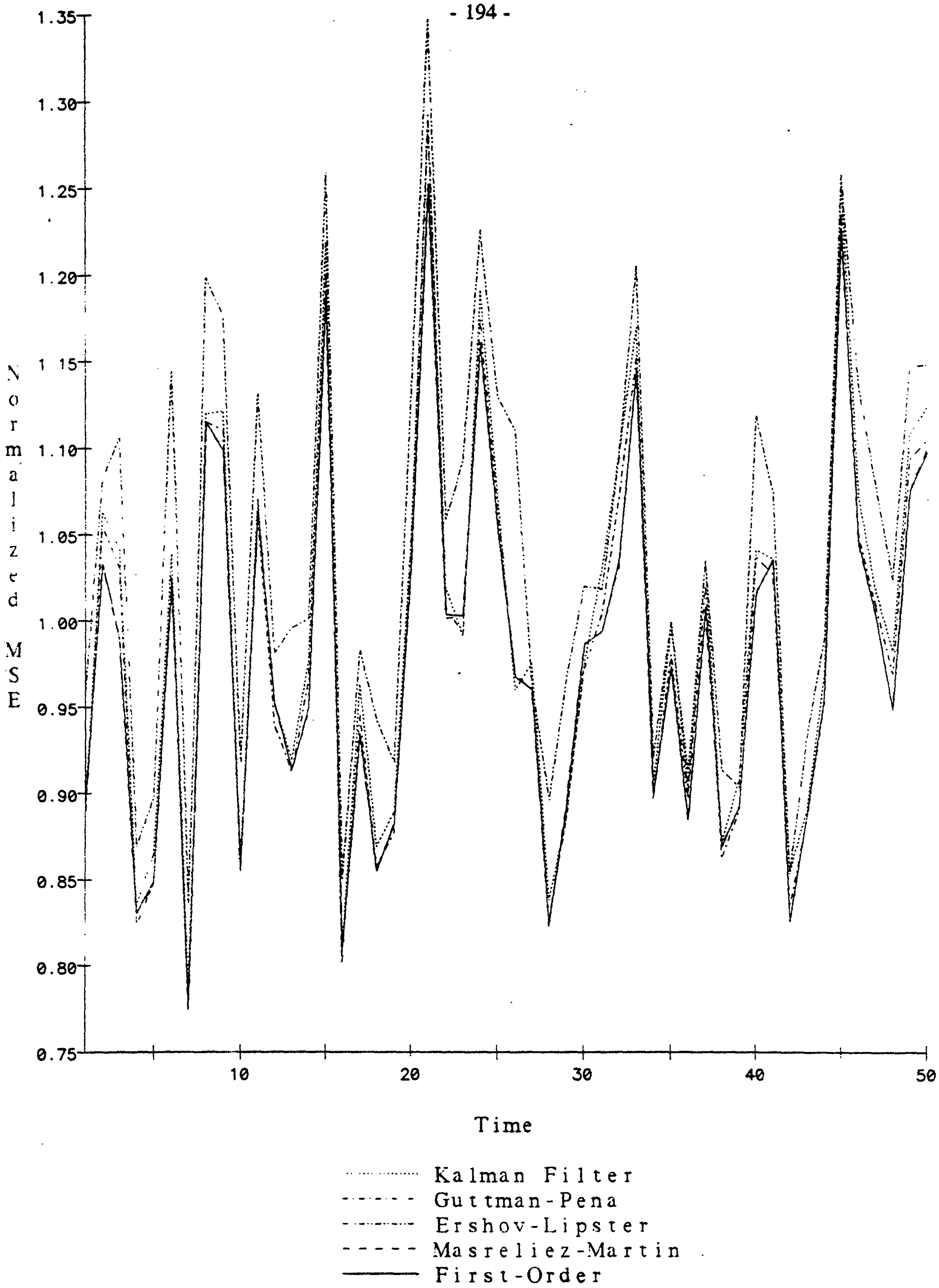

Figure 5.13 Normalized Mean Squared Error ($\tilde{R}_{out}$ = 4, $\bar{\varepsilon}$ = 0.01, F = 0.1, Gaussian)

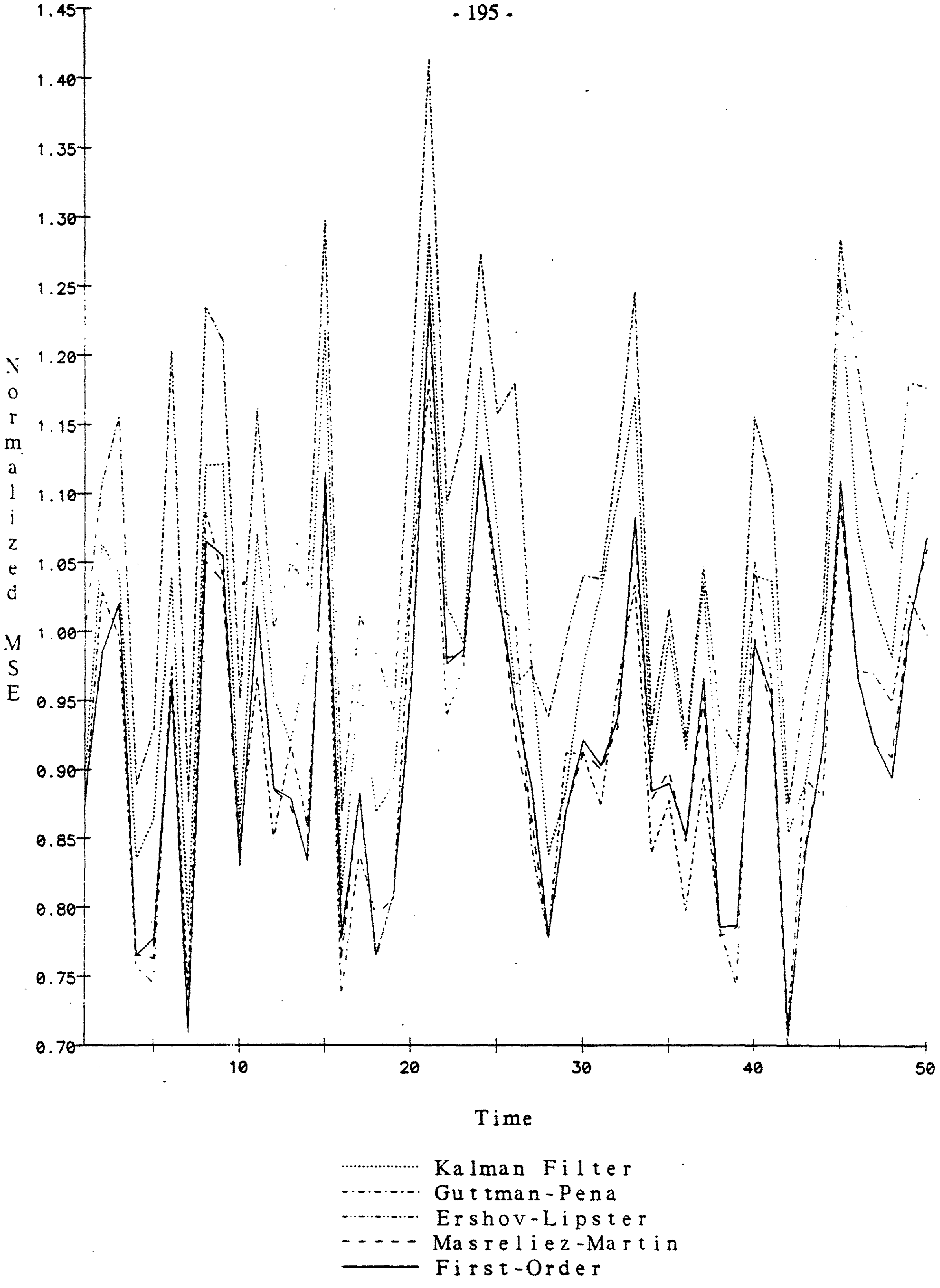

Figure 5.14 Normalized Mean Squared Error ($\tilde{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, F = 0.1, Gaussian)

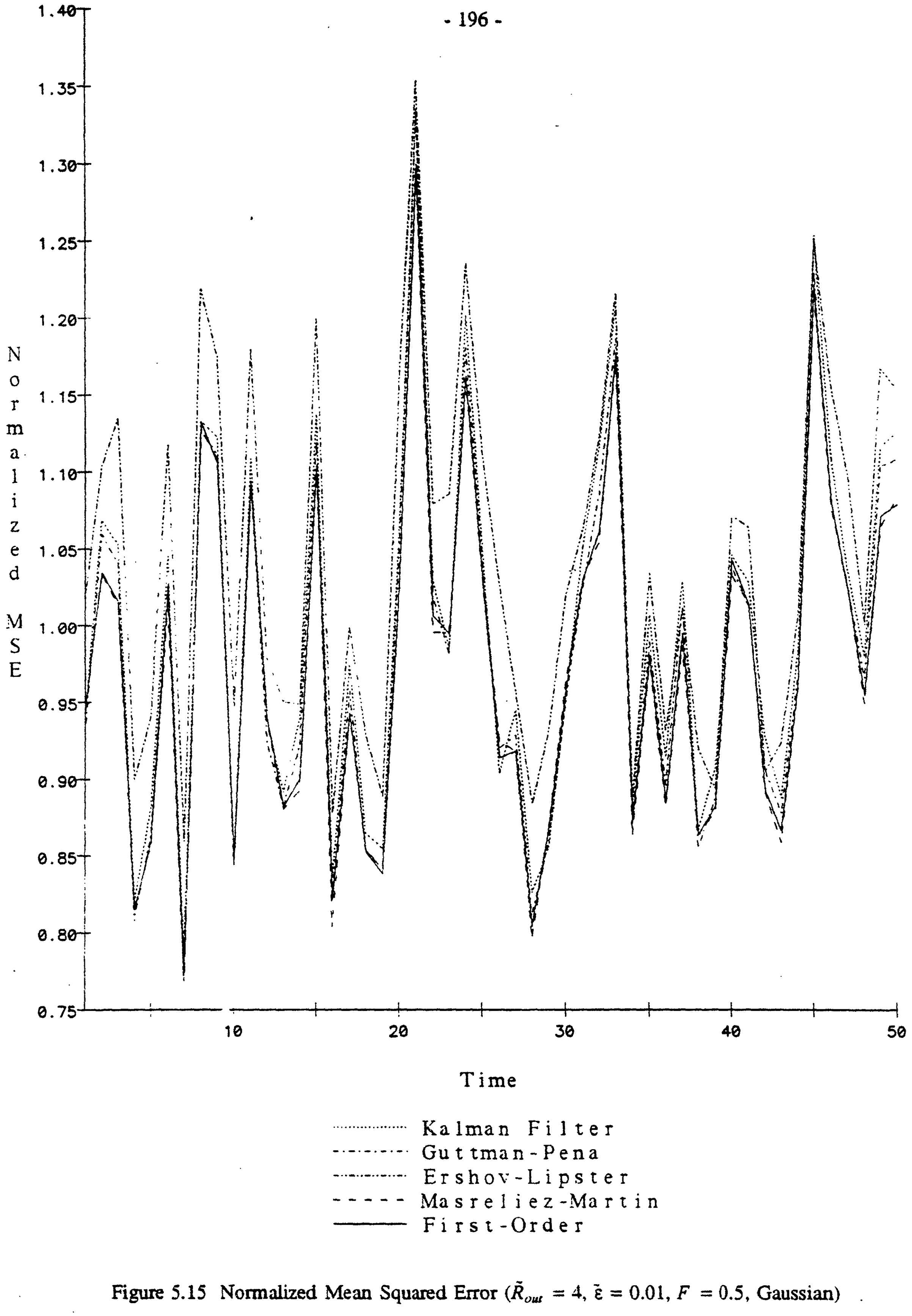

Figure 5.15 Normalized Mean Squared Error ($\tilde{R}_{out}$ = 4, $\bar{\varepsilon}$ = 0.01, F = 0.5, Gaussian)

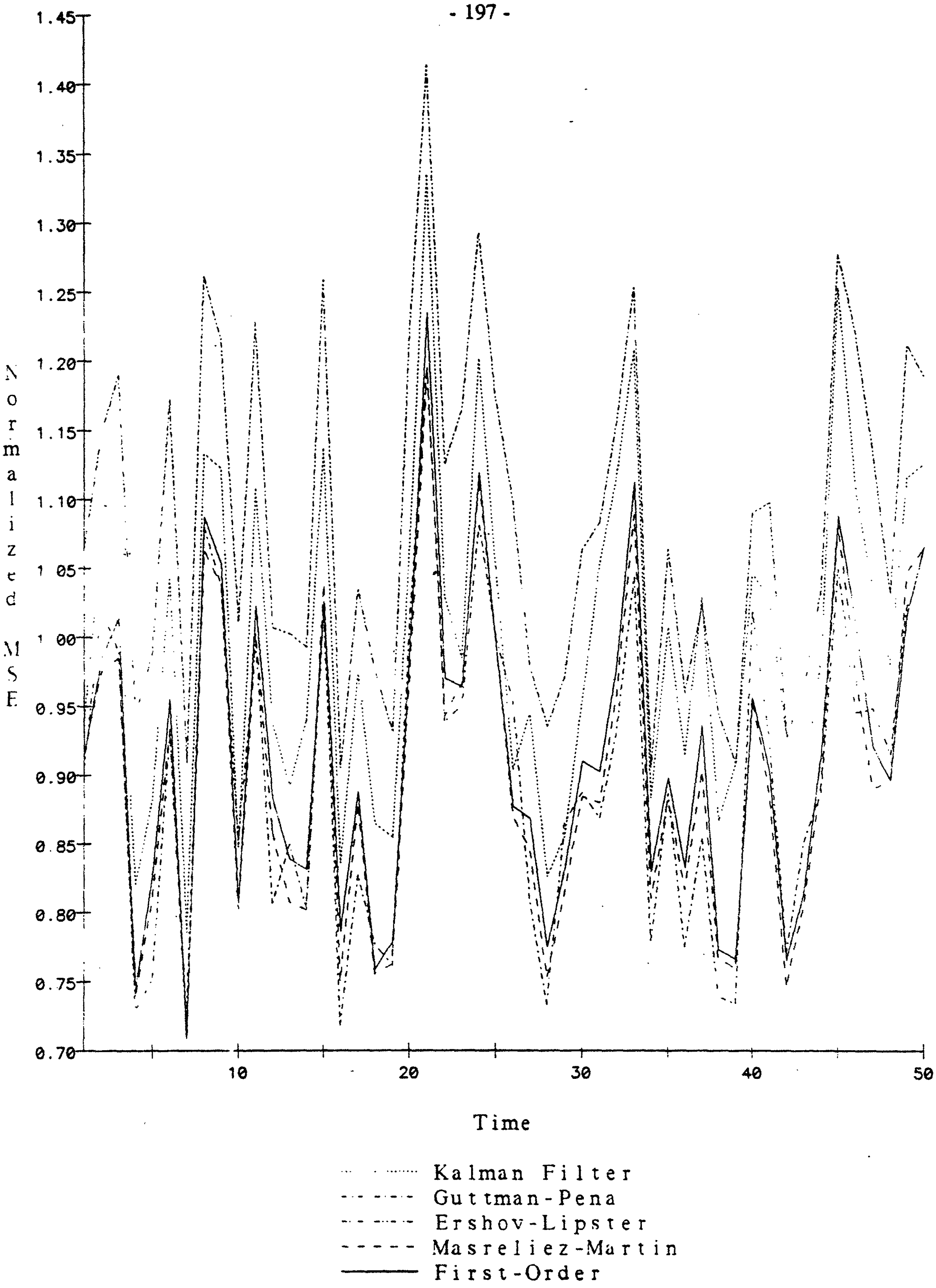

Figure 5.16 Normalized Mean Squared Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, F = 0.5, Gaussian)

Table 5.7 Mean Squared Estimation Error at $n = 20$ (Scale-Contaminated Gaussian)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.2519	1.2519	1.2519
$\tilde{R}_{out} = 6.25$	1.8070	1.8070	1.8070
$\tilde{R}_{out} = 9$	2.4899	2.4899	2.4899
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.1039	0.9881	0.9392
$\tilde{R}_{out} = 6.25$	1.2624	1.0878	1.0283
$\tilde{R}_{out} = 9$	1.3424	1.1448	1.0869
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9564	0.9564	0.9564
$\tilde{R}_{out} = 6.25$	1.0968	1.0968	1.0968
$\tilde{R}_{out} = 9$	1.1993	1.1993	1.1993
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.0208	0.9399	0.9261
$\tilde{R}_{out} = 6.25$	1.2263	1.0887	1.0489
$\tilde{R}_{out} = 9$	1.4176	1.2095	1.1356
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.0239	0.9500	0.9357
$\tilde{R}_{out} = 6.25$	1.2336	1.1023	1.0598
$\tilde{R}_{out} = 9$	1.4261	1.2222	1.1485

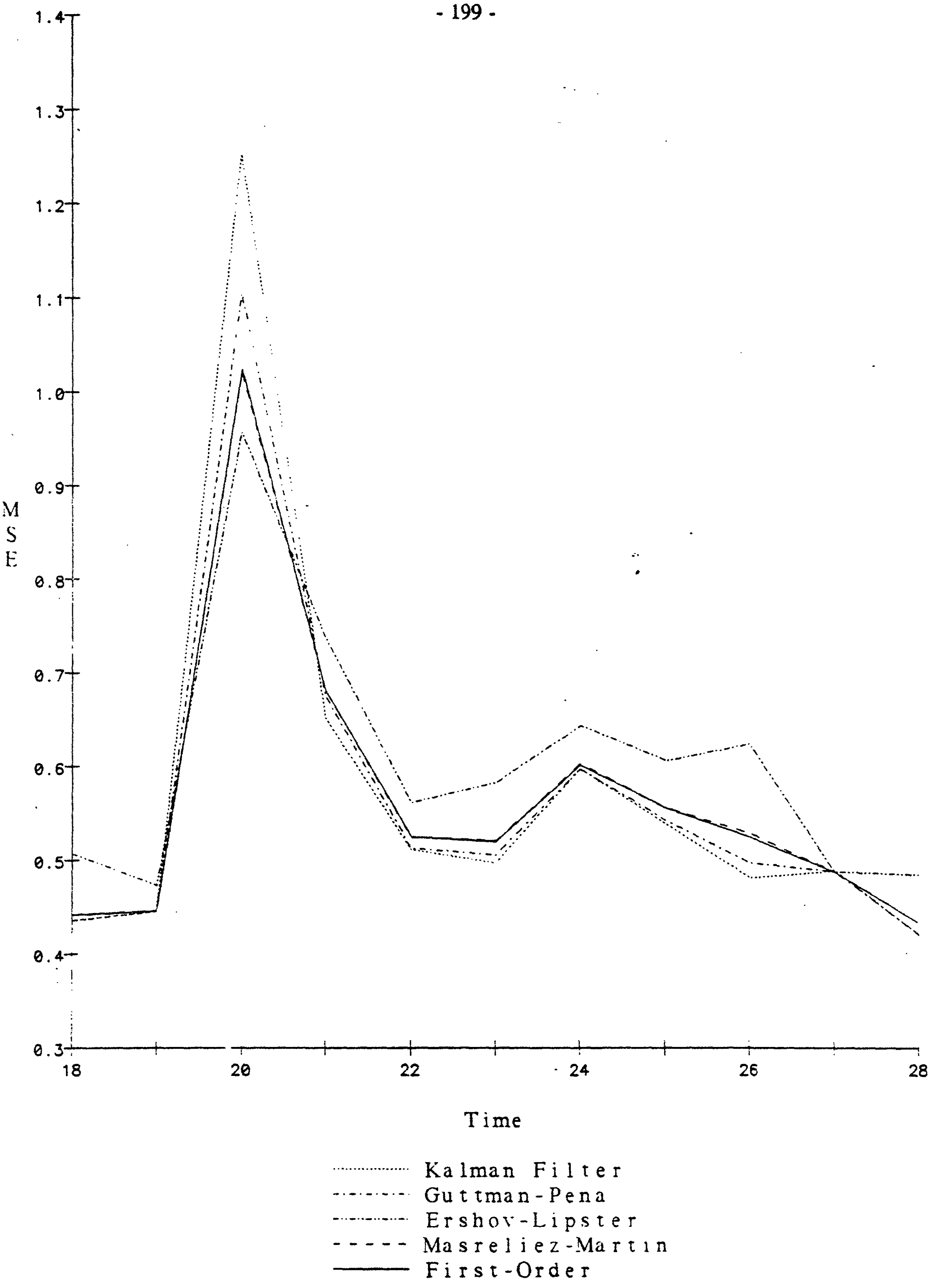

Figure 5.17 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.01, Scale-Contaminated)

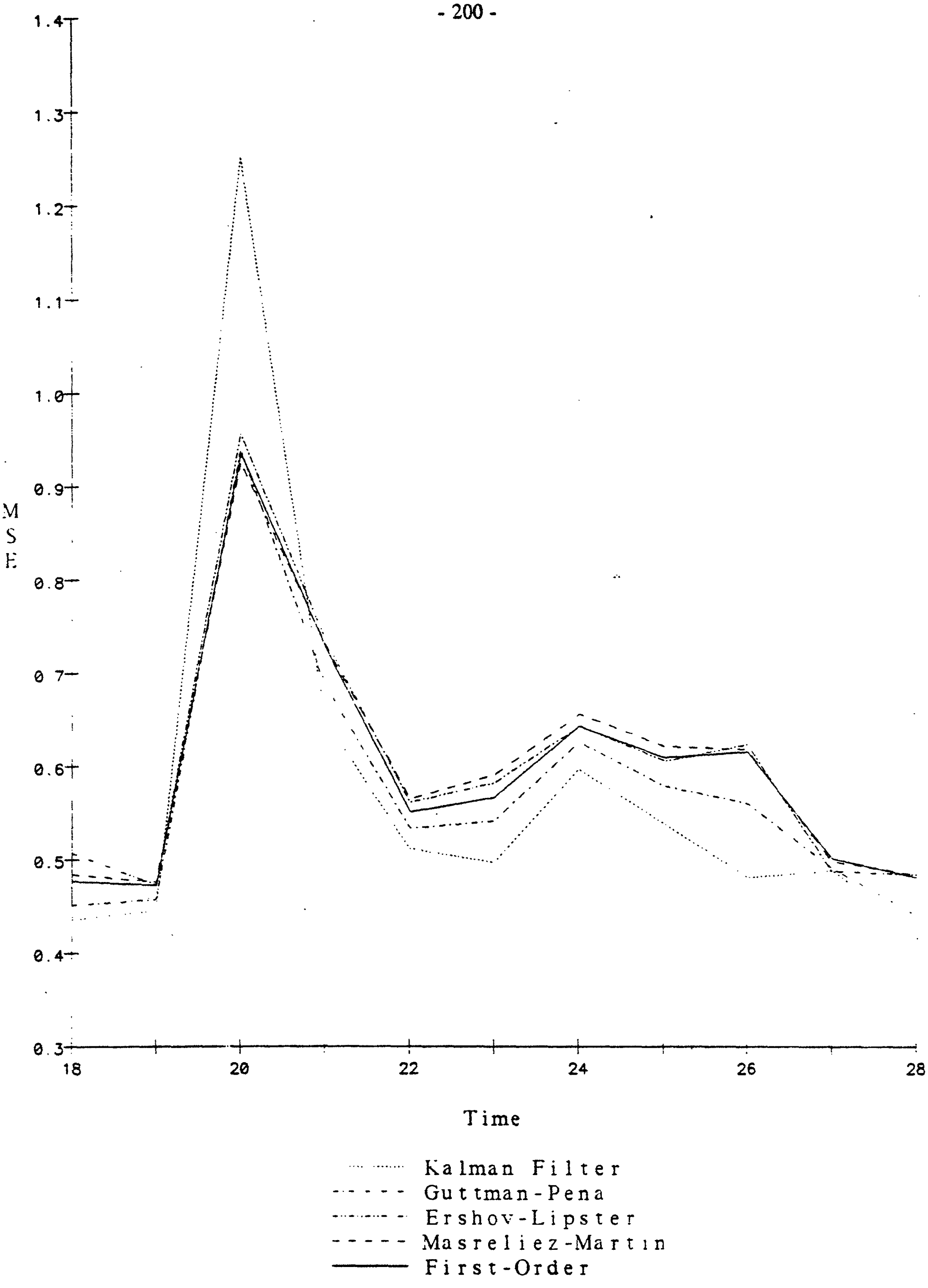

Figure 5.18 Mean Squared Estimation Error ($\tilde{R}_{out} = 4$, $\tilde{\varepsilon} = 0.10$, Scale-Contaminated)

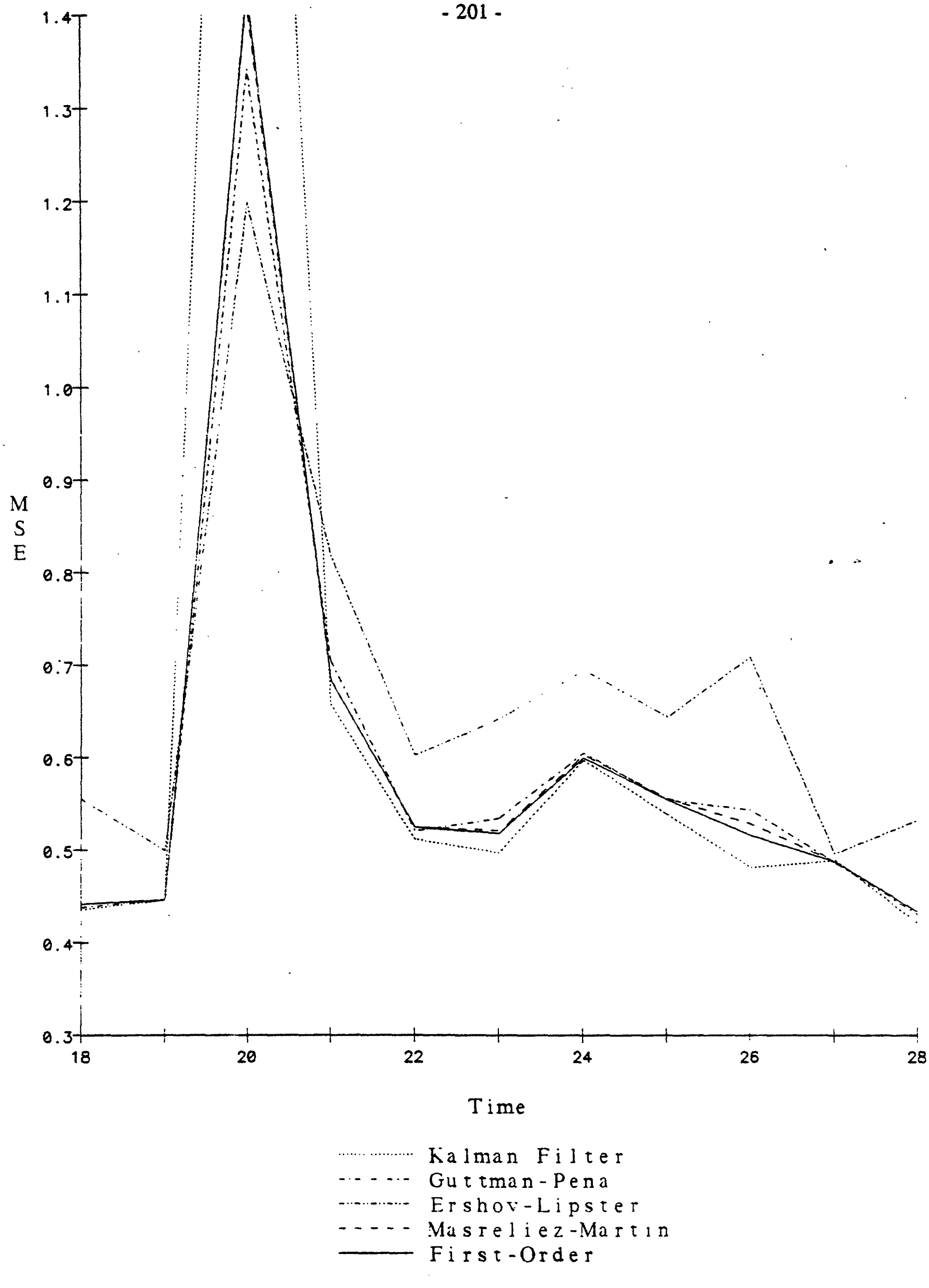

Figure 5.19 Mean Squared Estimation Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.01, Scale-Contaminated)

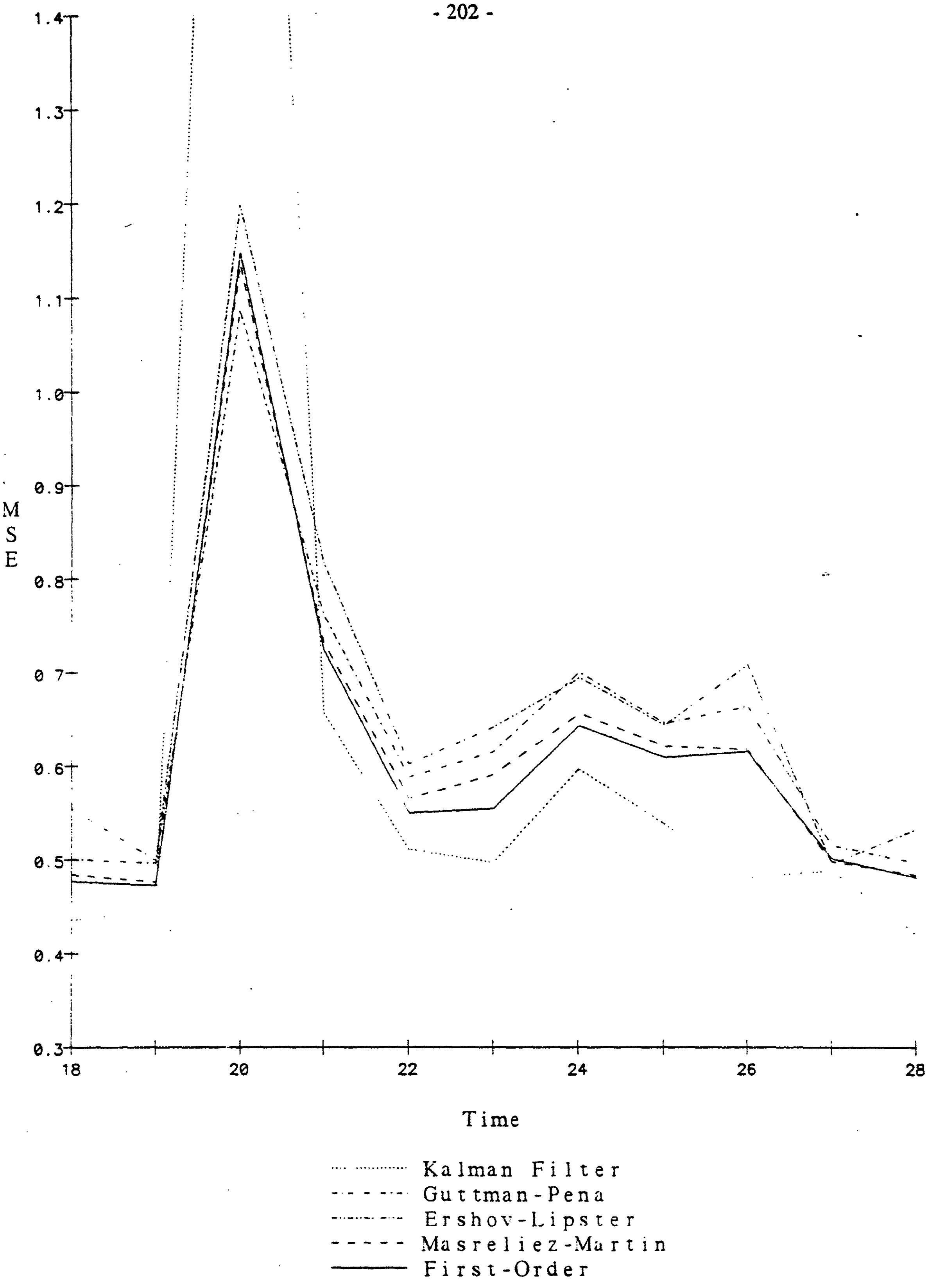

Figure 5.20 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.10, Scale-Contaminated)

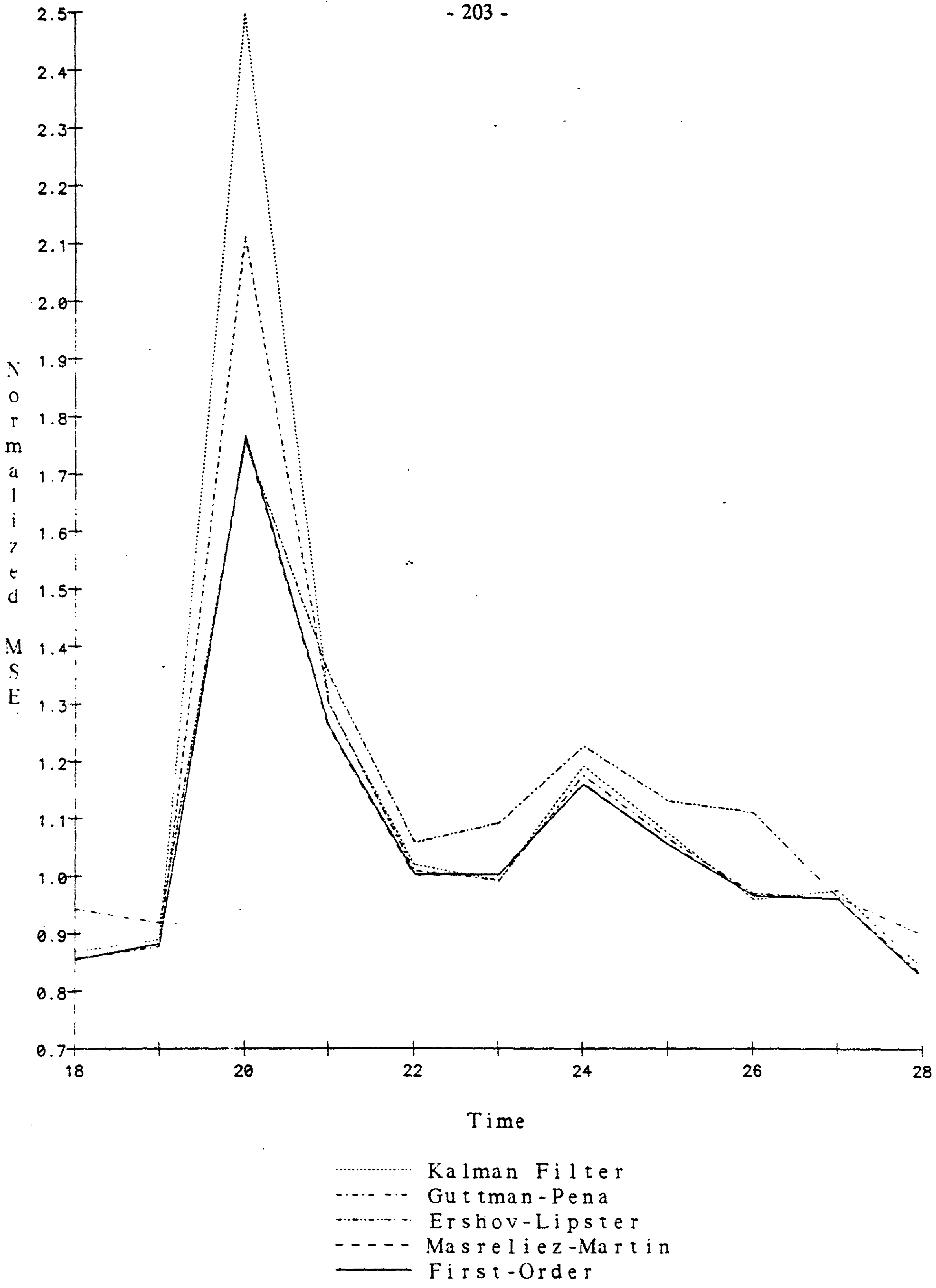

Figure 5.21 Normalized Mean Squared Error ($\tilde{R}_{out} = 4$, $\bar{\varepsilon} = 0.01$, Scale-Contaminated)

The Laplace Distribution. The Laplace (or double exponential) distribution is somewhat heavier tailed than the Gaussian distribution. Moreover, it is similar to Huber's least favorable distribution (for the no process noise case), at least in the tails. In this case, the observation noise obeyed the nominal Gaussian distribution except at $n = 20$, where the noise was Laplacian with variance $R_{out} = \tilde{R}_{out}$. The MSE at $n = 20$ appears in Table 5.8, and show the First-Order estimator to have the best performance at the outlier, for many parameter values. Figures 5.22-23 illustrate the performance of each estimator at and right after an outlier distributed according to the Laplace distribution. For small $\tilde{R}_{out}$ and $\tilde{\varepsilon}$, the estimators behave similarly, except that the Guttman-Peña estimator approaches closer to the performance of the Kalman Filter once the effects of the outlier have attenuated. For large $\tilde{R}_{out}$ and $\tilde{\varepsilon}$, on the other hand, the Masreliez-Martin and First-Order estimators perform virtually identically at the outlier, but the latter does better after the outlier. As with the nominal case, the behavior of the Guttman-Peña and Ershov-Lipster estimators are poor for large $\tilde{R}_{out}$ and $\tilde{\varepsilon}$ after the occurrance of an outlier. The performance of the Kalman Filter in the presence of an outlier is well illustrated by Figure 5.23: its MSE is lower than the robust (hence, suboptimal) estimators everywhere except at the outlier. If F were larger, the effects of the outlier would have taken longer to attenuate, but qualitatively, the respective performance of the estimators would not have changed.

Once again, the lag-one serial correlation of the residual does not change markedly from one estimator to the other in this case. The normalized mean squared errors for two sets of parameters appear in Figures 5.24-25; recall that proximity to unity, not absolute magnitude, is the performance criterion here.

Tukey's "Slash" Distribution. The scale-contaminated Gaussian and Laplace distributions are relatively light tailed, and do not highlight the differences among the various robust estimators analyzed here. Tukey's "slash" distribution has considerably heavier tails, and makes these differences quite apparent. The noise was normally distributed except at $n = 20$, where it obeyed the "slash" distribution. The MSE for $n = 18, \cdots, 28$ is plotted in Figures 5.26-29, and its values at $n = 20$ appear in Table 5.9. The Masreliez-Martin and First-Order estimators perform best at the outlier; while the former is somewhat better than the latter at $n = 20$, the reverse is true after the occurrance of the outlier, as suggested by the nominal case simulations. Moreover, the behavior of the Guttman-Peña and Ershov-Lipster estimators is very poor at the outlier for small values of $\tilde{R}_{out}$, while their performance following the outlier is poor for large values of that parameter.

The accuracy of the estimate of the second moment of the estimation error behaves similarly, as demonstrated by Figure 5.30-31. The Guttman-Peña and Ershov-Lipster estimators drastically underestimate the covariance at the time of an outlier when $\tilde{R}_{out}$ is small; they do better for large $\tilde{R}_{out}$, but in that case, the performance under nominal conditions is quite poor.

The Cauchy Distribution. Another very heavy-tailed noise distribution was the Cauchy distribution, whose variance is infinite. Here, the differences among the performance of the various robust estimators is highlighted most dramatically. Moreover, the deviations of some of the estimators from the state trajectory can get so large in this case, that the lasting effects of the outliers are more visible. Once again, the noise was nominal except at $n = 20$, where it obeyed a Cauchy distribution. The MSE at times $n = 20$ and $n = 21$ are given in Tables 5.10-11. Plots of the MSE at times

Table 5.8 Mean Squared Estimation Error at n = 20 (Laplace)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9094	0.9094	0.9094
$\tilde{R}_{out} = 6.25$	0.9094	0.9094	0.9094
$\tilde{R}_{out} = 9$	0.9094	0.9094	0.9094
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.7986	0.7273	0.7043
$\tilde{R}_{out} = 6.25$	0.7565	0.7059	0.6987
$\tilde{R}_{out} = 9$	0.7412	0.7091	0.7146
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.7169	0.7169	0.7169
$\tilde{R}_{out} = 6.25$	0.7400	0.7400	0.7400
$\tilde{R}_{out} = 9$	0.7648	0.7648	0.7648
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.7412	0.7026	0.6972
$\tilde{R}_{out} = 6.25$	0.7412	0.7026	0.6972
$\tilde{R}_{out} = 9$	0.7412	0.7026	0.6972
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.7396	0.6953	0.6951
$\tilde{R}_{out} = 6.25$	0.7396	0.6953	0.6951
$\tilde{R}_{out} = 9$	0.7396	0.6953	0.6951

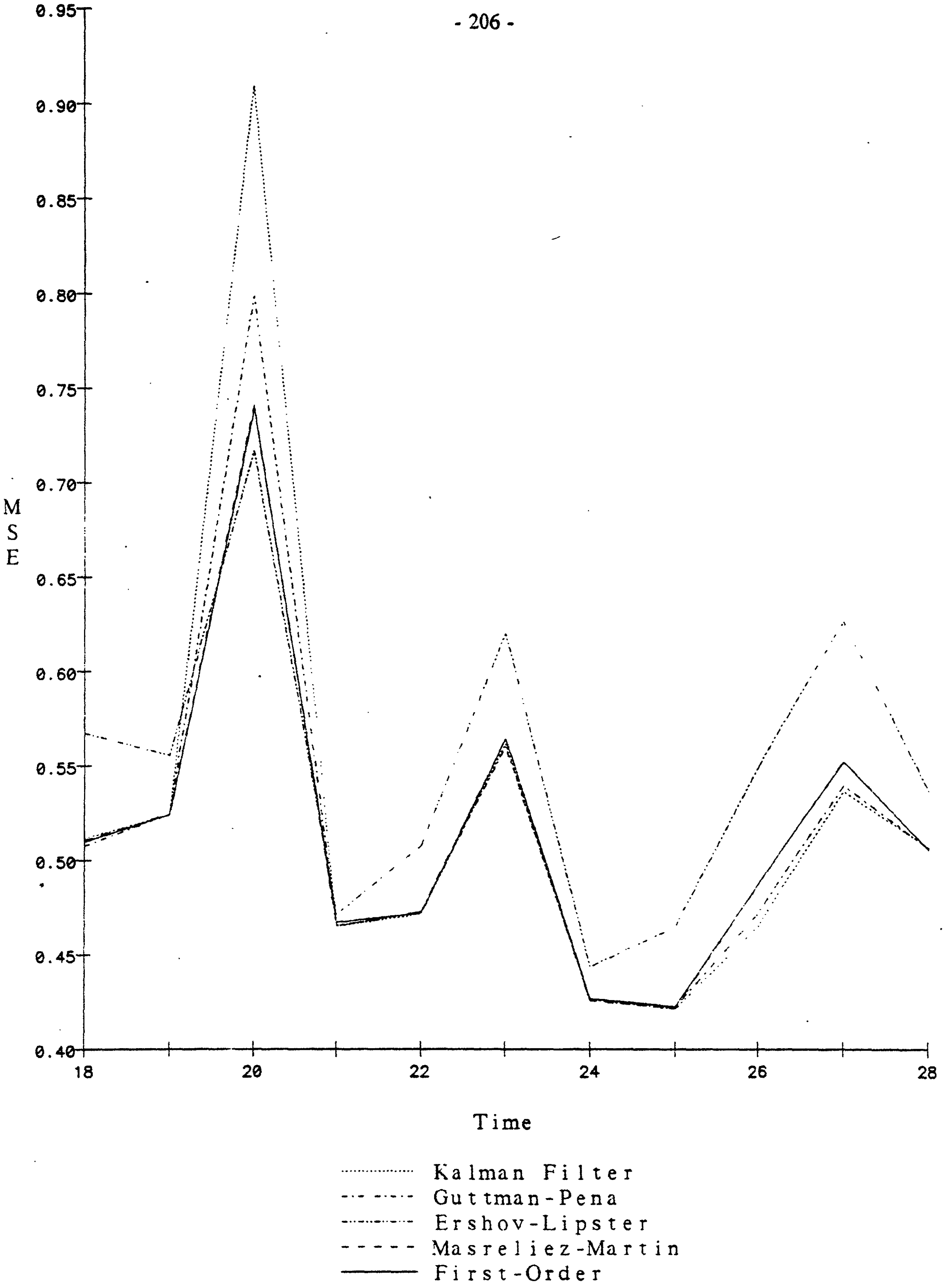

Figure 5.22 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.01, Laplace)

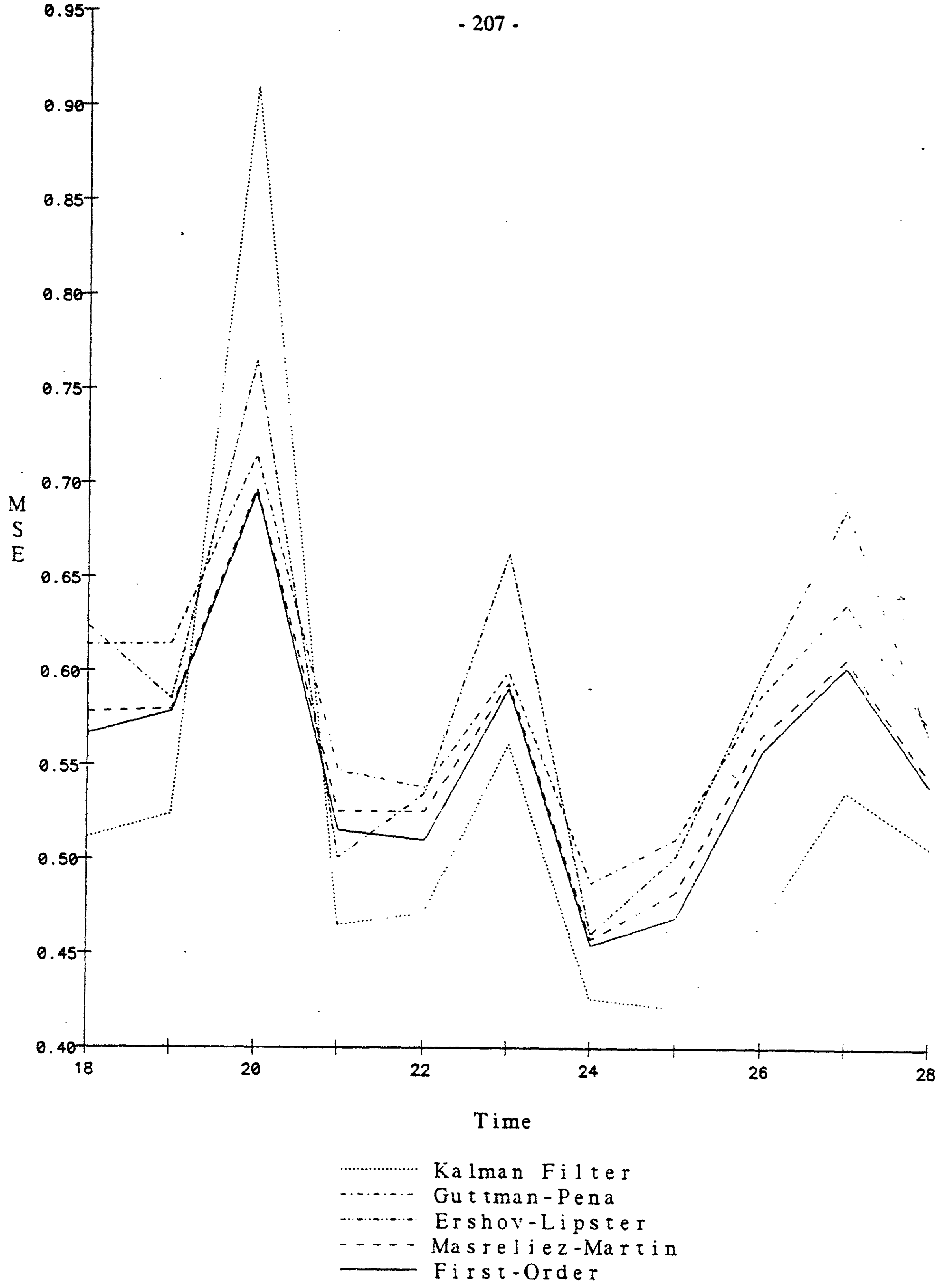

Figure 5.23 Mean Squared Estimation Error ($\tilde{R}_{out} = 9$, $\tilde{\varepsilon} = 0.10$, Laplace)

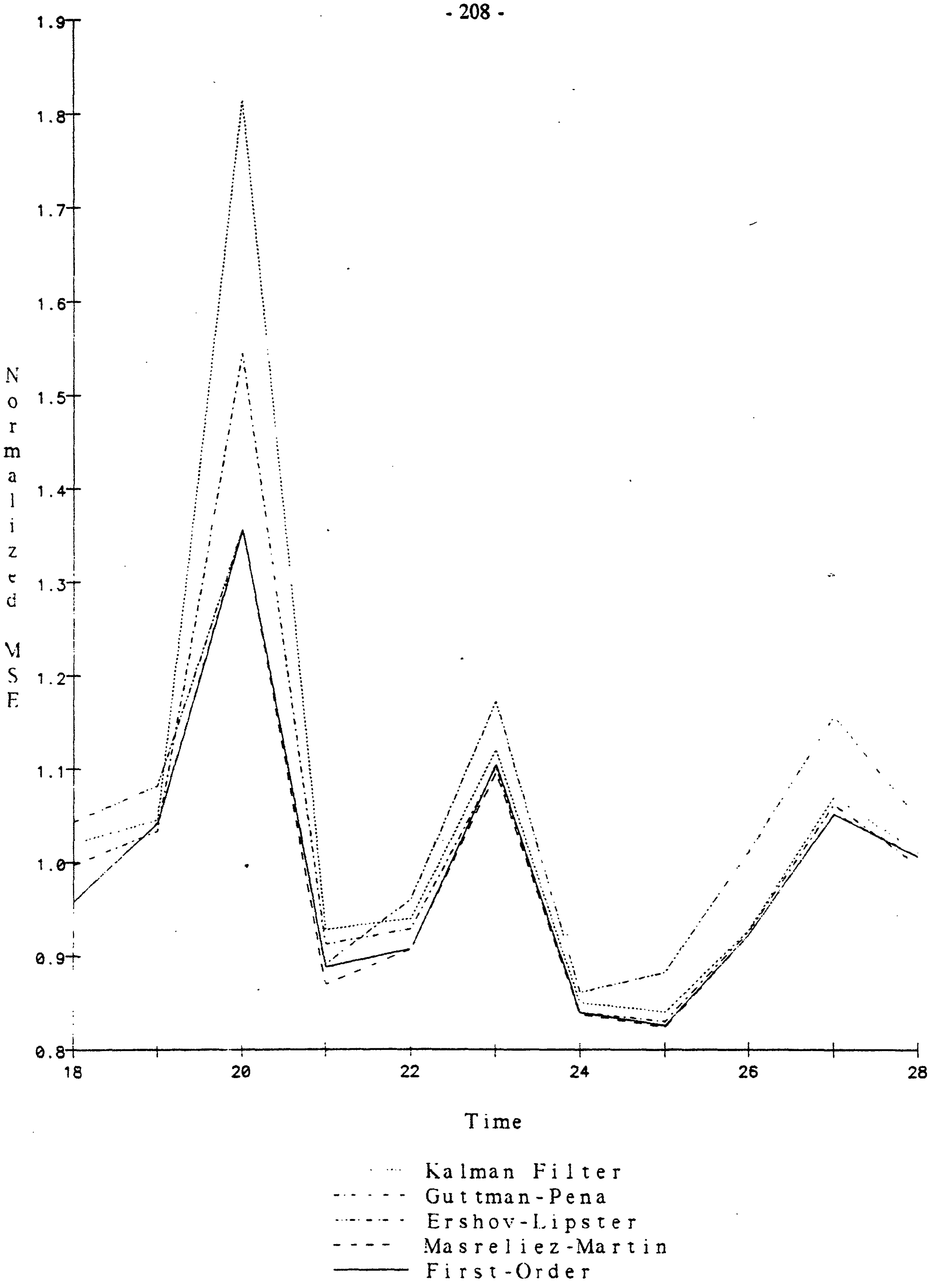

Figure 5.24 Normalized Mean Squared Error ($\tilde{R}_{out} = 4$, $\tilde{\varepsilon} = 0.01$, Laplace)

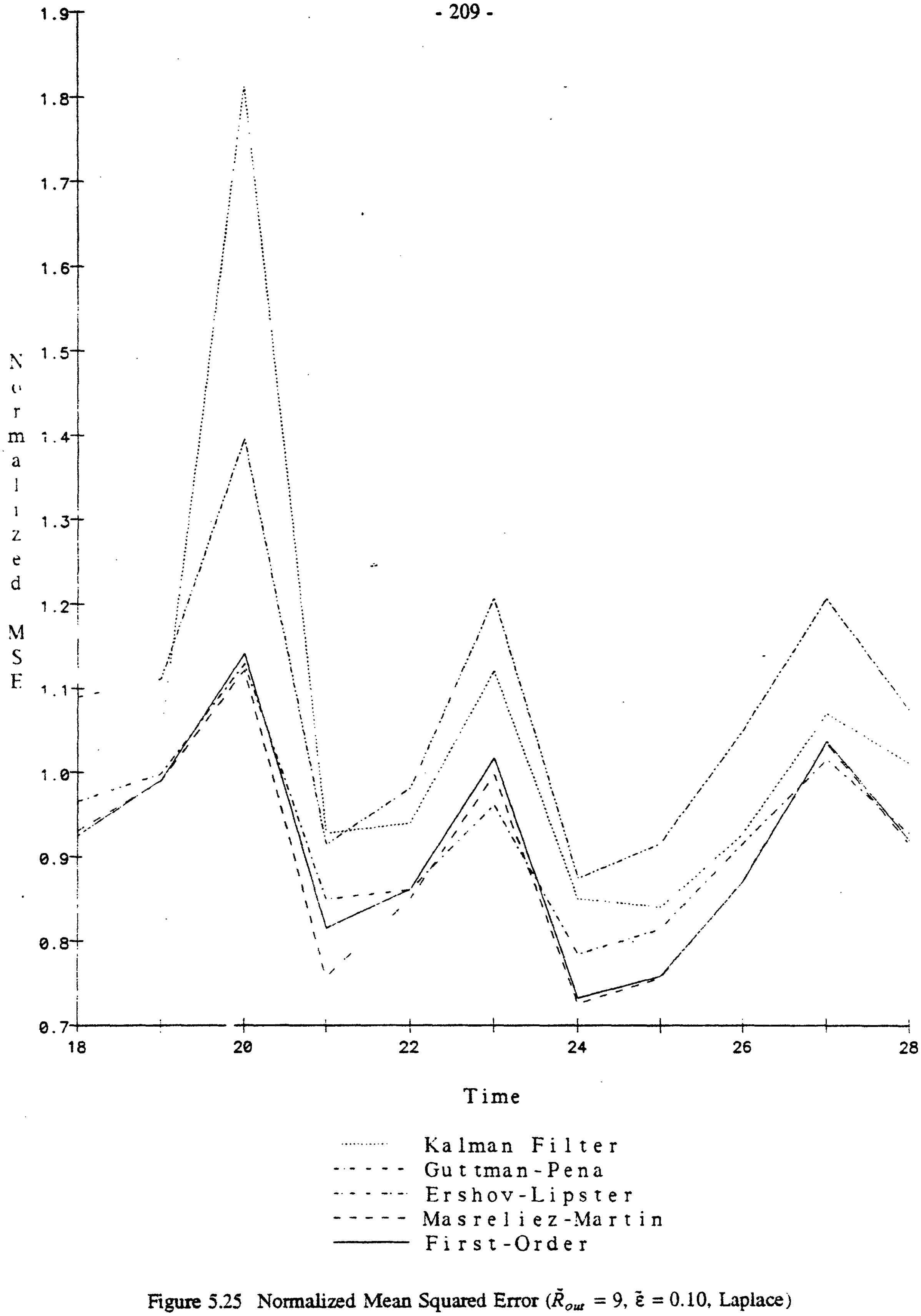

Figure 5.25 Normalized Mean Squared Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, Laplace)

Table 5.9 Mean Squared Estimation Error at n = 20 (Slash)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	23.2744	23.2744	23.2744
$\tilde{R}_{out} = 6.25$	23.2744	23.2744	23.2744
$\tilde{R}_{out} = 9$	23.2744	23.2744	23.2744
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	4.7842	4.6311	4.5733
$\tilde{R}_{out} = 6.25$	2.8361	2.7041	2.6616
$\tilde{R}_{out} = 9$	1.9830	1.8717	1.8404
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	4.5338	4.5338	4.5338
$\tilde{R}_{out} = 6.25$	2.6704	2.6704	2.6704
$\tilde{R}_{out} = 9$	1.8730	1.8730	1.8730
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.3204	1.0990	1.0254
$\tilde{R}_{out} = 6.25$	1.3204	1.0990	1.0254
$\tilde{R}_{out} = 9$	1.3204	1.0990	1.0254
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.3250	1.1072	1.0397
$\tilde{R}_{out} = 6.25$	1.3250	1.1072	1.0397
$\tilde{R}_{out} = 9$	1.3250	1.1072	1.0397

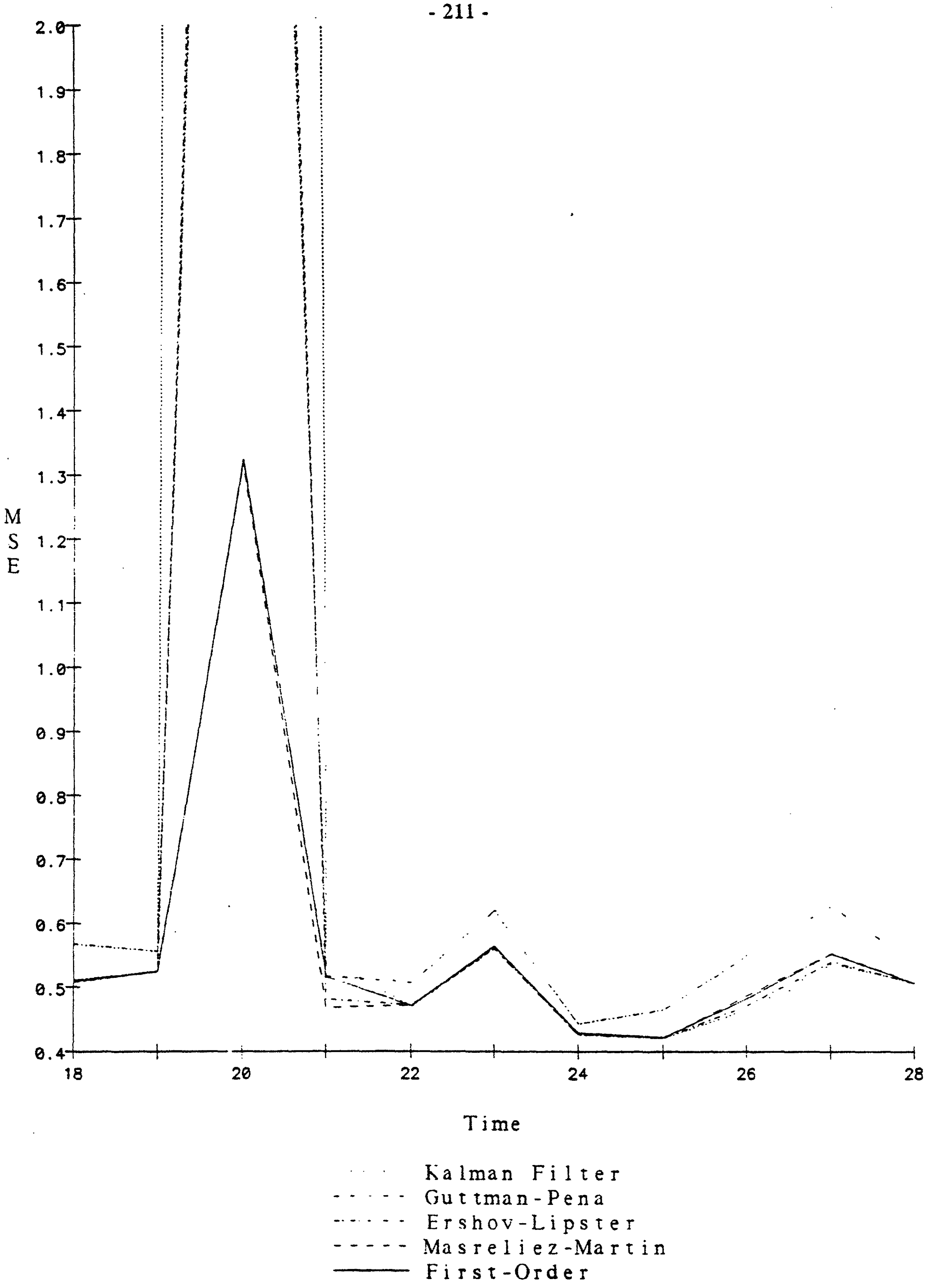

Figure 5.26 Mean Squared Estimation Error ($\bar{R}_{out} = 4$, $\bar{\varepsilon} = 0.01$, Slash)

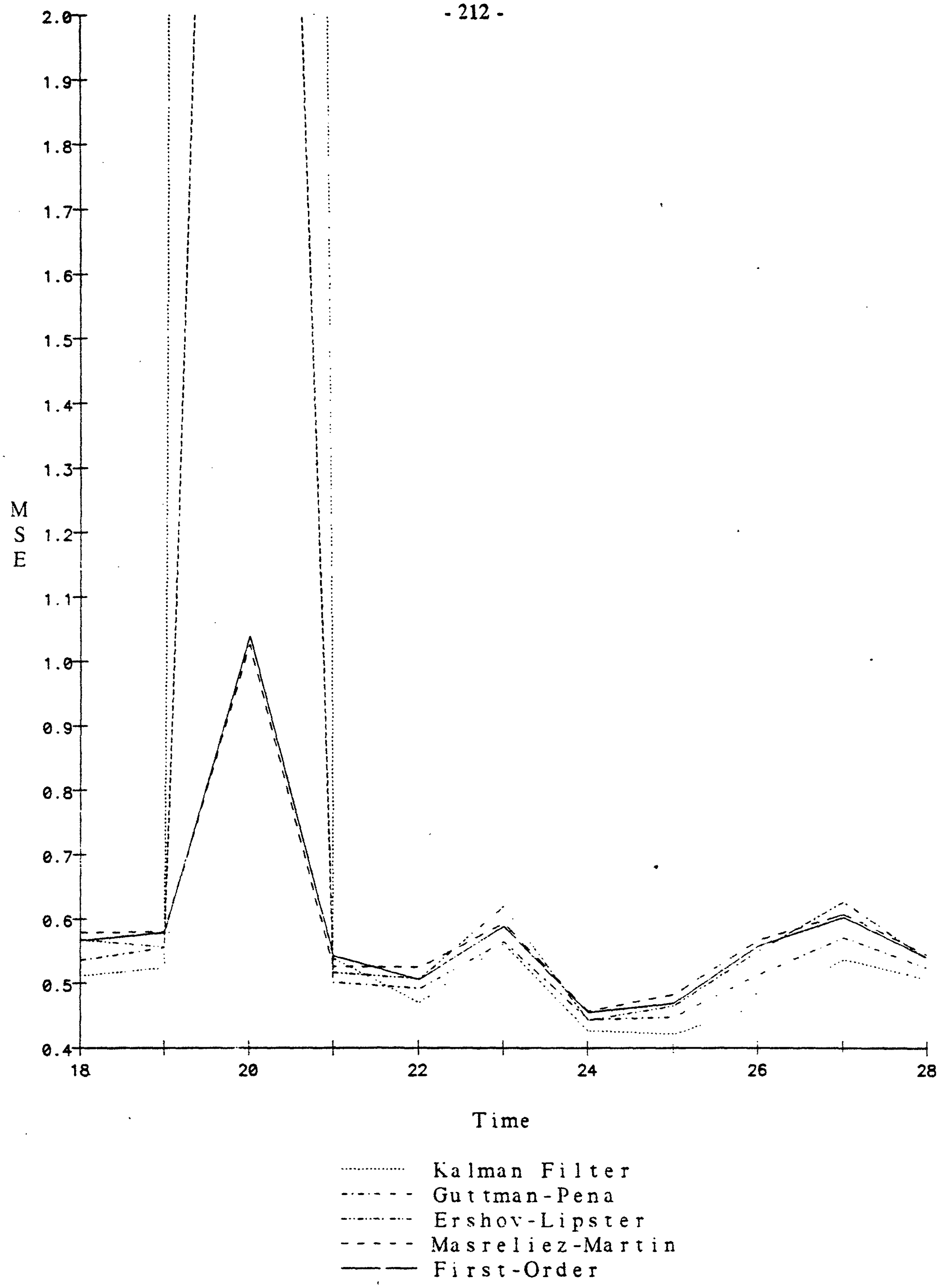

Figure 5.27 Mean Squared Estimation Error ($\breve{R}_{out}$ = 4, $\breve{\varepsilon}$ = 0.10, Slash)

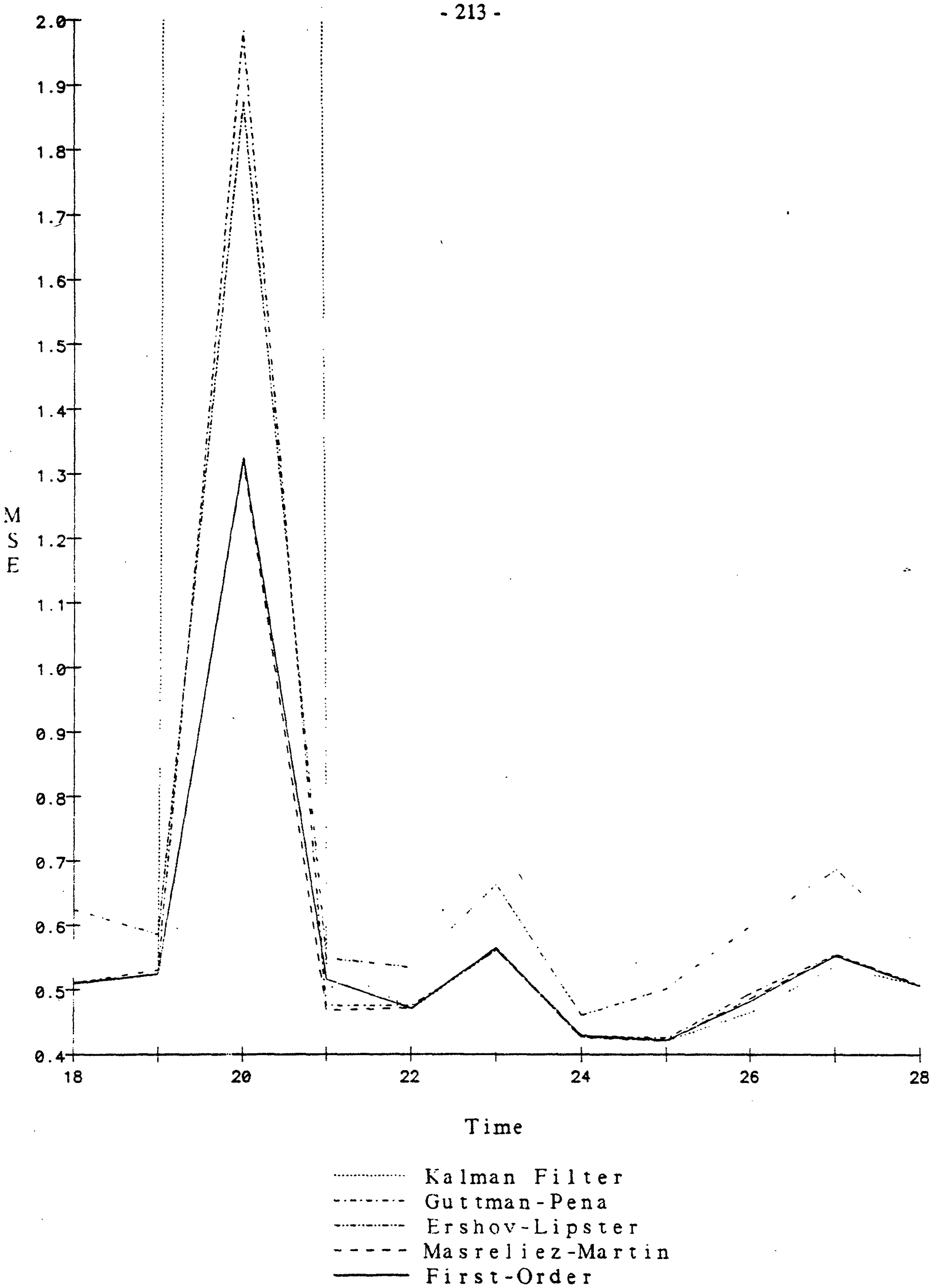

Figure 5.28 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.01, Slash)

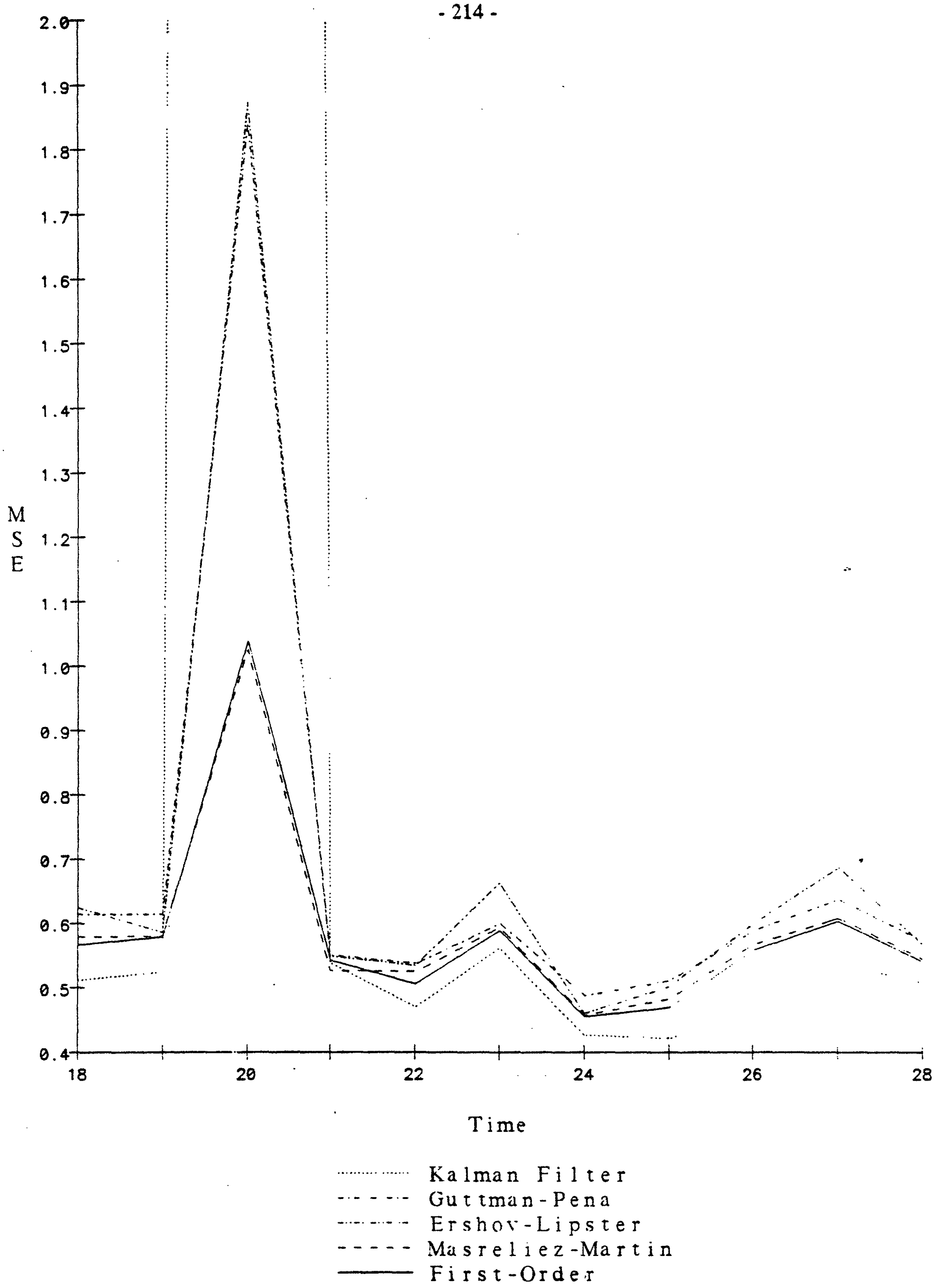

Figure 5.29 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, Slash)

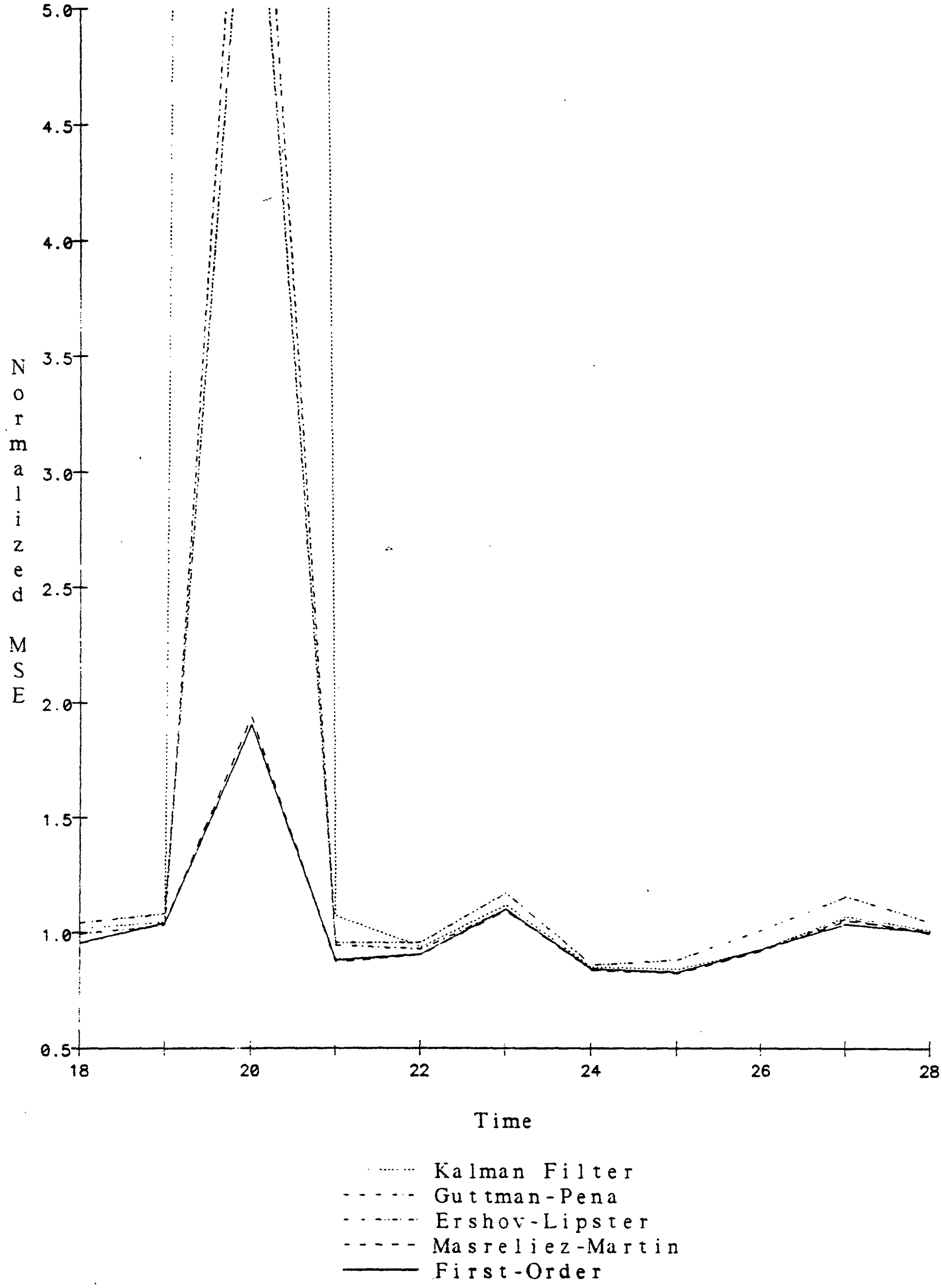

Figure 5.30 Normalized Mean Squared Error ($\tilde{R}_{out} = 4$, $\tilde{\varepsilon} = 0.01$, Slash)

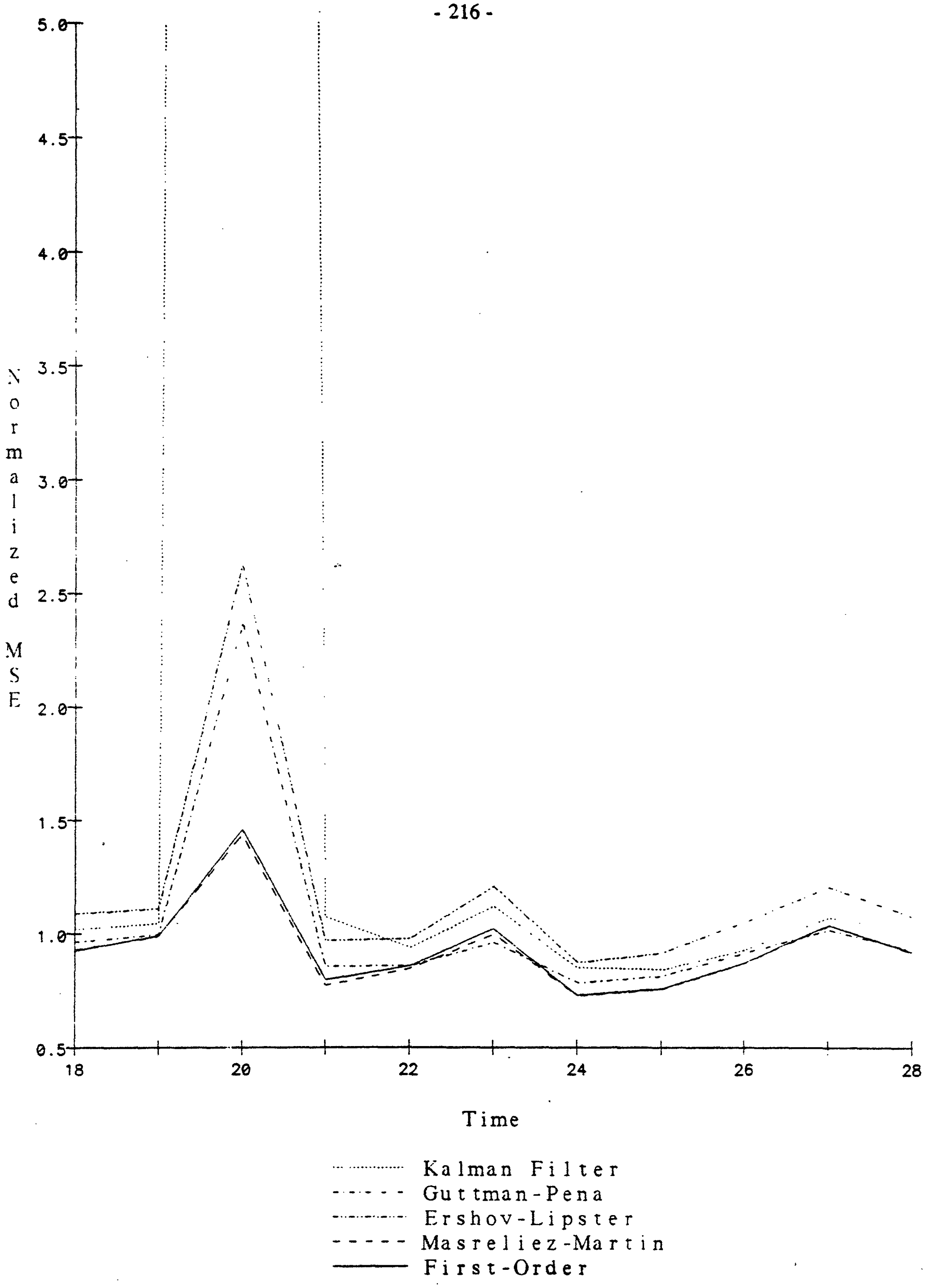

Figure 5.31 Normalized Mean Squared Error ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, Slash)

Table 5.10 Mean Squared Estimation Error at n = 20 (Cauchy)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	838.815	838.815	838.815
$\tilde{R}_{out} = 6.25$	838.815	838.815	838.815
$\tilde{R}_{out} = 9$	838.815	838.815	838.815
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	134.268	134.186	134.209
$\tilde{R}_{out} = 6.25$	63.8841	63.8453	63.8760
$\tilde{R}_{out} = 9$	33.6835	33.6659	33.6976
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	134.065	134.065	134.065
$\tilde{R}_{out} = 6.25$	63.7884	63.7884	63.7884
$\tilde{R}_{out} = 9$	33.6418	33.6418	33.6418
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9375	0.7914	0.7561
$\tilde{R}_{out} = 6.25$	0.9375	0.7914	0.7561
$\tilde{R}_{out} = 9$	0.9375	0.7914	0.7561
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.9394	0.7932	0.7561
$\tilde{R}_{out} = 6.25$	0.9394	0.7932	0.7561
$\tilde{R}_{out} = 9$	0.9394	0.7932	0.7561

Table 5.11 Mean Squared Estimation Error at n = 21 (Cauchy)

Kalman Filter			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	2.8063	2.8063	2.8063
$\tilde{R}_{out} = 6.25$	2.8063	2.8063	2.8063
$\tilde{R}_{out} = 9$	2.8063	2.8063	2.8063
Guttman-Peña			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.5053	1.5314	1.5522
$\tilde{R}_{out} = 6.25$	1.0731	1.1190	1.1588
$\tilde{R}_{out} = 9$	0.8381	0.9038	0.9564
Ershov-Lipster			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.5733	1.5733	1.5733
$\tilde{R}_{out} = 6.25$	1.1793	1.1793	1.1793
$\tilde{R}_{out} = 9$	0.9677	0.9677	0.9677
Masreliez-Martin			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5108	0.5398	0.5727
$\tilde{R}_{out} = 6.25$	0.5108	0.5398	0.5727
$\tilde{R}_{out} = 9$	0.5108	0.5398	0.5727
First-Order			
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	0.5117	0.5500	0.5770
$\tilde{R}_{out} = 6.25$	0.5117	0.5500	0.5770
$\tilde{R}_{out} = 9$	0.5117	0.5500	0.5770

$n = 18, \cdots, 28$ appear in Figures 5.32-35.

The effects of outliers on the residual sequences are illustrated well by Figures 5.36-37, where systematic excursions from whiteness are evident in the cases of the Kalman Filter and the Guttman-Peña and Ershov-Lipster estimators, but much less so in those of the Masreliez-Martin and First-Order estimators.

The normalized mean squared estimation error, plotted in Figures 5.38-39, illustrate once again that the covariance estimates are relatively good for the Masreliez-Martin and First-Order estimators, but not for the others.

Fixed-Amplitude Outliers. To show the influence of the magnitude of an outlier on the estimators, simulations were run with fixed-amplitude outliers at $n = 20$. The MSE is given in Table 5.12, and three cases (for magnitude equal to 2, 6, and 10 times the nominal standard deviation) are plotted in Figures 5.40-42. These show that, although the effect of the outlier may be controlled by choosing larger values for $\tilde{R}_{out}$, this is done at the expense of performance under nominal conditions.

Note finally that the model parameters used in these simulation exercises yield the influence-bounding function cutoffs k, window sizes ω, and error orders $\tilde{\varepsilon}^2 \omega^2$ given in Table 5.13.

5.5 Discussion

Simulation studies such as this one can provide valuable insight into the performance of various robust estimators under different noise distributions, but they seldom yield definitive conclusions or choices valid under all conditions. Consequently, a brief and informal discussion is presented here, concerning some of the lessons taught by the present effort. An important limitation of this simulation study is that it did not involve any comparisons with the performance of an optimal (defined in some sense) estimator; as a result, only the performance of various estimators relative to each other could be assessed.

From this limited vantage point, it can be stated that the Guttman-Peña estimator works very well when outliers are light-tailed, i.e. when the observation noise does not significantly deviate from normality. Despite the fact that the scale-contaminated Gaussian model leads to inflated covariances (as discussed in Section 1.2), this effect is only moderate for small modeled outlier covariance $\tilde{R}_{out}$ and fraction of contamination $\tilde{\varepsilon}$, and the Guttman-Peña estimator was found to have very good nominal performance in those cases. However, it broke down totally when the outliers have heavy-tailed, and values of $\tilde{R}_{out}$ and $\tilde{\varepsilon}$ large enough to mitigate the influence of Cauchy or "slash" outliers yielded severely degraded nominal performance.

The performance of the Ershov-Lipster estimator was somewhat disappointing, although this may in part be due to the choice of outlier detection test used here. Different tests and/or significance levels may yield improved performance, particularly under nominal conditions. In general, the Ershov-Lipster and Guttman-Peña estimators had qualitatively similar behavior, and both exhibited severely degraded performance in the presence of heavy-tailed outliers. This is a consequence of the fact that the adaptive/switching covariance (or gain) scheme employed by both estimators decreases but does not

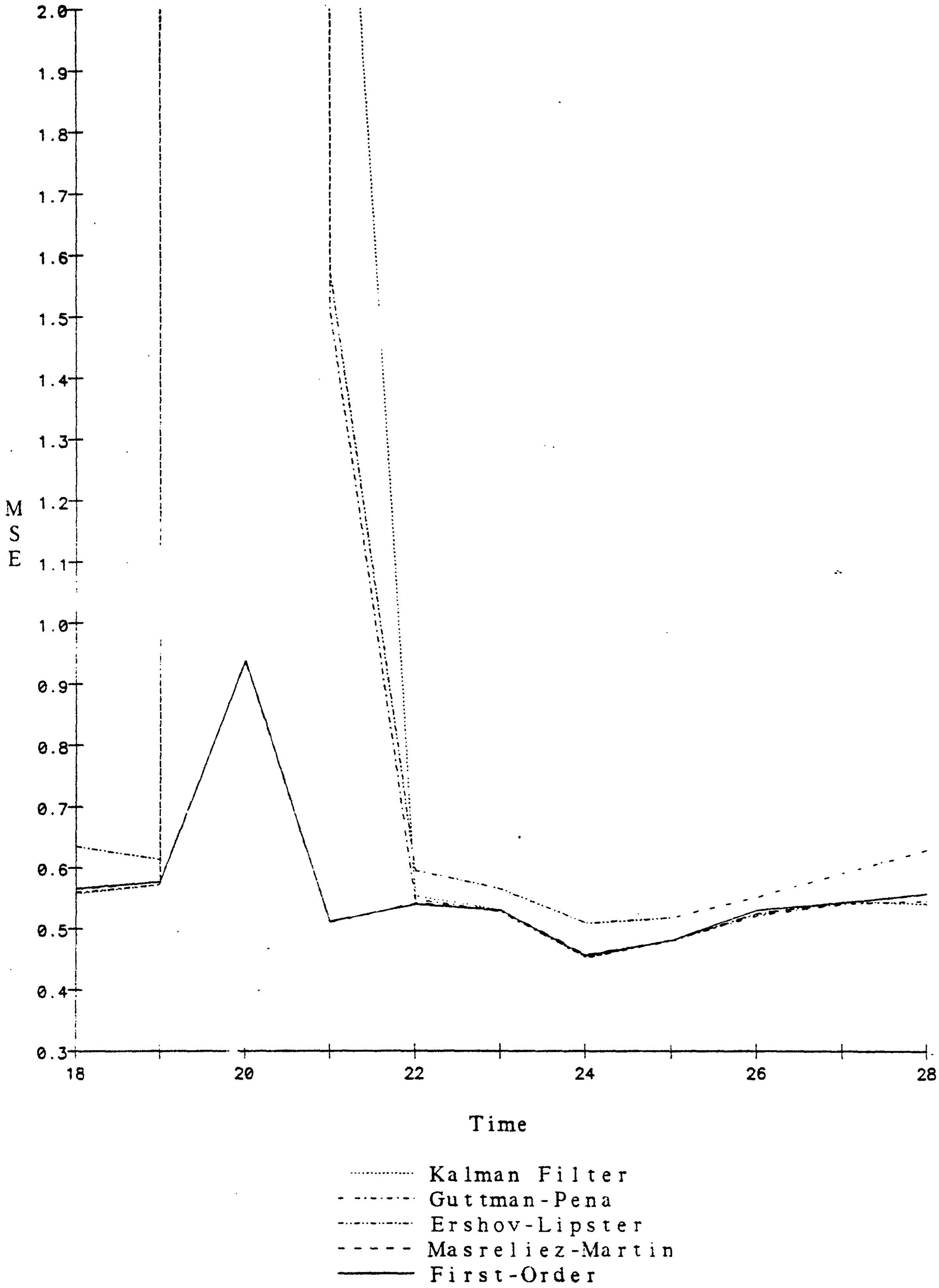

Figure 5.32 Mean Squared Estimation Error ($\tilde{R}_{out} = 4$, $\tilde{\varepsilon} = 0.01$, Cauchy)

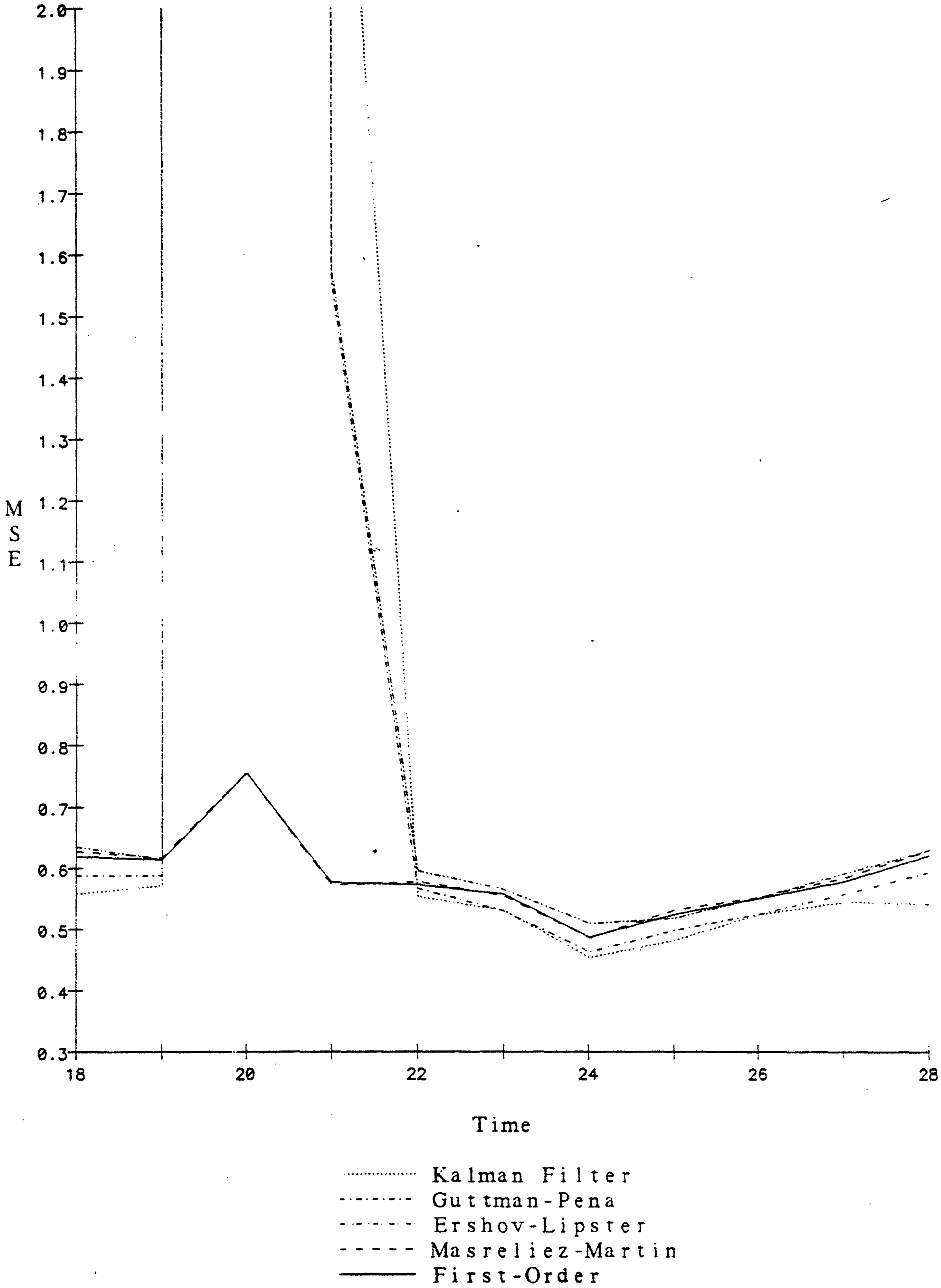

Figure 5.33 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.10, Cauchy)

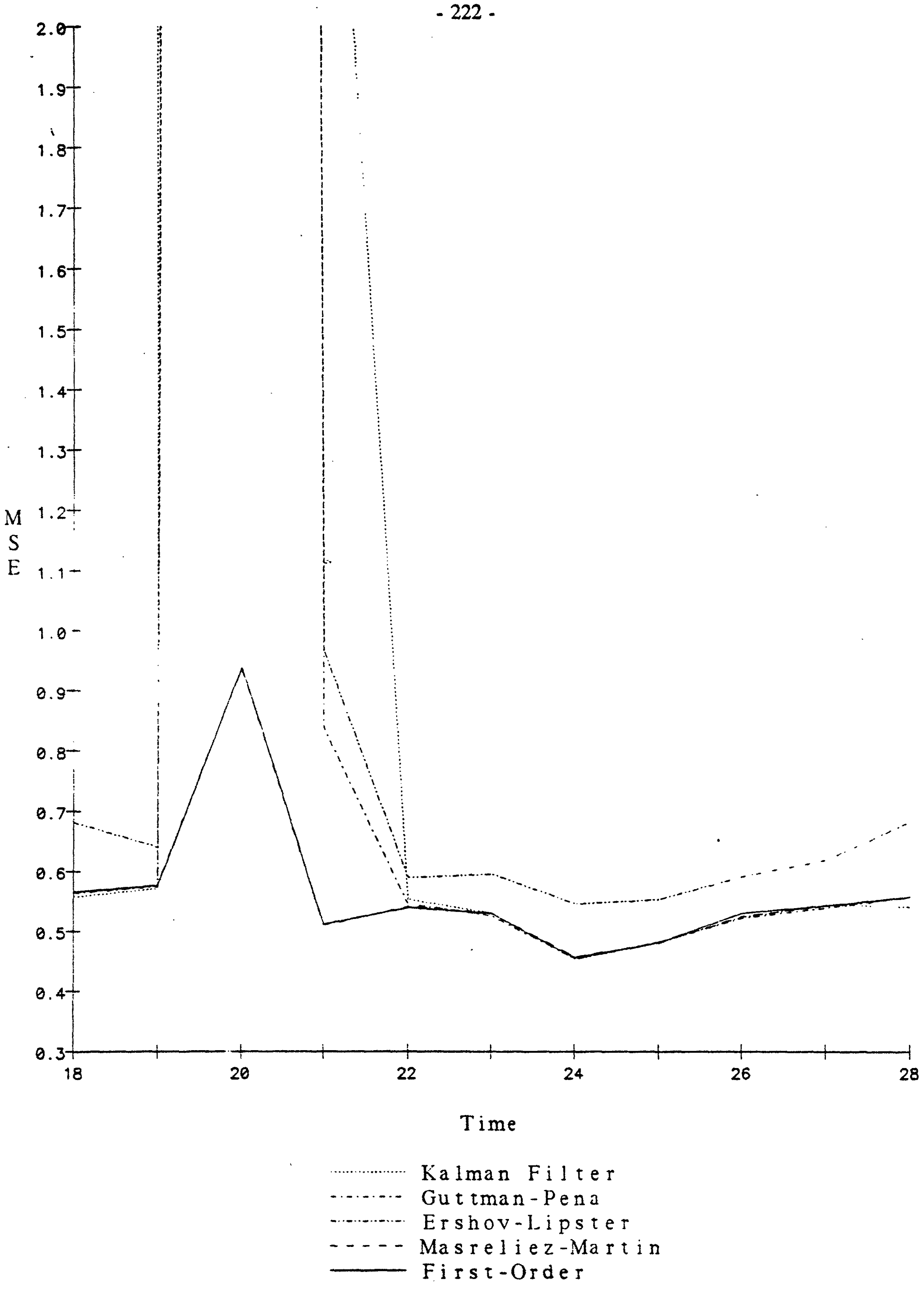

Figure 5.34 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.01, Cauchy)

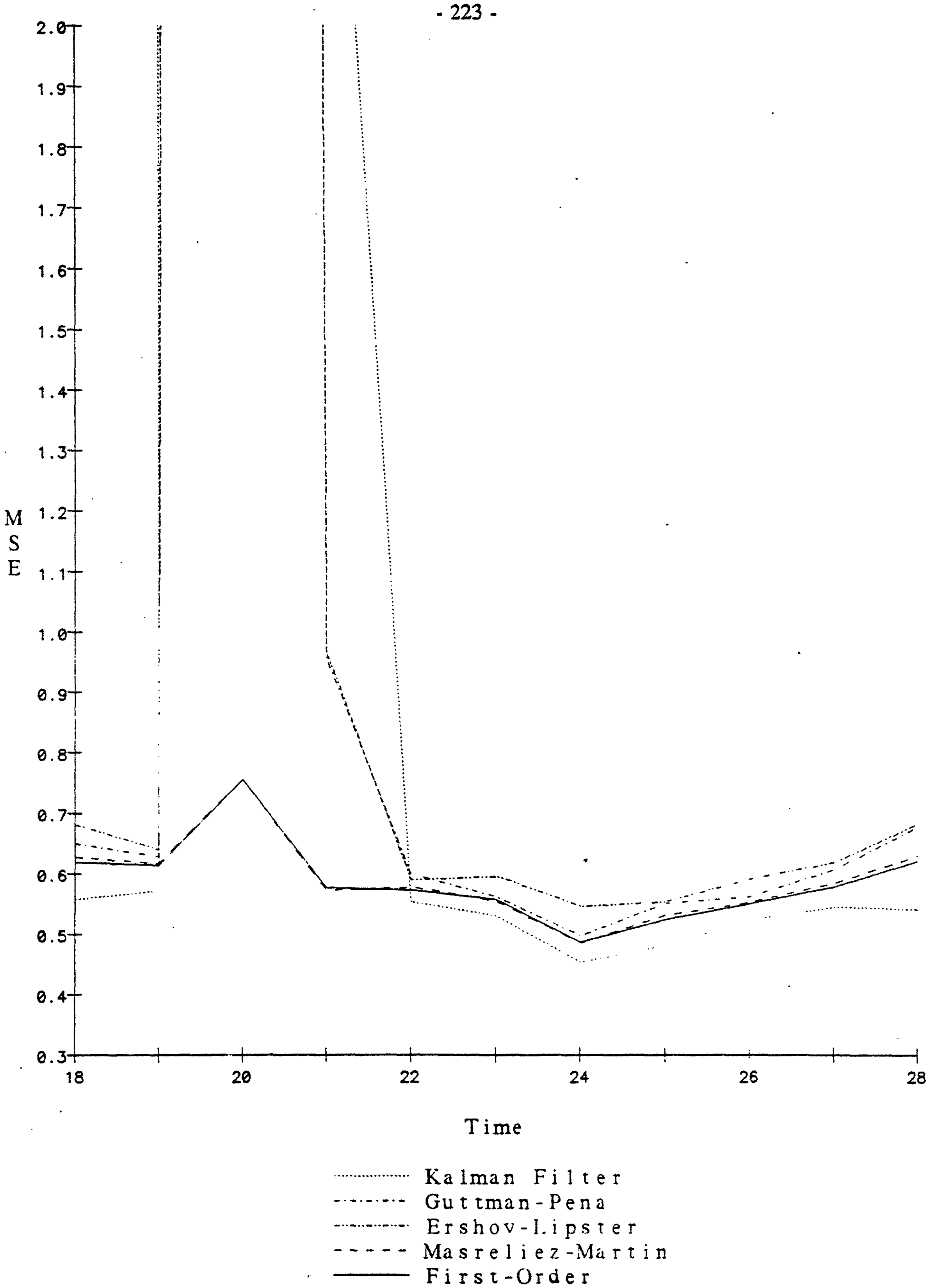

Figure 5.35 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.10, Cauchy)

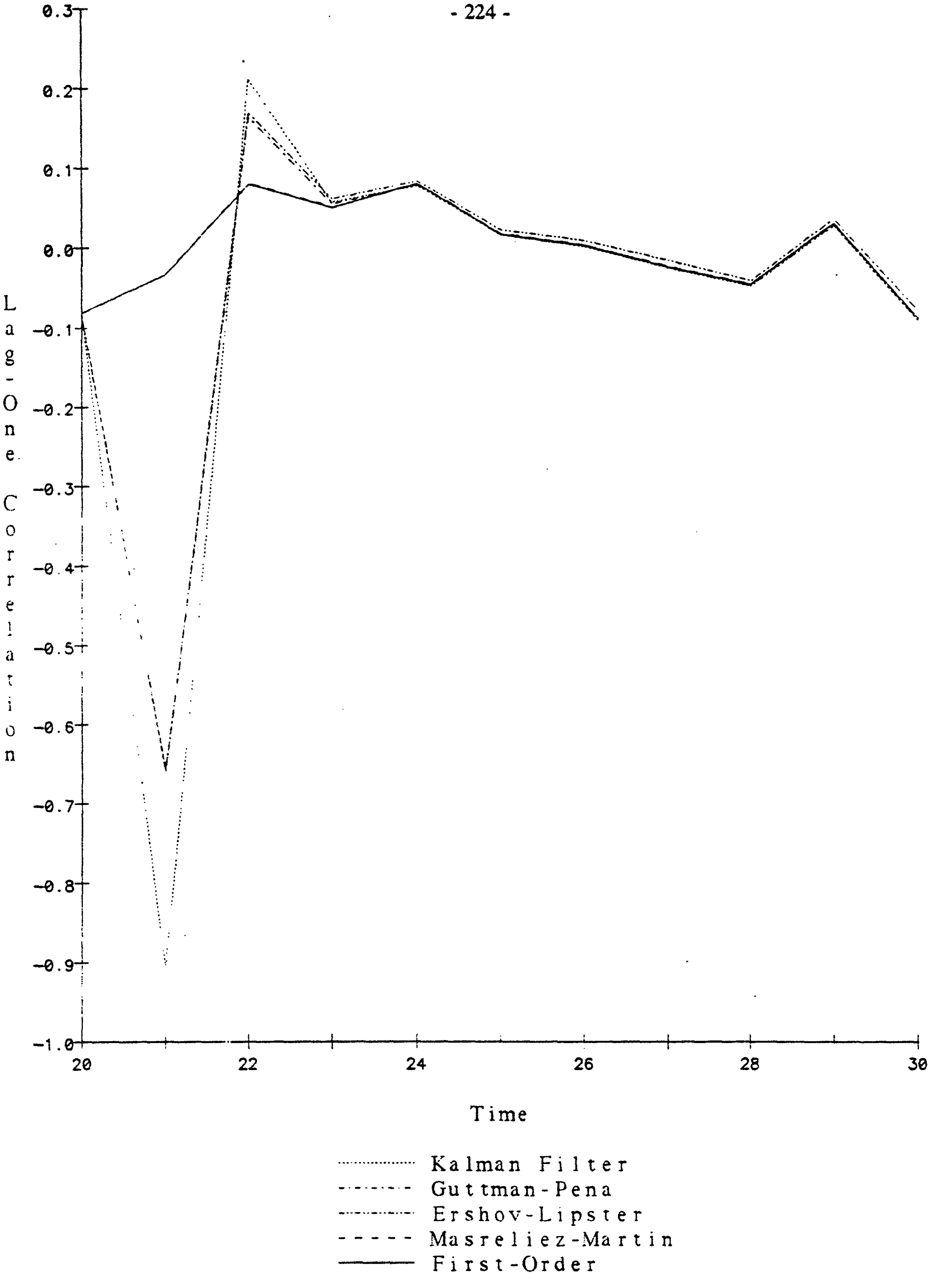

Figure 5.36 Lag-One Serial Correlation ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.01, Cauchy)

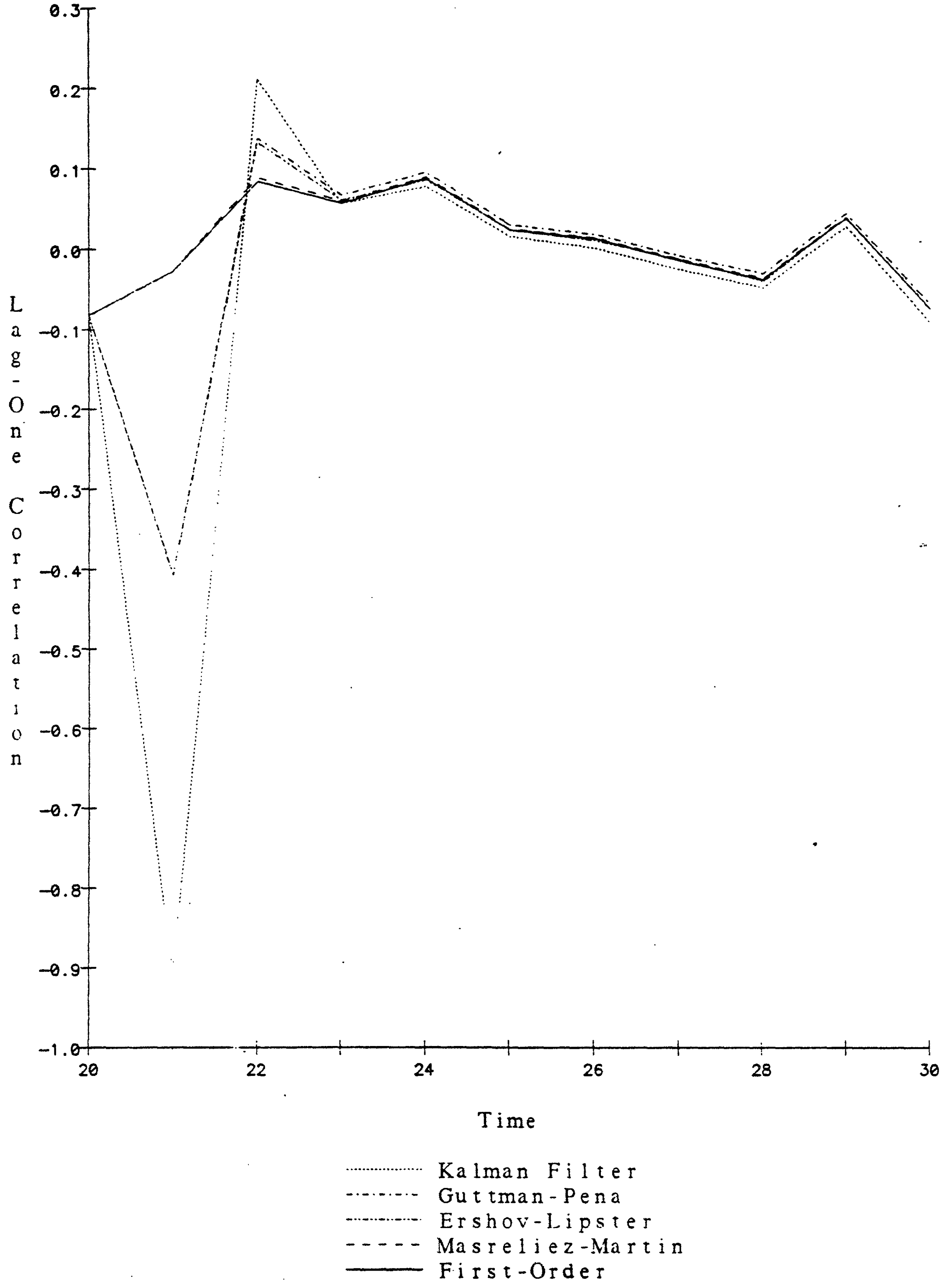

Figure 5.37 Lag-One Serial Correlation ($\bar{R}_{out}$ = 9, $\bar{\varepsilon}$ = 0.10, Cauchy)

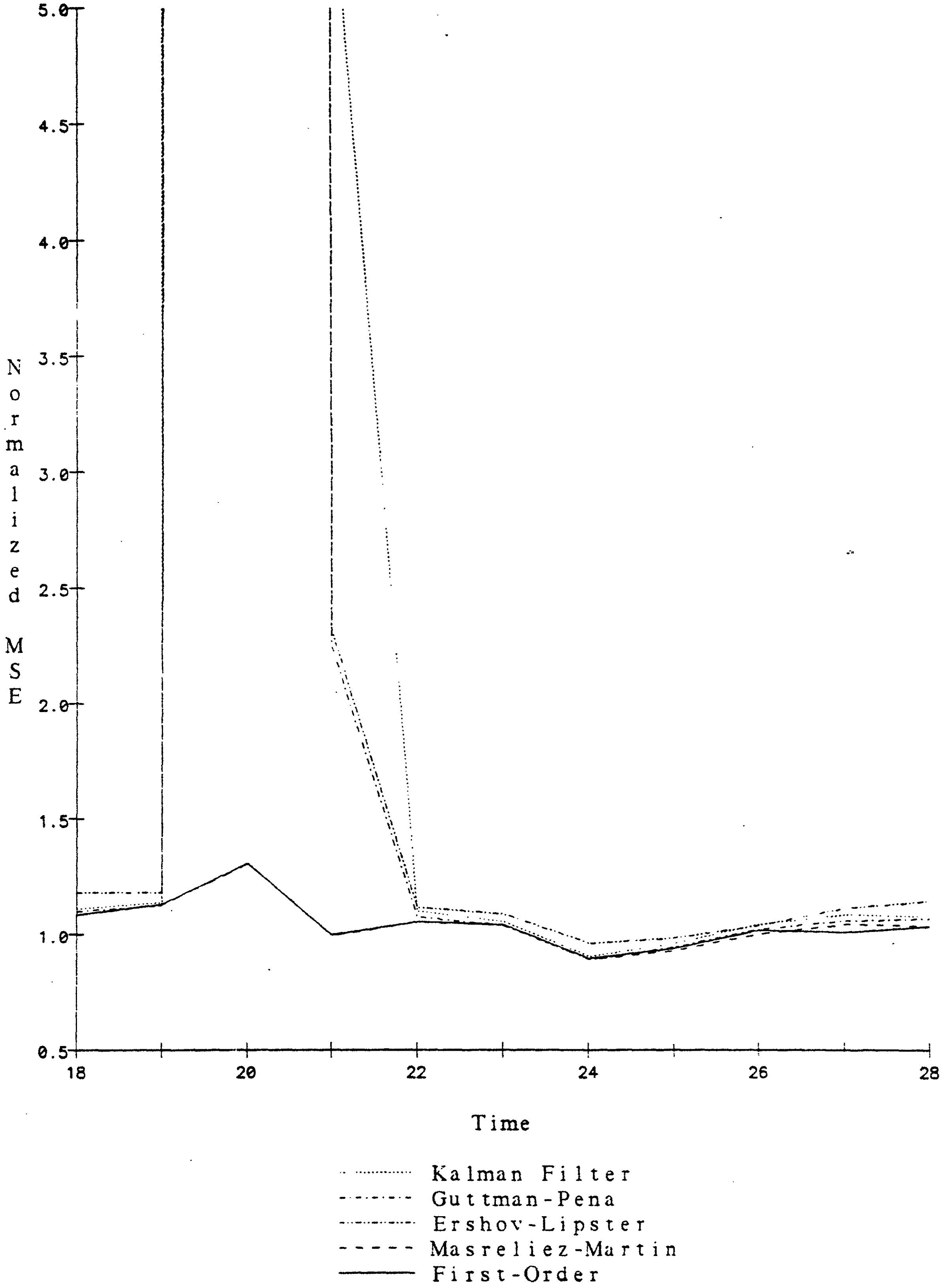

Figure 5.38 Normalized Mean Squared Error ($\tilde{R}_{out} = 4$, $\check{\varepsilon} = 0.01$, Cauchy)

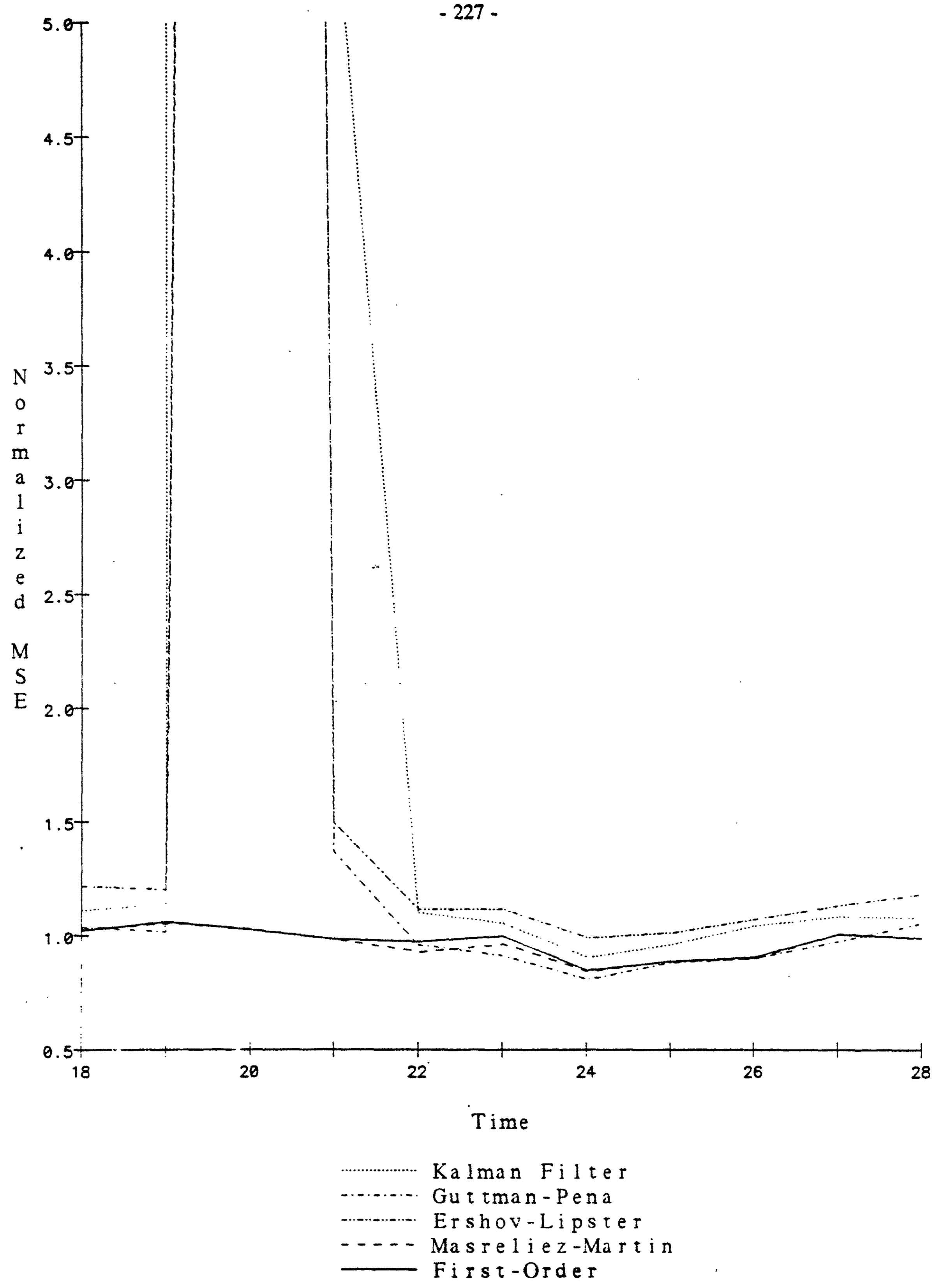

Figure 5.39 Normalized Mean Squared Error ($\tilde{R}_{out}$ = 9, $\tilde{\varepsilon}$ = 0.10, Cauchy)

Table 5.12 Mean Squared Estimation Error at $n = 20$ (Fixed Amplitude)

Kalman Filter

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.3652	2.6467	4.4308
$\tilde{R}_{out} = 6.25$	6.7173	9.5064	12.7979
$\tilde{R}_{out} = 9$	16.5920	20.8886	25.6876

Guttman-Peña

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.2957	1.9721	1.9943
$\tilde{R}_{out} = 6.25$	1.5976	1.3328	1.2497
$\tilde{R}_{out} = 9$	1.2368	1.2382	1.2404

Ershov-Lipster

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.5021	2.1554	1.8907
$\tilde{R}_{out} = 6.25$	1.3411	1.2693	1.2260
$\tilde{R}_{out} = 9$	1.2311	1.2345	1.2367

Masreliez-Martin

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.3766	2.0160	2.2701
$\tilde{R}_{out} = 6.25$	2.3281	2.3338	2.3338
$\tilde{R}_{out} = 9$	2.3338	2.3338	2.3338

First-Order

	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
$\tilde{R}_{out} = 4$	1.3749	2.0477	2.3411
$\tilde{R}_{out} = 6.25$	2.4006	2.3704	2.3419
$\tilde{R}_{out} = 9$	2.3320	2.3302	2.3300

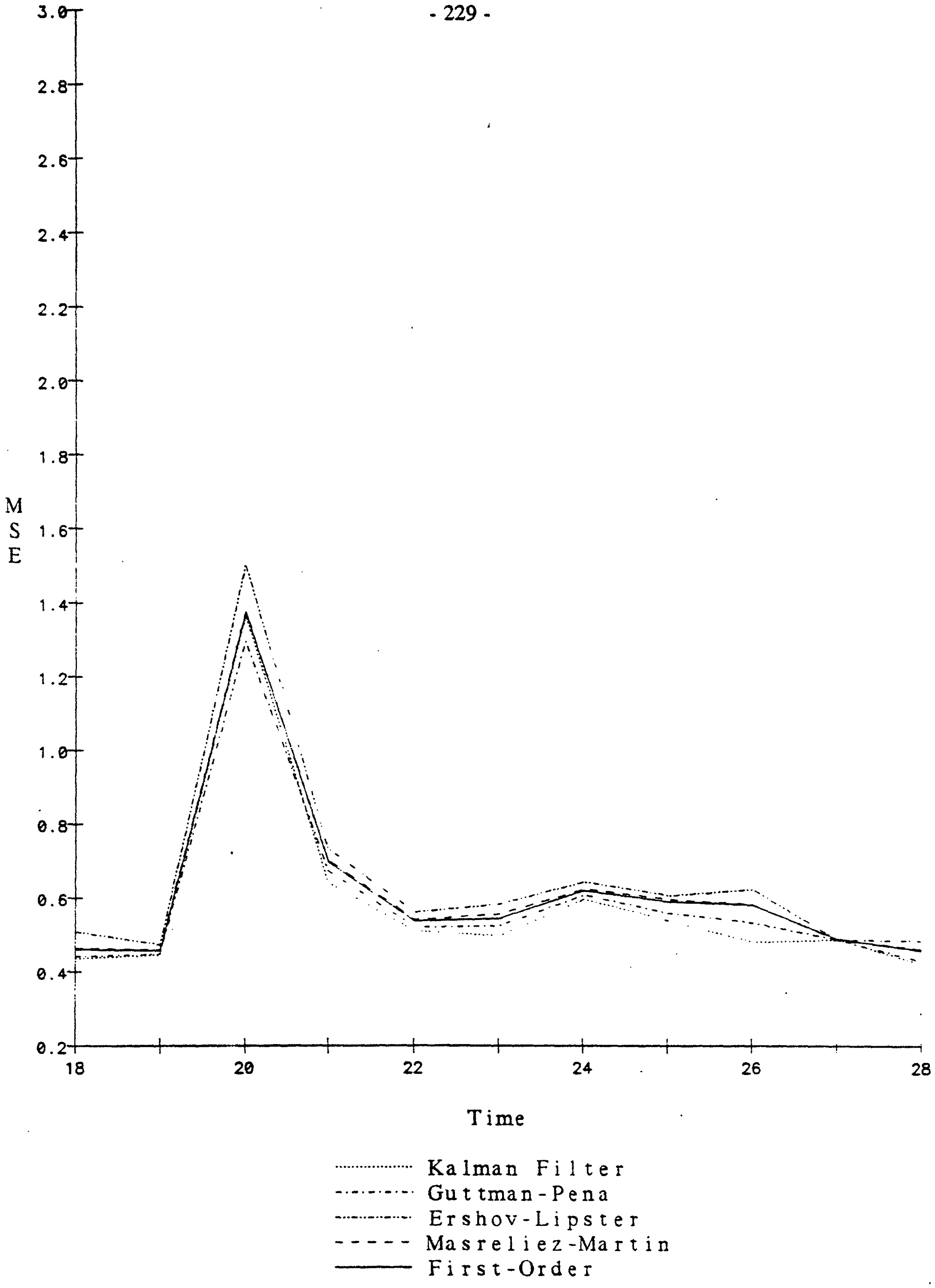

Figure 5.40 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 4, $\tilde{\varepsilon}$ = 0.05, Fixed)

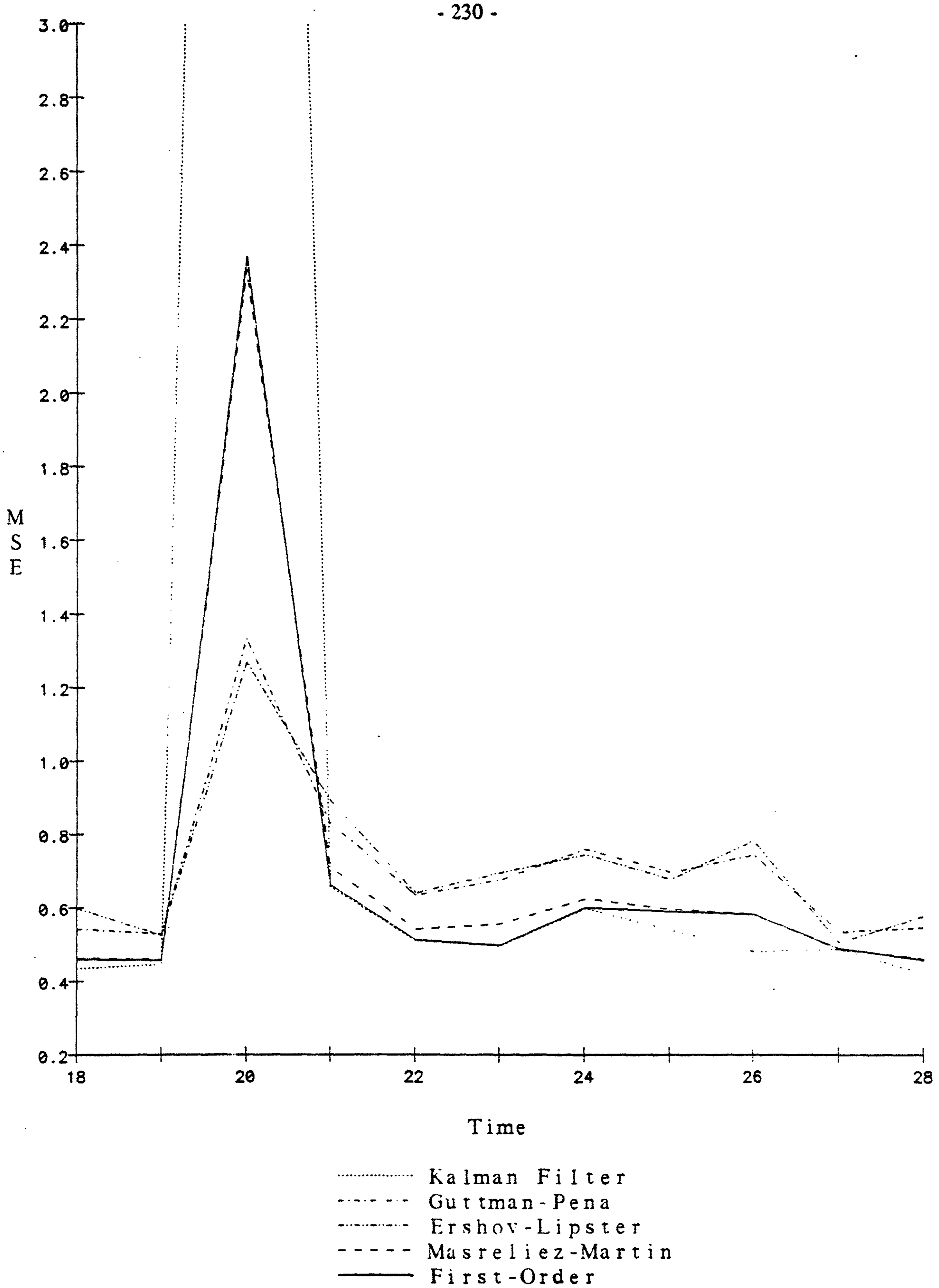

Figure 5.41 Mean Squared Estimation Error ($\bar{R}_{out}$ = 36, $\bar{\varepsilon}$ = 0.05, Fixed)

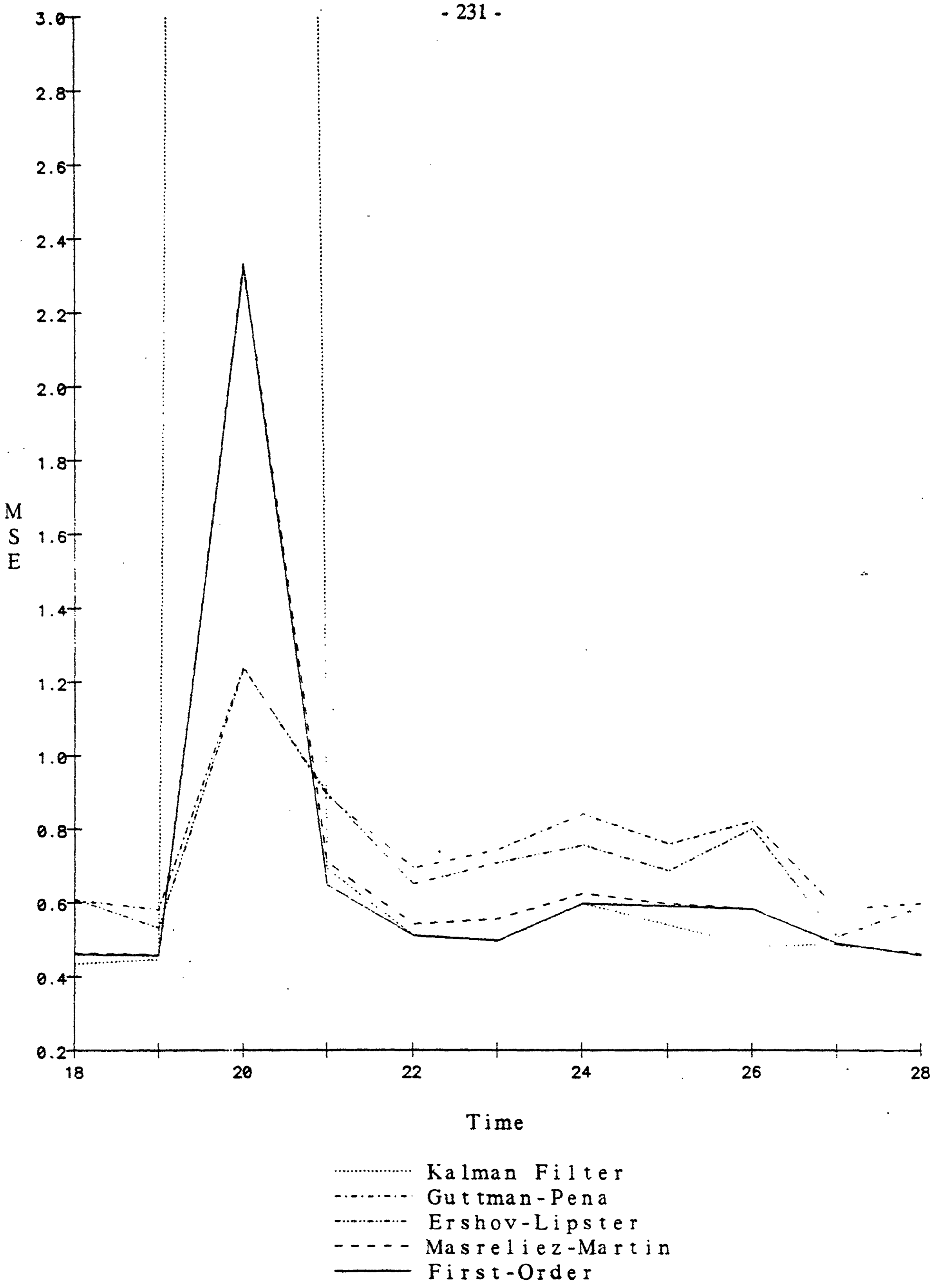

Figure 5.42 Mean Squared Estimation Error ($\tilde{R}_{out}$ = 100, $\bar{\varepsilon}$ = 0.05, Fixed)

Table 5.13 Parameters of the First-Order Estimator

	$F = 0.1$			$F = 0.5$		
	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$	$\tilde{\varepsilon} = 0.01$	$\tilde{\varepsilon} = 0.05$	$\tilde{\varepsilon} = 0.10$
k	1.9451	1.3983	1.1410	1.9451	1.3983	1.1410
β_1	-0.6425	-0.6425	-0.6425	-0.3109	-0.3109	-0.3109
δ	0.5260	0.5260	0.5260	0.7328	0.7328	0.7328
ω	7	5	4	15	10	7
$\tilde{\varepsilon}\omega$	0.07	0.25	0.40	0.15	0.50	0.70
$\tilde{\varepsilon}^2\omega^2$	0.0049	0.0625	0.1600	0.0225	0.2500	0.4900

bound the influence of very large observations.

In contrast, the Masreliez-Martin and First-Order estimators do bound the influence of large outliers, resulting in performance far superior to the Guttman-Peña and Ershov-Lipster estimators when the outliers are very heavy tailed. The Masreliez-Martin estimator has been shown empirically (*via* Monte Carlo studies) to perform very well under fairly broad conditions, but has not gained wide acceptance due to the rather arbitrary assumptions on which it is based. The present study confirms that this estimator performs quite well, and moreover this thesis suggests theoretical reasons to explain this performance.

The First-Order and Masreliez-Martin estimators perform comparably in most cases, and yet the former is considerably more complex than the latter. The derivation of the First-Order estimator suggests that, when the true outlier distribution is known exactly, or at least approximately, the additional complexity of the First-Order estimator may yield considerable improvement. However, when this distribution is very far from the modeled distribution used in the derivation of the estimator, the "correction" terms T_n^i in the estimator may not help and indeed may hurt the performance of the estimator. In other words, the choice between the Masreliez-Martin and First-Order estimators must depend on the particular application at hand.

It must also be noted that particular applications may dictate higher-order expansions than the first-order expansion used here. For example, the First-Order estimator is not robust to two or more outliers occurring in quick succession, as discussed in Section 4.2; the Masreliez-Martin estimator, on the other hand, is not sensitive to the configuration of outliers, but this is achieved at the expense of some nominal performance. Thus, the choice of which estimator to use will have to be based on the particular problem under consideration.

Leaving aside the variability among the robust filters tested here, and ignoring for a moment their respective strengths and weaknesses, the present simulation study shows once again that the Kalman Filter breaks down in the presence of outliers, thus confirming the need for robust recursive estimators when the noise significantly deviates from normality.

6. Conclusion

This thesis 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. It aims at deriving an estimator that is not only of practical value, but is also based on sound theory, so as to be useful for inference as well.

A brief summary of the thesis appears in Section 6.1, followed in Section 6.2 by a list of research topics motivated by the work described herein.

6.1 Summary

The relationship between point estimation and filtering is clear: 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. This relationship is at the root of the present thesis.

Huber's theory of minimax robust estimation is first reviewed in detail. It is shown that the Fisher Information is a more convenient measure of performance than the asymptotic variance, due to its convexity and other useful properties; conditions are derived for the existence of a minimax (in terms of Fisher Information) robust estimator of a location parameter.

Although Huber's method is batch, in the sense that the entire sample of past observations is needed at all time, an asymptotically equivalent recursive version of the robust estimator of location can be derived based on the stochastic approximation technique introduced by Robbins and Monro. The properties of such recursions are investigated, and proofs of consistency, asymptotic normality, and asymptotic efficiency are reviewed in detail.

These results are extended to the case where the "location parameter" varies according to a deterministic linear model. It is shown that this corresponds to estimating the state of a linear dynamic system when there is no process noise, and that the above asymptotic properties hold here as well, under relatively mild conditions.

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.

Using an asymptotic expansion around 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 Gaussian 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.

Because the derivation of a least favorable distribution in this case remains an open problem, the estimator derived here is not minimax. Several simplifications are proposed to make the estimator easier to use.

The results of a series of simulation experiments are presented, showing that some of the robust recursive estimators in the literature remain very sensitive to heavy-tailed noise. The First-Order estimator derived here 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.

6.2 Future Research Directions

Two principal limitations of the robust recursive estimator derived in this thesis have already been pointed out. Specifically,

(i) Equations (4.320) and (4.411) indicate that the estimator 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 non-robustness 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.

(ii) As discussed in Section 4.4, the least favorable distribution for this problem has not been found. Even if it were, there is no guarantee that the distribution and corresponding estimator would be a saddle point, and thus a solution to the minimax problem. While Approximation 4.4 is simple and appealing, this estimator is strictly speaking not optimal in the minimax sense -- indeed, it is somewhat more conservative. More research is needed to determine whether a minimax solution can be found, and how much better one would be than the estimator of Approximation 4.4.

In addition to the above, the following related problems are suggested as topics for future research:

(iii) *Patchy outliers.* As stated in Section 1.2, the derivation in Section 4.1 makes heavy use of the fact that outliers are rare and isolated. Yet, there are cases (e.g. cracking-grinding ice in the Arctic seas in the problem of signal processing for acoustic surveillance -- Wegman, 1986) where time series contain patchy outliers. One way to extend the present results to cover such outliers is by suitable time-aggregation: there is a considerable literature on the question of time scaling systems subject to wide-band (i.e. only approximately white) noise, including Blankenship and

Borkar (1977), Blankenship and Meyer (1977), and Blankenship and Papanicolaou (1977, 1978). It appears worthwhile to investigate the applicability of such techniques to the estimators proposed here. Besides patchy outliers, such methods might also allow the relaxation of another assumption of Section 1.2, namely the whiteness of the process and observation noises. The question of robustness against weakly colored noise remains relatively understudied at this writing.

(iv) *Process outliers.* As pointed out by Shirazi, Sannomiya and Nishikawa (1988), a robust estimator is expected to behave in quite opposite manners when confronted with process and observation outliers. In the latter case, as this thesis makes clear, the influence of the observation must be bounded and indeed downweighted, in favor of past information accumulated over time. In the former, on the other hand, it is desirable to emphasize the observation, and reduce the influence of past information, since a process noise outlier results in a shift in the state. In other words, the confidence an estimator accords to a large-valued innovation changes according to whether it is due to a process noise outlier or an observation noise outlier. This suggests that estimators resistant to process noise outliers must be constructed in ways very different from those described herein. The assumption that both kinds of outliers are very unlikely to occur together, within some short time period, will probably be necessary if estimators robust against both process and observation outliers are to be derived.

(v) *The continuous-time case.* It does not appear that the robust recursive estimation *via* stochastic approximation ideas have yet been applied to continuous-time systems. Yet, all the theoretical prerequisites seem to exist. The principal difference between the present results and their continuous-time analogue will probably be in Theorem 4.2, where a differential equation version of the conditional prior (in the spirit of Zakai, 1969: see Di Masi and Runggaldier, 1982) may result in a much simpler form. The same goes for discretely sampled continuous-time systems. Finally, it is worth noting that applications of time-scaling to discrete-time systems have been found to yield differential equations (Blankenship, 1981); thus, continuous-time results may be useful in accomodating colored noise as well as patchy outliers.

(vi) *Unknown model parameters.* The approach taken in the present thesis assumes that all model parameters are known, so that only the system state needs to be estimated -- in other words, this is a filtering problem. Yet, in many situations of practical importance, model parameters are unknown and need to be estimated simultaneously. The problem of robust model identification has been studied by Poljak and Tsypkin (1978, 1980), Poor (1986), and others, and can be combined with the filtering problem of this thesis. While model parameters may be estimated by using the residual of the robust estimator in a likelihood function, it is probable that a combined state-parameter estimation scheme will be more fruitful.

(vii) *Fault detection and identification.* Most techniques for detecting unmodeled changes in systems (variously refered to as faults, failures, jumps, etc.) are based on the detection of abrupt model fit degradation. They make heavy use of distributional assumptions, and even those that use the term "robust fault detection" are not *statistically* robust. As stated in Section 1.2, the principal motivation for this thesis was the absence of robust state estimators based on a sufficiently rigorous theoretical foundation to enable the use of their residuals for inference. It is easy to see, by convolving the conditional prior of the state (equation (4.160)) with the noise distribution, that

$$\begin{aligned} p(\underline{z}_n \mid \underline{z}_0, \cdots, \underline{z}_{n-1}) \\ &= (1-\varepsilon)^{\omega+1} \kappa_n \kappa_n^0 \, \mathrm{N}(\underline{z}_n ; \underline{\theta}_n^0, \Gamma_n^0) \\ &\quad + \varepsilon(1-\varepsilon)^{\omega} \kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i \, \mathrm{N}(\underline{z}_n ; \underline{\theta}_n^i, \Gamma_n^i) \\ &\qquad \int \mathrm{N}(\underline{z}_{i-1} - \xi ; \underline{v}_n^i + V_n^i M_n^i \Gamma_n^{i\ -1} (\underline{z}_n - \underline{\theta}_n^i), \\ &\qquad\qquad W_n^i - V_n^i M_n^i \Gamma_n^{i\ -1} M_n^i V_n^{i\,\mathrm{T}}) \, h(\xi) \, d\xi \\ &\quad + O_p(\omega^2 \varepsilon^2). \end{aligned} \qquad (6.1)$$

This expression trivially leads to the conditional distribution of the innovation, which can be utilized for fault detection and identification. Alternatively, a conditional residual can be computed from each conditional estimate, and used for inference.

(viii) *Non-linear models.* The estimates derived in this thesis can be applied to linearized versions of non-linear models, in analogy with the extended Kalman Filter (e.g. see Gelb, 1974, pp.182-190). Alternatively, more sophisticated approaches to non-linear filtering can be developed to recursively estimate the state of a non-linear dynamic system. The most difficult step is likely to be the propagation of the conditional prior distribution (Theorem 4.2), where, if the system can be represented by a continuous-time model, or at least a discretely sampled continuous-time model, Zakai's method may once again yield good results.

References

Abramowitz, M. and I.A. Stegun, eds., (n.d.) *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*, Dover Publications (New York), 9th printing.

Agee, W.S. and B.A. Dunn (1980) "Robust Filtering and Smoothing via Gaussian Mixtures," Tech. Rep. 73, Data Sciences Div., U.S. Army White Sands Missile Range (New Mexico).

Agee, W.S. and R.H. Turner (1979) "Application of Robust Regression to Trajectory Data Reduction," in R.L. Launer and G.N. Wilkinson (eds.), *Robustness in Statistics*, Academic Press (New York), 107-126.

Agee, W.S., R.H. Turner, and J.E. Gomez (1979) "Application of Robust Filtering and Smoothing to Tracking Data," Tech. Rep. 71, Data Sciences Div., U.S. Army White Sands Missile Range (New Mexico).

Akahira, M. and K. Takeuchi (1981) *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*, Springer-Verlag (Berlin and New York).

Alspach, D.L. (1974) "The Use of Gaussian Sum Approximations in Nonlinear Filtering," *Proc. Eighth Annual Princeton Conf. Information Sciences and Systems*, Dept. Electrical Engineering, Princeton University (Princeton, New Jersey), 479-483.

Anderson, B.D.O. and J.B. Moore (1979) *Optimal Filtering*, Prentice-Hall (Englewood Cliffs, New Jersey).

Anderson, B.D.O. and J.B. Moore (1981) "Detectability and Stabilizability of Time-Varying Discrete-Time Linear Systems," *SIAM J. Control Optimization*, 19, 1, 20-32.

Andrews, D.F. (1974) "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523-531.

Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey (1972) *Robust Estimates of Location - Survey and Advances*, Princeton University Press (Prinçeton, New Jersey).

Bachman, G. and L. Narici (1966) *Functional Analysis*, Academic Press (New York).

Bickel, P.J. (1981) "Minimax Estimation of the Mean of a Normal Distribution When the Parameter Space is Restricted," *Ann. Stat.*, 9, 6, 1301-1309.

Bickel, P.J. (1983) "Minimax Estimation of the Mean of a Normal Distribution Subject to Doing Well at a Point." in M.H. Rizvi, J.S. Rustagi, and D. Siegmund (eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, Academic Press (New York and London).

Bickel, P.J. and J.R. Collins (1983) "Minimizing Fisher Information over Mixtures of Distributions," *Sankhyā*, A, 25, 1, 1-19.

Blackwell, D. and M.A. Girshick (1954) *Theory of Games and Statistical Decisions*, John Wiley (New York).

Blankenship, G. (1981) "Singularly Perturbed Difference Equations in Optimal Control Problems," *IEEE Trans. Automatic Control*, AC-26, 4, 911-917.

Blankenship, G. and V. Borkar (1977) "A Robustness Property of the Separation Principle," *Proc. IEEE Conf. Decision and Control* (New Orleans, Louisiana), 135-140.

Blankenship, G. and D. Meyer (1977) "Linear Filtering with Wide-Band Noise Disturbances," *Proc. IEEE Conf. Decision and Control* (New Orleans, Louisiana), 580-584.

Blankenship, G. and G.C. Papanicolaou (1977) "Stability and Control of Stochastic Systems with Wide-Band Noise Disturbances," *Proc. Joint Automatic Control Conf.* (San Francisco, California), 1056 (abstract only).

Blankenship, G. and G.C. Papanicolaou (1978) "Stability and Control of Stochastic Systems with Wide-Band Noise Disturbances," *SIAM J. Appl. Math.*, 34, 3, 437-476.

Block, H.D. (1956) "On Stochastic Approximation," ONR Report PB 134 310, Dept. of Mathematics, Cornell University (Ithaca, New York).

Blum, J.R. (1954a) "Approximation Methods which Converge with Probability One," *Ann. Math. Stat.*, 25, 2, 382-386.

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

Boncelet, C.G. Jr. and B.W. Dickinson (1983) "An Approach to Robust Kalman Filtering," *Proc. 22nd IEEE Conf. Decision and Control* (San Antonio, Texas), vol.1, 304-305.

Brown, L.D. (1971) "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems," *Ann. Math. Stat.*, 42, 855-903.

Bryson, A.E. Jr. and Y.-C. Ho (1975) *Applied Optimal Control: Optimization, Estimation, and Control*, John Wiley (New York).

Burkholder, D.L. (1956) "On a Class of Stochastic Approximation Processes," *Ann. Math. Stat.*, 27, 4, 1044-1059.

Caines, P.E. and D.Q. Mayne (1970) "On the Discrete Time Matrix Riccati Equation of Optimal Control," *Int. J. Control*, 12, 5, 785-194.

Casella, G. and W. Strawderman (1981) "Estimating a Bounded Normal Mean," *Ann. Stat.*, 9, 868-876.

Chernoff, H. (1952) "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *Ann. Math. Stat.*, 23, 493-507.

Chernoff, H. (1972) *Sequential Analysis and Optimal Design*, Society for Industrial and Applied Mathematics (Philadelphia, Pennsylvania).

Chung, K.L. (1954) "On a Stochastic Approximation Method," *Ann. Math. Stat.*, 25, 3, 463-483.

Collins, J.R. (1976) "Robust Estimation of a Location Parameter in the Presence of Assymetry," *Ann. Stat.*, 4, 68-85.

Conway, J.B. (1985) *A Course in Functional Analysis*, Springer-Verlag (Berlin and New York).

Cox, D.R. and D.V. Hinkley (1974) *Theoretical Statistics*, Chapman and Hall (London).

Dahlquist, G. and A. Björk (1974) *Numerical Methods*, Prentice-Hall (Englewood Cliffs, New Jersey).

Derman, C. (1956) "Stochastic Approximation," *Ann. Math. Stat.*, 27, 4, 879-886.

Derman, C. and J. Sacks (1959) "On Dvoretzky's Stochastic Approximation Theorem," *Ann. Math. Stat.*, 30, 2, 601-606.

Deyst, J.J. and C.F. Price (1968) "Conditions for Asymptotic Stability of the Discrete Minimum-Variance Linear Estimator," *IEEE Trans. Automatic Control*, 13, 702-705.

Di Masi, G.B. and W.J. Runggaldier (1982) "On Measure Transformations for Combined Filtering and Parameter Estimation in Discrete Time," *Systems and Control Letters*, 2, 1, 57-62.

Di Masi, G.B., W.J. Runggaldier, and B. Barozzi (1983) "Generalized Finite-Dimensional Filters in Discrete Time," in R.S. Bucy and J.M.F. Moura (eds.), *Nonlinear Stochastic Problems* (Proc. NATO Advanced Study Inst. on Nonlinear Stochastic Problems, Arma ̧cao de Pera, Portugal), 267-277.

Donoho, D.L. (1978) "The Asymptotic Variance Formula and Large-Sample Criteria for the Design of Robust Estimators," unpublished senior thesis, Department of Statistics, Princeton University (Princeton, New Jersey).

Doraiswami, R. (1976) "A Decision Theoretic Approach to Parameter Estimation," *IEEE Trans. Automatic Control*, AC-21, 860-866.

Duncan, D.B. and S.D. Horn (1972) "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *J.A.S.A.*, 67, 340, 815-821.

Dvoretzky, A. (1956) "On Stochastic Approximation," *Proc. Third Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 1, 39-55.

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.

Epling, M.L. (1964) "A Multivariate Stochastic Approximation Procedure," Tech. Rep. 5, Dept. of Statistics, Stanford University (Stanford, California).

Ershov, A.A. (1978a) "Robust Filtering Algorithms," *Automation and Remote Control*, 39, 7, 992-996.

Ershov, A.A. (1978b) "Stable Methods of Estimating Parameters," *Automation and Remote Control*, 39, 8, 1152-1181.

Ershov, A.A. and R.Sh. Lipster (1978) "Robust Kalman Filter in Discrete Time," *Automation and Remote Control*, 39, 3, 359-367.

Evans, J., P. Kersten, and L. Kurz (1976) "Robustized Recursive Estimation with Applications," *Information Sciences*, 11, 69-92.

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

Feller, W. (1966) *An Introduction to Probability Theory and its Applications*, John Wiley (New York).

Gauss, C.F. (1821) "Göttingische gelehrte Anzeigen," in *Werke*, Königlische Gesellschaft der Wissenschaften zu Göttingen, reprinted by Georg Olms Verlag (Hildesheim, 1973), vol.4.

Gebski, V. and D. McNeil (1984) "A Refined Method of Robust Smoothing," *J.A.S.A.*, 79, 387, 616-623.

Gelb, A. (1974) *Applied Optimal Estimation*, M.I.T. Press (Cambridge, Massachusetts).

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.

Gross, A.M. (1977) "Confidence Intervals for Bisquare Regression Estimates," *J.A.S.A.*, 72, 341-354.

Guilbo, E.P. (1979) "Robust Adaptive Stochastic Approximation-Type Algorithms," in A. Niemi (ed.), *A Link Between Science and Applications of Automatic Control* (Proc. Seventh Triennal World Cong. Int. Federation Automatic Control, Helsinki), Pergamon Press (Oxford and New York), vol.3, 2153-2157.

Guttman, I. and D. Peña (1984) "Robust Kalman Filtering and its Applications," Tech. Rep. 2766, Mathematics Research Center, University of Wisconsin-Madison.

Guttman, I. and D. Peña (1985) "Comment: Robust Filtering," *J.A.S.A.*, 80, 389, 91-92.

Hager, W.W. and L.L. Horowitz (1976) "Convergence and Stability Properties of the Discrete Riccati Operator Equation and the Associated Optimal Control and Filtering Problems," *SIAM J. Control Optimization*, 14, 2, 295-312.

Hahn, W. (1963) *Theory and Application of Liapunov's Direct Method*, Prentice-Hall (Englewood Cliffs, New Jersey).

Halmos, P.R. (1964) *Measure Theory*, D. van Nostrand (Princeton, New Jersey).

Hampel, F.R. (1973) "Robust Estimation: a Condensed Partial Survey," *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 27, 87-104.

Hampel, F.R. (1974) "The Influence Curve and its Role in Robust Estimation," *J.A.S.A.*, 69, 383-393.

Harrison, P.J. and C.F. Stevens (1971) "A Bayesian Approach to Short-Term Forecasting," *Operations Research Quarterly*, 22, 4, 341-362.

Harrison, P.J. and C.F. Stevens (1976) "Bayesian Forecasting," *J. Royal Statistical Society*, B, 38, 3, 205-228.

Hewer, G.A., R.D. Martin, and J. Zeh (1987) "Robust Preprocessing for Kalman Filtering of Glint Noise," *IEEE Trans. Aerospace and Electronic Systems*, AES-23, 1, 120-128.

Hodges, J.L. Jr. and E.L. Lehmann (1956) "Two Approximations to the Robbins-Monro Process," *Proc. Third Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 1, 95-104.

Householder, A.S. (1964) *The Theory of Matrices in Numerical Analysis*, Blaisdell Publishing Company (New York).

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

Huber, P.J. (1967) "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 1, 221-233.

Huber, P.J. (1969) *Thèorie de l'Infèrence Statistique Robuste,* Presses de l'Université de Montréal (Montréal).

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

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

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

Jaeckel, L.A. (1971) "Robust Estimates of Location: Symmetry and Assymetric Contamination," *Ann. Math. Stat.*, 42, 3, 1020-1034.

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

Kallianpur, G. (1954) "A Note on the Robbins-Monro Stochastic Approximation Method," *Ann. Math. Stat.*, 25, 2, 386-388.

Kalman, R.E. (1960) "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Engineering -- Trans. ASME*, 82, 35-45.

Kalman, R.E. and R.S. Bucy (1961) "New Results in Linear Filtering and Prediction Theory," *J. Basic Engineering -- Trans. ASME*, 83, 95-108.

Kalman, R.E. and J.E. Bertram (1960) "Control System Analysis and Design via the 'Second Method' of Lyapunov: II. Discrete-Time Systems," *J. Basic Engineering -- Trans. ASME*, 82, 2, 394-400.

Kassam, S.A. and H.V. Poor (1985) "Robust Techniques for Signal Processing: A Survey," *Proc. IEEE*, 73, 3, 433-481.

Kendall, M. and A. Stuart (1977-79) *The Advanced Theory of Statistics*, Macmillan (New York), 4th edition.

Kirlin, R.L. and A. Moghaddamjoo (1986) "Robust Adaptive Kalman Filtering for Systems with

Unknown Step Inputs and Non-Gaussian Measurement Errors," *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-34, 2, 252-263.

Kitagawa, G. (1987) "Non-Gaussian State-Space Modeling of Nonstationary Time Series," *J.A.S.A.*, 82, 400, 1032-1050.

Künsch, H.R. (1986) "Discussion [of Martin and Yohai, 1986]," *Ann. Stat.*, 14, 3, 824-826.

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

La Salle, J. and S. Lefschetz (1961) *Stability by Liapunov's Direct Method, with Applications*, Academic Press (New York and London).

Le Cam, L.M. (1953) "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates," *University of California Publications on Statistics*, University of California Press (Berkeley and Los Angeles), 1, 277-329.

Le Cam, L.M. (1956) "On the Asymptotic Theory of Estimation and Testing Hypotheses," *Proc. Third Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 1, 129-156.

Le Cam, L.M. (1970) "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates," *Ann. Math. Stat.*, 41, 3, 802-828.

Lehmann, E.L. (1952) "On the Existence of Least Favorable Distributions", *Ann. Math. Stat.*, 23, 3, 408-416.

Lehmann, E.L. (1959) *Testing Statistical Hypotheses*, John Wiley (New York).

Levin, I.K. (1980) "Accuracy Analysis of a Robust Filter of a Certain Type by the Method of Convex Hulls," *Automation and Remote Control*, 5, 660-669.

Levit, B. Ya (1979) "On the Theory of the Asymptotic Minimax Property of Second Order," *Theory Probability and its Applications*, 24, 435-437.

Levit, B. Ya (1980) "On Asymptotic Minimax Estimates of the Second Order," *Theory Probability and its Applications*, 25, 552-568.

Loève, M. (1951) "On Almost Sure Convergence," *Proc. Second Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 279-303.

Loève, M. (1963) *Probability Theory*, D. van Nostrand (Princeton, New Jersey).

Mallows, C.L. (1978) "Problem 78-4: Minimizing an Integral," *SIAM Review*, 10, 1, 183.

Mallows, C.L. (1980) "Some Theory of Nonlinear Smoothers," *Ann. Stat.*, 8, 4, 695-715.

Marazzi, A. (1980) "Robust Bayesian Estimation for the Linear Model," Res. Rep. 27, Fachgruppe für Statistik, Eidgenössiche Technische Hochschule, Zürich.

Marazzi, A. (198?) "Robust Bayesian Estimation for the Linear Model," unpub. mss.

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

Martin, R.D. (1979) "Approximate Conditional-Mean Type Smoothers and Interpolators," in Th. Gasser and M. Rosenblatt (eds.), *Smoothing Techniques for Curve Estimation* (Proc. Workshop, Heidelberg), Springer-Verlag (Berlin and New York), 117-143.

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

Martin, R.D. and A.E. Raftery (1987) "Comment: Robustness, Computation, and Non-Euclidean Models," *J.A.S.A.*, 82, 400, 1044-1050.

Martin, R.D. and V.J. Yohai (1986) "Influence Functionals for Time Series," *Ann. Stat.*, 14, 3, 781-818.

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).

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

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.

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.

Mataušek, M.R. and S.S. Stanković (1980) "Robust Real-Time Algorithm for Identification of Non-Linear Time-Varying Systems," *Int. J. Control*, 31, 1, 79-94.

McGarty, T.P. (1975) "Bayesian Outlier Rejection and State Estimation," *IEEE Trans. Automatic Control*, AC-20, 682-687.

Meinhold, R.J. and N.D. Singpurwalla (1983) "Understanding the Kalman Filter," *Amer. Statistician*, 37, 2, 123-127.

Métivier, M. (1982) *Semimartingales*, Walter de Gruyter (Berlin and New York).

Meyr, H. and G. Spies (1984) "The Structure and Performance of Estimators for Real-Time Estimation of Randomly Varying Time Delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-32, 1, 81-94.

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.

Morris, J.M. (1976) "The Kalman Filter: A Robust Estimator for Some Classes of Linear Quadratic Problems," *IEEE Trans. Information Theory*, IT-22, 5, 526-534.

Mosteller, F. and J.W. Tukey (1977) *Data Analysis and Regression*, Addison-Wesley (Reading, Massachusetts).

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.

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

Poljak, B.T. and Ja.Z. Tsypkin (1978) "Robust Identification," in Rajbman (ed.), *Identification and System Parameter Estimation*, North-Holland (Amsterdam and London), 203-224.

Poljak, B.T. and Ja.Z. Tsypkin (1980) "Robust Identification," *Automatica*, 16, 53-63.

Poor, H.V. (1986) "Discussion [of Martin and Yohai, 1986]," *Ann. Stat.*, 14, 3, 829-831.

Prescott, P. (1978) "Selection of Trimming Proportions for Robust Adaptive Trimmed Means," *J.A.S.A.*, 73, 361, 133-140.

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.

Rey, W.J.J. (1983) *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag (Berlin and New York).

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

Royden, H.L. (1968) *Real Analysis*, The Macmillan Company (New York).

Sacks, J. (1958) "Asymptotic Distribution of Stochastic Approximation Procedures," *Ann. Math. Stat.*, 29, 373-405.

Sacks, J. and D. Ylvisaker (1972) "A Note on Huber's Robust Estimation of a Location Parameter," *Ann. Math. Stat.*, 43, 4, 1068-1075.

Schmetterer, L. (1961) "Stochastic Approximation," *Proc. Fourth Berkeley Symp. Mathematical Statistics and Probability*, University of California Press (Berkeley and Los Angeles), 1, 587-609.

Shirazi, M.N., N. Sannomiya, and Y. Nishikawa (1988) "Robust ε-contaminated Gaussian Filtering of Discrete-Time Linear Systems," *Int. J. Control*, 48, 5, 1967-1977.

Sorenson, H.W. and D.L. Alspach (1971) "Recursive Bayesian Estimation Using Gaussian Sums," *Automatica*, 7, 465-479.

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.

Stanković, S.S. and B. Kovacević (1979) "Comparative Analysis of a Class of Robust Real-Time Identification Methods," in R. Isermann (ed.), *Identification and System Parameter Estimation* (Proc. Fifth IFAC Symposium, Darmstadt), vol.1, 763-770.

Stepiński, T. (1982) "Comparative Study of Robust Methods of Vehicle State Estimation," in G.A. Bekey and G.N. Saridis (eds.), *Identification and System Parameter Estimation* (Proc. Sixth IFAC Symposium, Washington, D.C.), vol.1, 829-834.

Stigler, S.M. (1973) "Simon Newcomb, Percy Daniell and the History of Robust Estimation 1885-1920," *J.A.S.A.*, 68, 872-879.

Stockinger, N. and R. Dutter (1983) "Robust Time Series Analysis: An Overview," Res. Rep. 9, Institut für Statistik, Technische Universität Graz.

Titterington, D.M., A.F.M. Smith, and U.E. Makov (1985) *Statistical Analysis of Finite Mixture Distributions*. John Wiley (New York).

Tollet, I.H. (1976) "Robust Forecasting for the Linear Model with Emphasis on Robustness Toward Occasional Outliers," *Proc. IEEE Int. Conf. Cybernetics and Society* (Washington, D.C.), 600-605.

Tsaknakis, H. and P. Papantoni-Kazakos (1988) "Outlier Resistant Filtering and Smoothing," *Information and Computation*, 79, 2, 163-192.

Tsai, C. and L. Kurz (1982) "A Robustized Maximum Entropy Approach to System Identification," in R.F. Drenick and F. Kozin (eds.), *System Modeling and Optimization* (Proc. 10th IFIP Conf., New York), 276-284.

Tsai, C. and L. Kurz (1983) "An Adaptive Robustizing Approach to Kalman Filtering," *Automatica*, 19, 3, 279-288.

Tukey, J.W. (1960) "A Survey of Sampling from Contaminated Distributions," in I. Olkin *et al.* (eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press (Stanford, California).

Tukey, J.W. and D.H. Laughlin (1963) "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1," *Sankhyā*, A, 25, 331-352.

VandeLinde, V.D., R. Doraiswami, and H.O. Yurtseven (1972) "Robust Filtering for Linear Systems," *Proc. IEEE Conf. Decision and Control* (New Orleans, Louisiana), 652-656.

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

Wald, A. (1949) "Note on the Consistency of the Maximum Likelihood Estimate," *Ann. Math. Stat.*, 20, 4, 595-601.

Wald, A. (1950) *Statistical Decision Functions*, John Wiley (New York).

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

Wegman, E.J. (1986) "Discussion [of Martin and Yohai, 1986]," *Ann. Stat.*, 14, 3, 836-837.

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

West, M., P.J. Harrison, and H.S. Migon (1985) "Dynamic Generalized Linear Models and Bayesian Forecasting," *J.A.S.A.*, 80, 389, 73-83.

Willems, J.L. (1970) *Stability Theory of Dynamical Systems*, Nelson (London)

Wolfowitz, J. (1952) "On the Stochastic Approximation Method of Robbins and Monro," *Ann. Math. Stat.*, 23, 3, 457-461.

Wolfowitz, J. (1956) "On Stochastic Approximation Methods," *Ann. Math. Stat.*, 27, 4, 1151-1156.

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

Yurtseven, H.Ö. (1979) "Multistage Robust Filtering for Linear Systems," *Proc. 18th Conf. Decision and Control* (Fort Lauderdale, Florida), 500-501.

Yurtseven, H.Ö. and A.S.C. Sinha (1978) "Two-Stage Exact Robust Filtering for a Single-Input Single-Output System," *Proc. Joint Automatic Control Conf.* (Philadelphia, Pennsylvania), vol.4, 165-173.

Zakai, M. (1969) "On the Optimal Filtering of Diffusion Processes," Z. *Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 11, 230-243.