Robust Kalman Filters Using Generalized Maximum Likelihood-Type Estimators

Mital Arun Gandhi

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Lamine Mili, Chair
Amir Zaghloul
Daniel Stilwell
Joseph Wang
Martin Day

November 20, 2009
Falls Church, Virginia

Keywords: Kalman Filtering, Estimation, Robust Statistics

Robust Kalman Filters Using Generalized Maximum Likelihood-Type Estimators

Mital Arun Gandhi

ABSTRACT

Estimation methods such as the Kalman filter identify best state estimates based on certain optimality criteria using a model of the system and the observations. A common assumption underlying the estimation is that the noise is Gaussian. In practical systems though, one quite frequently encounters thick-tailed, non-Gaussian noise. Statistically, contamination by this type of noise can be seen as inducing outliers among the data and leads to significant degradation in the KF. While many nonlinear methods to cope with non-Gaussian noise exist, a filter that is robust in the presence of outliers and maintains high statistical efficiency is desired. To solve this problem, a new robust Kalman filter framework is proposed that bounds the influence of observation, innovation, and structural outliers in a discrete linear system. This filter is designed to process the observations and predictions together, making it very effective in suppressing multiple outliers. In addition, it consists of a new prewhitening method that incorporates a robust multivariate estimator of location and covariance. Furthermore, the filter provides state estimates that are robust to outliers while maintaining a high statistical efficiency at the Gaussian distribution by applying a generalized maximum likelihood-type (GM) estimator. Finally, the filter incorporates the correct error covariance matrix that is derived using the GM-estimator's influence function.

This dissertation also addresses robust state estimation for systems that follow a broad class of nonlinear models that possess two or more equilibrium points. Tracking state transitions from one equilibrium point to another rapidly and accurately in such models can be a difficult task, and a computationally simple solution is desirable. To that effect, a new robust extended Kalman filter is developed that exploits observational redundancy and the nonlinear weights of the GM-estimator to track the state transitions rapidly and accurately.

Through simulations, the performances of the new filters are analyzed in terms of robustness to multiple outliers and estimation capabilities for the following applications: tracking autonomous systems, enhancing actual speech from cellular phones, and tracking climate transitions. Furthermore, the filters are compared with the state-of-the-art, i.e. the H_∞-filter for tracking an autonomous vehicle and the extended Kalman filter for sensing climate transitions.

Dedication

To my parents, Arun and Smita Gandhi

Acknowledgments

I wish to express sincere appreciation and gratitude to my advisor, Dr. Lamine Mili, for providing me with the scholarly training, direction, and invaluable support and encouragement throughout these years. All of it has been essential to reach this milestone. I also thank Professors Amir Zaghloul, Daniel Stilwell, Joseph Wang, and Martin Day for their time and efforts to be on my committee, and for the perspective that they have provided to my research.

Above all, I am indebted to my parents, Arun and Smita Gandhi, for teaching me to value knowledge and for their unconditional support and love throughout these years. I am also grateful to my elder sister, Sheetal, for her encouraging and inspiring words along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the major objectives of filtering and estimation in the context of statistical signal processing is to extract signals of interest from noise. The actual system is represented by a continuous or discrete linear model via static or dynamic equations with unknown states and certain assumptions on the statistical properties of the signal and noise. For the discrete case, the recursive Kalman filter (KF) is one such method that provides maximum likelihood (ML-) estimates under the following assumptions: (a) the system dynamics and observation functions are linear; (b) a quadratic performance criterion is minimized; and (c) the observation and system process noises, which affect the observations and state predictions respectively, follow Gaussian probability distributions.

In practical systems though, the assumed model is only an approximate one and the two types of noises may be contaminated by unknown thick-tailed, non-Gaussian probability distributions that may induce observation and innovation outliers in the signal. Because classical parametric methods require an exact knowledge of the noise probability distribution, they are not able to suppress these outliers. While filtering in non-Gaussian noise has been an active area of research, a robust and highly efficient method to suppress multiple simultaneously occurring outliers of all types is not available and is the subject of this work. We begin with a description of outlier characteristics in Section 1.1. Classical and modern filtering techniques are then reviewed in Section 1.2, followed by the goal of this research in Section 1.3. A summary of novel contributions is provided in Section

1.4 and organization of this dissertation follows in Section 1.6.

1.1 Outliers in State Space Models and Linear Regression

This work is focused on discrete dynamic models with Gaussian noise that is contaminated by noise processes whose distributions may be unknown, asymmetric, or thick-tailed. One example of such contamination is impulsive noise, a time-varying disturbance that is characterized by relatively large amplitude and short duration spikes, deviates strongly from the white, zero-mean, Gaussian assumption, and is very difficult if not impossible to reject by classical techniques like the KF [64, 116, 150, 151]. Statistically, such contamination of the signal may induce outliers among the data. It can occur through various sources in engineering problems, such as discontinuities from hardware switching in digital control systems [153], faults in the sensors of a control system including target estimation and tracking applications [8, 126], and co-channel interference and fading in wireless communications [15, 114, 116, 151], just to name a few. Furthermore, different sensors have fundamental limitations subject to the associated physical medium, which may lead to outliers. Random electrical noises, typically introduced into the signal via sensors and circuits in the system, may also induce outliers. In cellular phone applications, co-channel and fading interferences and discontinuities from the demodulation process [116, 151] induce impulsive noise [64, 151] and may be a source of outliers in the speech signal. In this case, the impulses are often overlapped over several samples and may even completely corrupt the speech segment, yielding missing data.

1.1.1 Types of Outliers in Linear Regression and Time Series Models

Formally, the occurrence of these outliers can be discussed in the context of time series analysis, regression analysis with independent, identically distributed (i.i.d.) observations, and survey data analysis. We consider the former two in this work; Barnett and Lewis [10] have considered the last case.

A mathematically strict, unique, and generally accepted definition of outliers is not apparent in the literature [59]. Barnett and Lewis [10] defined them as "a patch of observations which appears to be inconsistent with the remainder of that set of data." So, loosely speaking, outliers do not follow the pattern of the majority of the data, perhaps because they are generated by a mechanism other than that of the rest of the data.

Outliers have been classified into various types in the literature [10, 21, 59, 66, 77, 147, 155], such as level change (LC), transient change (TC), variance change (VC), reallocation (RE) and seasonal outliers (SLS). Maronna, Martin, and Yohai [78] indicated two types of outliers in linear regression and time series models, namely isolated and patchy outliers. The former type was introduced by Fox [40, 95] as Type I outliers and affects the observation vector $\mathbf{z}_k$, expressed as

$$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{e}_k, \tag{1.1}$$

directly through the observation noise $\mathbf{e}_k$. The patchy type, coined as Type II outliers by Fox [40, 95], affects the propagated state $\mathbf{x}_k$ and can occur via the process noise $\mathbf{w}_k$ in the system dynamics model given by

$$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{w}_k + \mathbf{u}_k. \tag{1.2}$$

In engineering, these two types are also known as additive outliers and innovation outliers following the works of Masreliez and Martin [85, 84]. However, we call the former observation outliers for the sake of consistency with the type of noise that causes the outlier. Besides these two forms, we also recognize and consider in this work structural outliers that affect $\mathbf{z}_k$ and $\mathbf{x}_k$ through errors in the matrices $\mathbf{H}_k$ and $\mathbf{F}_k$ in (1.1) and (1.2). We also consider outliers that may arise in the control vector $\mathbf{u}_k$ in (1.2). As shown in Table 1.1, the observations, predictions, and filter's error covariance matrix can all be affected by one or more of these outliers.

In the linear regression framework, Hampel [55] described two types of outliers, vertical outliers and bad leverage points, and attributed them to gross errors such as measurement faults and model inadequacy or failure. A vertical outlier is defined as a data point whose projection on the design

Table 1.1: Types of noise, outliers, and affected components of the model.

Types of noises	**Names of outliers in this work**	**Names of outliers in the literature**	**Affected model components**
Observation noise, $\mathbf{e}_k$	Observation outlier	Isolated [79], Type I [40, 95], Additive [85]	$\mathbf{z}_k$
System process noise, $\mathbf{w}_k$, and control vector, $\mathbf{u}_k$	Innovation outlier	Patchy [79], Type II [40, 95], Innovation [85]	$\hat{\mathbf{x}}_{k\|k-1}$
Structural errors in $\mathbf{H}_k$ and $\mathbf{F}_k$	Structural outlier	—	$\mathbf{z}_k$ $\hat{\mathbf{x}}_{k\|k-1}$ $\mathbf{\Sigma}_{k\|k-1}$ $\mathbf{\Sigma}_{k\|k}$

space falls within the bulk of the data whereas a bad leverage point's projection is distant; it has been shown that the latter may cause severe detrimental effects on the maximum likelihood-type (M-) estimators. When the time series signal is treated in the linear regression setting, observation and innovation outliers can now be seen as vertical outliers [121], and structural outliers become bad leverage points. How these three types of outliers affect the linear regression estimator and methods to mitigate these effects are discussed in more detail in Chapter 4; here, suffice it to say that we are interested in a filter that gives robust estimates when one or more of these outliers occur simultaneously.

1.1.2 Distributional Structure for the Outliers

More formally, the concept of qualitative robustness described by Hampel [55] can be used to model the occurrence of outliers in a system. Let F be a probability distribution assumed for majority of the m i.i.d. observations $\{z_1,\ldots,z_m\}$, and G the distribution that the observations actually follow. Hampel [55] proposed to regard an estimator as a system with inputs G or F, and associated outputs L_G or L_F. One can then consider two related probability distribution spaces $N_\epsilon(F)$ and $N_\delta(L_F)$, defined as balls with radii ϵ and δ centered at F and L_F. Each space is endowed with a metric to form a metric space, as depicted in Figure 1.1. An estimator $\hat{x}$, where $\hat{x}$ denotes an estimator of the parameter x, is said to be qualitatively robust if a small deviation between G and

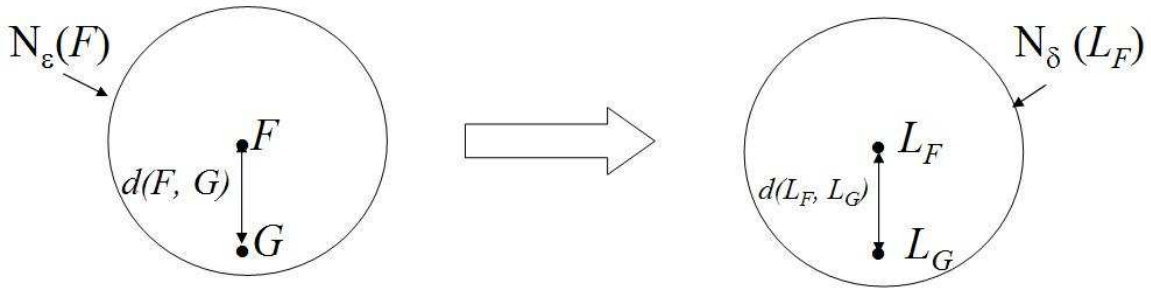


Figure 1.1: The probability distribution spaces $N_\epsilon(F)$ and $N_\delta(L_F)$ are shown, where it is desired that a small deviation between G and F yields a small deviation between L_G and L_F.

F yields a small deviation between L_G and L_F. In other words, if G is in a close neighborhood of F, then L_G remains in a close neighborhood of L_F.

A simpler way to explain the occurrence of outliers without having to define metric spaces is the ϵ-contaminated model, which induces a topological neighborhood around the target distribution F and yields a probability distribution G of the data set as follows [62]:

$$G = (1 - \epsilon)F + \epsilon H, \tag{1.3}$$

where H is an unknown distribution for the outliers. In this work, this model is used to investigate how outliers can be generated via different mechanisms and to measure the goodness of an estimator from a robustness perspective (see Section 3.4). The reader is referred to the work of Becker and Gather [44] for a discussion of some other outlier generating models.

1.2 Literature Review

1.2.1 Classical Filtering Techniques

Next, we briefly discuss some classical and modern techniques used in estimating the states of a dynamic system. Using the system's observations and a linear or nonlinear dynamic model in the continuous or discrete time, a state estimator calculates the best state values in a certain

sense at each time step. The Luenberger estimator and the Kalman filter are two such linear estimators. The former is useful to estimate the state of a system with deterministic noise and a known dynamic model [76]. The method ensures stability and convergence by correcting the current state estimate by an amount proportional to the prediction error, which is the difference between the predicted output and actual observation. If the system model is unknown or time-varying, it must be identified. In this case, the prediction errors can be a result of model identification, state estimation, or both, and complex adaptive estimator designs may be applied.

For a stochastic system with additive observation and system process noises and known model parameters, the most popular technique is attributed to R.E. Kalman from the early 1960s [67, 69]. The recursive Kalman filter is simply the solution to Gauss' least squares estimation problem and builds on the work of Norbert Wiener in estimating the underlying signal from a noisy time series [2, 16, 45, 88, 110]. At each time k, the state vector $\mathbf{x}_k \in \Re^{n\times 1}$ is related to the system's dynamics and the observation vector $\mathbf{z}_k \in \Re^{m\times 1}$ via (1.1) and (1.2). In these equations, $\mathbf{w}_k \in \Re^{n\times 1}$ is the system process noise vector at time k, $\mathbf{e}_k \in \Re^{m\times 1}$ is the observation noise vector at time k, $\mathbf{u}_k \in \Re^{n\times 1}$ is the input control vector at time k, $\mathbf{F}_k \in \Re^{n\times n}$ is the state transition matrix at time k, and $\mathbf{H}_k \in \Re^{m\times n}$ is the observation matrix at time k. In the KF, two fundamental assumptions underlying the characteristics of the system and noise are the following: (a) the system follows a linear Markov process, implying the true state is independent of all earlier states given the immediately previous state and (b) the system and observation noise processes are white, zero-mean, and Gaussian, that is

$$\mathbf{w}_k \sim N[\mathbf{0}, \mathbf{W}_k], \tag{1.4}$$

$$\mathbf{e}_k \sim N[\mathbf{0}, \mathbf{R}_k], \tag{1.5}$$

where $\mathbf{W}_k$ and $\mathbf{R}_k$ are positive definite. For stability and convergence, the filter includes a correction factor in the estimation equations that is obtained using the covariance matrices of the noise. Because the KF is easy to implement, it has been found widespread popularity for many different applications. However, when the fundamental assumptions are violated, the filter may provide

strongly biased solutions or even diverge [69, 133].

1.2.2 Modern Filtering Techniques

Many nonlinear methods have been proposed in the literature to handle non-Gaussian noises and outliers arising via different mechanisms affecting the observation and system processes; for example, see [18, 22, 26, 30, 31, 32, 71, 84, 85]. In 1970, Bucy proposed one of the earliest nonlinear filters [18] to handle non-Gaussian noises. However, this filter is computationally intensive with increasing order of state variables and assumes the noise probability distribution is known *a priori*. In the mid-1970s, Masreliez and Martin [84, 85] pioneered the application of robust statistics [55, 62, 78, 121] to handle symmetric, ϵ-contaminated Gaussian noise in $\mathbf{e}_k$ and $\mathbf{w}_k$ by means of separate filters and stochastic approximation.

Since then, many methods have been proposed to handle observation outliers, namely Christensen and Soliman's filter using the least absolute value criterion [22]; Doblinger's adaptive KF scheme [26]; and filters by Durovic, Durgaprasad, and Kovacevic [30, 31, 71] utilizing the M-estimators. However, these methods do not iterate at each time step when solving the underlying nonlinear estimator, implying that the predictions are assumed to be accurate and are used to suppress observation outliers that deviate from them. As a result, when innovation and observation outliers occur simultaneously, these filters yield unreliable results. On the other hand, assuming the observations are accurate would also lead to erroneous estimates. Hence, a filter is needed that does not rely completely on either the predictions or the observations; instead, it should process them simultaneously via an iterative solution for the underlying estimator. Finally, the classical KF error covariance matrix has been inaccurately retained in these filters. The only exception is the method proposed by Durovic and Kovacevic [31], which uses the covariance matrix for M-estimators from Huber [62]. In general, this matrix needs to be replaced by that corresponding to the underlying nonlinear estimator.

Yet another method in the time domain to suppress outliers is the moving median filter, a technique that is quite popular in speech enhancement applications [73, 150]. The method simply

replaces the point in the center of a window by the sample median of all points within that window. But, for it to be effective, the window must be at least twice as long as the corrupted segment. So, when outliers occur sequentially over several samples, the filter yields degraded estimates. Furthermore, correlation among the speech samples lead to poor results. To account for this correlation, some works in the literature [41, 103] have suggested using the Kalman filter. However, the KF may not estimate the model parameters well in the presence of outliers.

Finally, frequency domain methods have also been suggested to suppress noise. One such popular method used in speech processing is the so-called spectral subtraction technique [6, 12, 75, 100]. However, these methods are generally ineffective against outliers since the signal's spectral content is highly altered by the outliers, as seen for real speech in Section 6.7.

In contrast to methods dealing with arbitrarily large outliers, the H_∞-filter is a technique from robust control that by design may accommodate modeling errors and uncertainties due to unknown-but-bounded noise [29, 50, 51, 52, 131, 133, 132, 152]. This filter is reviewed in further detail in Chapter 2, where we will see that its robustness is complementary to the one proposed in this dissertation because the H_∞-filter minimizes worst-case estimation error but does not handle well outliers.

1.3 Research Objective

Detecting and suppressing multiple simultaneously occurring observation, innovation, and structural outliers is a challenging and difficult problem, one for which a robust and efficient solution is not available in the literature. Indeed, applying the classical Kalman filter and many of the modern filters described in Section 1.2.2 may yield strongly inaccurate results. As stated in Section 1.1.2, the outliers can be described by means of the ϵ-contaminated model as points that deviate from a target Gaussian distribution due to a contaminant following an arbitrary distribution H. Hence, the mechanism generating the outliers in a given system is unknown in general, and therefore, an optimal estimator design using a maximum likelihood approach is not possible. Furthermore, the

class of M-estimators cannot be used either as they are not robust to structural outliers.

The goal in this research is to develop a robust state estimator that is able to handle any of the three types of outliers. Three measures of such robustness are the maximum bias, breakdown point, and influence function [62]. An estimator is qualitatively robust if the maximum possible bias b_{max} is bounded when the sample is contaminated by at least one outlier, where bias is defined as the difference between the parameter's true value and its estimate. The breakdown point of an estimator is the maximum fraction of outliers for which the estimator has a bounded bias. Thus, an estimator is considered to be robust if it has a bounded bias under contamination, yielding a positive breakdown point. Finally, the effects of an infinitesimal contamination on the estimator $\hat{x}$ at a distribution F is given by the influence function. As discussed in Section 3.4.4, the maximum bias curve provides an integrated assessment of these measures with bias under contamination.

Besides being robust, the filter should also be a good estimator in classical statistical terms, characterized by properties of consistency, unbiasedness, rate of convergence, and efficiency. First, the estimator should converge towards the true value of the parameter to be estimated when the number of measurements increases to infinity. This property is called Fisher consistency. Second, a good estimator should have a fast rate of convergence towards the true value. Third, the estimator should be unbiased, i.e. its mean value should be equal to the true value for any sample size. Fourth, the variance of the estimates should be in the vicinity of the Crámer-Rao lower bound at the assumed parametric model. When the lower bound is attained, the estimator is said to be efficient at that model. In summary, we are interested in a filter that has a positive breakdown point (robust) and continues to maintain good performance for additive, zero-mean, Gaussian observation noise (highly efficient).

1.4 Summary of Novel Contributions

To achieve our objective, we initiate a new broad class of filters that are of a maximum likelihood-type and are robust to all types of outliers. This class, which includes the KF as a particular case,

is casted within a general linear regression framework that allows us to make use of any robust estimator whose covariance matrix can be derived. Specifically, the three key steps of the approach are as follows: (a) create a redundant observation vector, (b) perform robust prewhitening, and (c) solve the underlying estimator. These steps are described in detail in Sections 4.4, 4.7, and 4.8.4, respectively, and the new scheme is summarized in Section 4.9. Note that observation redundancy is required for an estimator to be capable of suppressing the outliers, i.e. have a positive breakdown point, and can be achieved in practice by simply placing more sensors in the system. To process the observations together, we convert the classical recursive filter into a batch-mode linear regression form in the first step of the generalized maximum likelihood-type (GM-) Kalman filter (GM-KF).

The second contribution of this research is the second step of the GM-KF: a new prewhitening procedure to robustly uncorrelate the noise when outliers are present in the predictions and observations. The procedure utilizes a robust estimator of location and covariance, such as the Projection Statistics (PS), to identify and down-weight the outliers before prewhitening the data set; as discussed in Sections 4.7 and 6.1, the method helps achieve robustness while maintaining high statistical efficiency in the state estimates.

The third contribution of this work is the use of the GM-estimator in the final step of the proposed filter, resulting in a method that is robust to both vertical outliers and bad leverage points in the linear regression framework. In practice, this means that the GM-KF can suppress all three types of outliers given sufficient redundancy in the observations. The unconstrained nonlinear optimization in the GM-estimator has been solved using the Iteratively Re-weighted Least Squares (IRLS) algorithm, whose convergence rate is also derived in Section 5.4.

The fourth contribution of this work is the development of a new state estimation error covariance matrix required in the GM-KF. This asymptotic matrix is derived using its relationship to the influence function evaluated at the same model. Following the work of Hampel [55] and Fernholz [37], the associated derivations are given in Sections 5.1 - 5.3.

The fifth contribution of this research is a new filter that is applicable to systems undergoing nonlinear dynamics with one or more stable equilibrium points. As seen in Section 7.6, we develop

this filter by applying a robust prewhitening and estimation procedure in the extended Kalman filter (EKF) methodology. The resulting technique, namely the GM-EKF, is able to suppress outliers and accurately sense any shifts in the states of a nonlinear system between the equilibrium points. We have also derived the influence function of the nonlinear GM-estimator in Section 7.7, useful to obtain the error covariance matrix of the GM-EKF.

1.5 Applications and Results

The performance of the GM-KF is demonstrated in Chapter 6 in terms of improved robustness against outliers while maintaining high statistical efficiency under Gaussian noise. In particular, through simulations of various engineering applications, we show that the GM-KF has a breakdown point that is no larger than 33% and is able to suppress multiple, concomitant outliers. For details, the reader is referred to Sections 6.1 - 6.6. In addition, it is shown in Section 6.1 that the GM-KF achieves a statistical efficiency of 95% asymptotically with appropriately chosen parameters.

The first application we consider is tracking autonomous systems using global positioning satellite data links. In Section 6.3, the filter is applied on a model to track an unmanned ground vehicle moving in a two-dimensional terrain. In Section 6.4, the filter is used to follow an unmanned aerial vehicle in a search-and-rescue operational scenario. Outliers can be introduced in these models in many ways, including lost data link between the transmitter-receiver pair due to physical limitations and faulty sensors yielding grossly inaccurate measurements. The simulations demonstrate favorable results for the GM-KF in comparison to the H_∞-filter in the presence of various outliers.

Second, we consider a model that represents the dynamic behavior of a helicopter under typical loading and flight conditions at an airspeed of 135 knots. The GM-KF is applied to this model to robustly estimate the helicopter's dynamics, including its horizontal velocity, vertical velocity, pitch rate, and pitch angle, in the presence of all three types of outliers. Details on this application can be found in Section 6.5.

Third, in Section 6.7, we consider a cellular phone application in which actual speech is corrupted

by outliers generated by interference and fading channels. In this case, using the GM-Kalman filter to estimate the parameters of an autoregressive model leads to a new robust Linear Predictive Coding (LPC) implementation that is capable of suppressing outliers in the speech signal.

Finally, we apply the GM-EKF as a simple, computationally efficient, and robust solution to detect and follow climate transitions in a Langevin model, characterized by a double-well potential possessing two stable equilibrium points. Because the time-scale in this model is considered at geological time scales, it is pertinent to be able to sense and accurately track the shifts in the state of this system rapidly and reliably given the model and incoming observations. This is exactly what a GM-EKF is able to achieve, as seen in Section 7.9. By contrast, the traditional extended Kalman filter exhibits poor performance, and other proposed solutions are very computationally intensive with complicated design methodologies.

1.6 Organization of the Dissertation

We discuss the classical recursive Kalman filter, its characteristics and weaknesses, and other filtering techniques in more detail in Chapter 2. Some concepts from classical and robust statistics are then presented in Chapter 3. In Chapter 4, the robust GM-Kalman filter is developed. Various statistical properties and mathematical results for the GM-KF are derived in Chapter 5. Chapter 6 presents applications and performance results of the GM-Kalman filter. The GM-EKF is developed in Chapter 7 and applied to track climate transitions in a nonlinear model with multiple equilibrium points. Finally, conclusions are drawn and future research paths are outlined in Chapter 8.

Chapter 2

Review of Classical Kalman and Other Filtering Techniques

Statistical signal processing has its roots in the areas of probability, statistics, linear algebra, signals and systems theory, and digital signal processing. In this chapter, we review some classical and modern filtering techniques [2, 63, 128, 144]. Particularly, we study the implementation and characteristics of the Kalman filter, a very popular method first developed in the papers by Kalman [67], and Kalman and Bucy [18]. Then, we discuss the H_∞-filter from robust control literature and other nonlinear methods, including the extended Kalman filter, hidden Markov models, and particle filters.

2.1 Linear Discrete Dynamic Systems

A discrete linear Gauss-Markov system is described by means of a dynamic state equation and an observation equation with conditions on noise and initial values [2, 16, 45, 88]. Let the state of the system be a stochastic vector $\mathbf{x}_k \in \Re^{n\times 1}$. At every time k, suppose that $\mathbf{x}_k$ is observed indirectly via an observation vector $\mathbf{z}_k \in \Re^{m\times 1}$. Let the dynamics of $\mathbf{x}_k$ and $\mathbf{z}_k$ be described for $k \in Z^+$,

where Z^+ is the set of non-negative integers, by the following:

$$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{w}_k + \mathbf{u}_k, \tag{2.1}$$

$$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{e}_k, \tag{2.2}$$

where $\mathbf{w}_k \in \Re^{n\times 1}$ is the system process noise vector; $\mathbf{e}_k \in \Re^{m\times 1}$ is the observation noise vector; $\mathbf{u}_k \in \Re^{n\times 1}$ is the input control vector; $\mathbf{F}_k \in \Re^{n\times n}$ is the state transition matrix; and $\mathbf{H}_k \in \Re^{m\times n}$ is the observation matrix. The system dynamics is assumed to be a Markov process, i.e. the true state is independent of all earlier states given the immediately previous state. The system and observation noises are assumed to be i.i.d., Gaussian random processes, i.e.

$$\mathbf{w}_k \sim N[\mathbf{0}, \mathbf{W}_k], \tag{2.3}$$

$$\mathbf{e}_k \sim N[\mathbf{0}, \mathbf{R}_k], \tag{2.4}$$

where $\mathbf{W}_k$ and $\mathbf{R}_k$ are positive definite covariance matrices. Assuming the noise, observation, control, and state vectors are mutually uncorrelated yields the following relations:

$$E[\mathbf{z}_i\mathbf{e}_j^T] = \mathbf{0} \;\forall\; i,j \tag{2.5}$$

$$E[\mathbf{z}_i\mathbf{w}_j^T] = \mathbf{0} \;\forall\; i,j \tag{2.6}$$

$$E[\mathbf{z}_i\mathbf{u}_j^T] = \mathbf{0} \;\forall\; i,j \tag{2.7}$$

$$E[\mathbf{x}_i\mathbf{e}_j^T] = \mathbf{0} \;\forall\; i \le j \tag{2.8}$$

$$E[\mathbf{x}_i\mathbf{w}_j^T] = \mathbf{0} \;\forall\; i \le j \tag{2.9}$$

$$E[\mathbf{x}_i\mathbf{u}_j^T] = \mathbf{0} \;\forall\; i \le j \tag{2.10}$$

$$E[\mathbf{e}_i\mathbf{w}_j^T] = \mathbf{0} \;\forall\; i,j \tag{2.11}$$

$$E[\mathbf{e}_i\mathbf{u}_j^T] = \mathbf{0} \;\forall\; i,j \tag{2.12}$$

$$E[\mathbf{e}_i\mathbf{e}_j^T] = \mathbf{R}_i\delta_{ij} \;\forall\; i,j \tag{2.13}$$

$$E[\mathbf{w}_i\mathbf{w}_j^T] = \mathbf{W}_i\delta_{ij} \;\forall\; i,j \tag{2.14}$$

$$E[\mathbf{w}_i \mathbf{u}_j^T] = \mathbf{0} \; \forall \; i, j \tag{2.15}$$

$$E[\mathbf{u}_i \mathbf{u}_j^T] = \mathbf{0} \; \forall \; i, j \tag{2.16}$$

Finally, it should be noted that the Kalman filter requires uniform complete observability and uniform complete controllability in the underlying time-varying system. The reader is referred to the work of Kalman [68] for a more complete discussion on these properties.

2.2 The Linear Kalman Filter

In statistical signal processing terminology, smoothers are methods that estimate past states given the preceding model and observations until time k, whereas filters estimate the state vector $\mathbf{x}_k$ given noisy observations until time k. The former includes methods such as Kalman Smoothing, Expectation Propagation, Variational Lower Bounds, Two Filter Smoothing, and Particle Smoothing. Examples of filters include

1. Those assuming Gaussian probability distributions, such as
   (a) Kalman filter;
   (b) Extended Kalman filter;
   (c) Linear Update filter (also known as the unscented Kalman filter).
2. Those assuming a mixture of Gaussian probability distributions, such as
   (a) Assumed Density filter;
   (b) Gaussian Sum filter.
3. Non-parametric online methods, such as
   (a) Histogram filter;
   (b) Particle filter;
   (c) Other variants of the Particle filter.

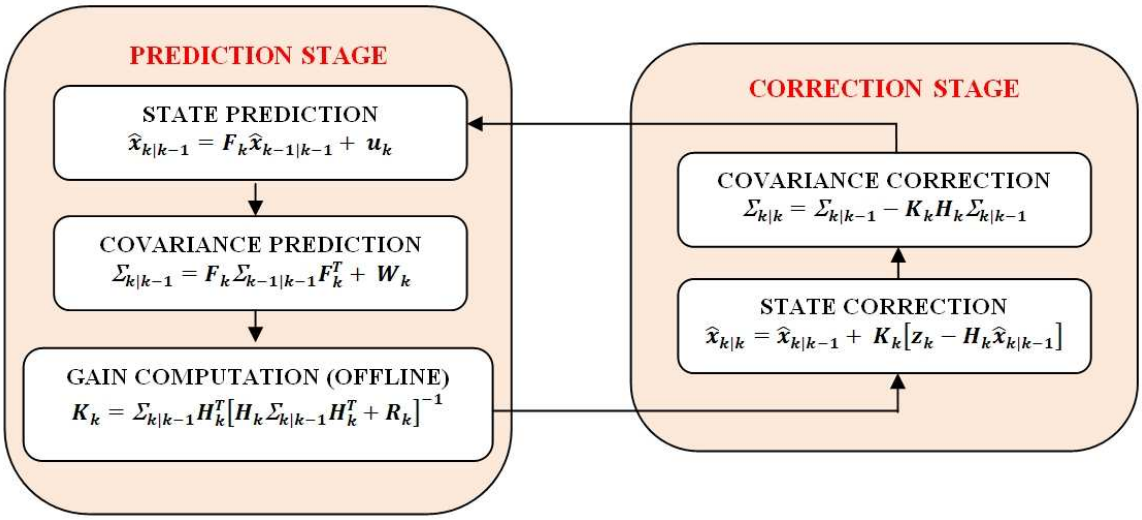


Figure 2.1: Prediction and correction stages of the Kalman filter recursion.

From these methods, we are most interested in the Kalman filter, a classical statistical method that incorporates the least squares estimator developed by Legendre in 1805 and Gauss in 1809 [59, 67]. It was adopted as early as the 1930s by the diagnostic school of thought [112], and established as an optimal solution under Gaussian noise in the context of Tukey's hypothesis testing methods in the 1940s [112]. The filter's recursion equations using a state-space approach [65] are given below, and also derived via a statistical approach in Section 2.2.1:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{u}_k, \tag{2.17}$$

$$\mathbf{\Sigma}_{k|k-1} = \mathbf{F}_k \mathbf{\Sigma}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{W}_k, \tag{2.18}$$

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1} \mathbf{H}_k^T [\mathbf{H}_k \mathbf{\Sigma}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k]^{-1}, \tag{2.19}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k [\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1}], \tag{2.20}$$

$$\mathbf{\Sigma}_{k|k} = \mathbf{\Sigma}_{k|k-1} - \mathbf{K}_k \mathbf{H}_k \mathbf{\Sigma}_{k|k-1}. \tag{2.21}$$

The initial state vector $\mathbf{x}_0$ is assumed to be a normally distributed vector random variable $N(\mathbf{x}_0, \mathbf{\Sigma}_0)$. The initialization of the covariance matrix $\mathbf{\Sigma}_0$ can be arbitrary, as long as it is non-zero, as the filter will eventually converge and "forget" initialization errors [2, 63, 65]. The recursion can be understood in two stages: prediction and correction, described visually in Figure 2.1 and as a

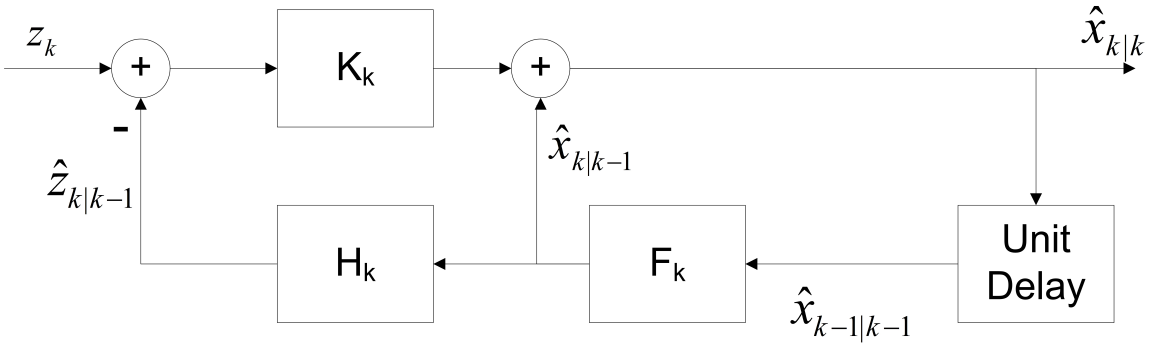


Figure 2.2: Kalman filter recursion in a block diagram.

block diagram in Figure 2.2. Equations 2.17 - 2.18 represent the time update portion of the filter recursion, in which a prediction is made at time k given the information at time $k-1$ for the state and its covariance matrix. Equations 2.19 and 2.21 then correct this state and covariance matrix prediction by using the latest observation at time k. Note that the state transition matrix $\mathbf{F}_k$, the observation matrix $\mathbf{H}_k$, and the noise covariance matrices, $\mathbf{W}_k$ and $\mathbf{R}_k$, may change at each time step, but are assumed to remain constant in this work.

2.2.1 Kalman Filter as a Bayesian Statistical Estimator

An understanding of the KF is not complete without highlighting its statistical nature. The recursion in (2.1) and (2.2) represents the standard innovations form of the filter [2, 65]. It can also be seen as a recursive Bayesian estimator that is optimal in a minimum mean-squared sense at the Gaussian distribution. Let us denote by $\mathbf{X}_k \equiv \{\mathbf{x}_0, \cdots, \mathbf{x}_k\}$ and $\mathbf{Z}_k \equiv \{\mathbf{z}_0, \cdots, \mathbf{z}_k\}$. The objective is to infer the posterior probability density function (PDF) of the state vector given the observations, expressed as

$$p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{Z}_{k-1}). \tag{2.22}$$

Assuming a linear Gauss-Markov system, it can be shown that the conditional PDFs at each time instant are Gaussian, and can be represented as follows:

$$\hat{\mathbf{x}}_{k|k-1} = E[\mathbf{x}_k|\mathbf{Z}_{k-1}] \tag{2.23}$$

$$\hat{\mathbf{x}}_{k|k} = E[\mathbf{x}_k|\mathbf{Z}_k] \tag{2.24}$$

$$\boldsymbol{\Sigma}_{k|k-1} = E[(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^T|\mathbf{Z}_{k-1}] \tag{2.25}$$

$$\boldsymbol{\Sigma}_{k|k} = E[(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k})^T|\mathbf{Z}_k] \tag{2.26}$$

By virtue of the Gauss-Markov theorem [137], the conditional mean is dependent only on the latest observation, i.e. $\hat{\mathbf{x}}_{k|k} = E[\mathbf{x}_k|\mathbf{z}_k]$, and therefore, the filter can be implemented recursively. Note that the above solution is optimal when $\mathbf{x}_0$, $\mathbf{w}_k$, and $\mathbf{e}_k$ are jointly Gaussian. To obtain the conditional mean, we start with Bayes' rule, expressed for any two random variables $\mathbf{x}$ and $\mathbf{z}$ as

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x}), \tag{2.27}$$

which can be rewritten as

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})}, \tag{2.28}$$

with the probability of $\mathbf{z}$ given by

$$p(\mathbf{z}) = \int p(\mathbf{z}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{2.29}$$

Following (2.28), the desired posterior PDF can then be expressed as

$$\begin{array}{ccccc} p(state|data) & \propto & p(data|state) & \times & p(state) \\ \downarrow & & \downarrow & & \downarrow \\ p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{Z}_{k-1}) & \propto & p(\mathbf{z}_k|\mathbf{x}_k, \mathbf{Z}_{k-1}) & \times & p(\mathbf{x}_k|\mathbf{Z}_{k-1}) \\ \downarrow & & \downarrow & & \downarrow \\ \text{Posterior PDF} & & \text{Likelihood} & & \text{Prior PDF} \end{array}$$

Next, using the information for state $\mathbf{x}_{k-1}$ from its posterior PDF at time $k-1$, i.e.

$$p(\mathbf{x}_{k-1|k-1}|\mathbf{Z}_{k-1}) \sim N(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{\Sigma}_{k-1|k-1}), \tag{2.30}$$

the task is to first predict the state for time k to obtain $\hat{\mathbf{x}}_{k|k-1}$; this estimate is then updated to obtain $\hat{\mathbf{x}}_{k|k}$. For a normally distributed random variable X, the following result from statistics [94] is the key to deriving the prediction equation:

$$X \sim N(\boldsymbol{\mu}, \mathbf{\Sigma}) \Rightarrow \mathbf{C}X \sim N(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\mathbf{\Sigma}\mathbf{C}^T). \tag{2.31}$$

Using this result, the prior PDF at time k is given by

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{Z}_{k-1}) &\sim N(\mathbf{F}_k\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{u}_k, \mathbf{F}_k\mathbf{\Sigma}_{k-1|k-1}\mathbf{F}_k^T + \mathbf{W}_k), && (2.32) \\ &\sim N(\hat{\mathbf{x}}_{k|k-1}, \mathbf{\Sigma}_{k|k-1}). && (2.33) \end{aligned}$$

The state prediction is then the expected value of the prior PDF expressed in (2.33). Next, we need to calculate the desired posterior PDF $p(\mathbf{z}_k|\mathbf{x}_k, \mathbf{Z}_{k-1})$. Let the innovations obtained from the predicted state $\hat{\mathbf{x}}_{k|k-1}$ be denoted by

$$\begin{aligned} \mathbf{i}_{k|k-1} &= \mathbf{z}_k - \hat{\mathbf{z}}_{k|k-1} && (2.34) \\ &= \mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1} && (2.35) \\ &= \mathbf{z}_k - \mathbf{H}_k[\mathbf{F}_k\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{u}_k]. && (2.36) \end{aligned}$$

Since all components in (2.36) are known, observing $\mathbf{z}_k$ is equivalent to observing $\mathbf{i}_{k|k-1}$ leading to

the following equivalent expression:

$$p(\mathbf{x}_k|\mathbf{i}_{k|k-1}, \mathbf{Z}_{k-1}) \propto p(\mathbf{i}_{k|k-1}|\mathbf{x}_k, \mathbf{Z}_{k-1}) \times p(\mathbf{x}_k|\mathbf{Z}_{k-1}), \tag{2.37}$$

where the likelihood is now given by

$$p(\mathbf{i}_{k|k-1}|\mathbf{x}_k, \mathbf{Z}_{k-1}). \tag{2.38}$$

Finally, using Bayes' rule, the desired posterior PDF can be expressed as

$$p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{Z}_{k-1}) = \frac{p(\mathbf{i}_{k|k-1}|\mathbf{x}_k, \mathbf{Z}_{k-1}) \times p(\mathbf{x}_k|\mathbf{Z}_{k-1})}{\int_{all\mathbf{x}_k} p(\mathbf{i}_{k|k-1}, \mathbf{x}_k|\mathbf{Z}_{k-1}) d\mathbf{x}_k}. \tag{2.39}$$

This equation can be reduced to obtain the posterior distribution using additional standard techniques from multivariate statistics, as derived in Meinhold and Singpurwalla [87]. For a normally distributed likelihood function, the resulting posterior PDF as shown in [87] is given by a multivariate normal distribution with mean vector

$$\boldsymbol{\mu} = \hat{\mathbf{x}}_{k|k} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{u}_k + \boldsymbol{\Sigma}_{k|k-1} \mathbf{H}_k^T (\mathbf{R}_k + \mathbf{H}_k \boldsymbol{\Sigma}_{k|k-1} \mathbf{H}_k^T)^{-1} \mathbf{i}_{k|k-1}, \tag{2.40}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{k|k} = \boldsymbol{\Sigma}_{k|k-1} - \boldsymbol{\Sigma}_{k|k-1} \mathbf{H}_k^T (\mathbf{R}_k + \mathbf{H}_k \boldsymbol{\Sigma}_{k|k-1} \mathbf{H}_k^T)^{-1} \mathbf{H}_k \boldsymbol{\Sigma}_{k|k-1}. \tag{2.41}$$

Upon inspection, it is apparent that these expressions resemble exactly the KF implementation given in (2.20) - (2.21).

2.2.2 Characteristics of the Kalman Filter

Several important properties of the Kalman filter have been discussed in the literature [2, 63, 65], of which some are highlighted in this section. It is well-known that the filter is an unbiased and

recursive Bayesian, or maximum a posteriori (MAP), estimator. When $\mathbf{x}_0$, $\mathbf{w}_k$, and $\mathbf{e}_k$ are jointly Gaussian and the mean-squared error is the minimization criterion, the filter estimate $\hat{\mathbf{x}}_{k|k}$ is also an ML-estimate. Even if the Gaussianity assumptions on $\mathbf{x}_0$, $\mathbf{w}_k$, and $\mathbf{v}_k$ do not hold, the filter is still the best affine estimator, or minimum variance filter, i.e. it minimizes the variance of the estimation error among the entire class of linear filters. Furthermore, the KF is computationally fast due to its recursive and linear nature. In fact, $\mathbf{\Sigma}$ is fixed *a priori* and can be computed offline, independent of the actual errors.

But, the Kalman filter and its standard variations have non-negligible drawbacks. First, optimal performance is not attained if accurate estimates of the covariance matrices are not available, the noise process does not follow a Gaussian PDF, or the system model parameters are corrupted by outliers [49, 65, 135]. One suggestion to make the filter more robust to unmodeled errors in the system dynamics is to inflate the noise covariance matrix $\mathbf{W}_k$, which results in a larger gain matrix $\mathbf{K}_k$ through the error covariance matrix $\mathbf{\Sigma}_{k|k-1}$. However, as will be seen in Chapter 6, artificially increasing the gain does not identify and suppress the various types of outliers satisfactorily and the filter's performance degrades. Second, by not iterating in the solution, the filter completely trusts the predictions and observations. As a result, any errors and deviations from assumptions are not captured, causing the least squares estimator underlying the KF to yield strongly biased results in the presence of just a single observation, innovation, or structural outlier. Furthermore, Tukey [148] demonstrated that just two deviating observations among a set of 1000 points are sufficient for L_2-norm to be less efficient than L_1-norm. So again, while optimal at the Gaussian distribution in the mean-squared sense, the KF is not robust or highly efficient under contamination.

Realizing this problem, a lot of efforts have been directed towards researching alternative estimation methods [138, 139]. Even in 1805, Legendre's seminal work on least squares had referenced the need for outlier rejection. Laplace's work on the median was an early publication to consider outliers in a somewhat formal manner [138]. But, the formal introduction of robust statistical theory is attributed to Huber's seminal work in 1964 [62]. In it, Huber proposed to design classes of estimators that may not be optimal under classical assumptions, but whose bias and variance

remain bounded when the assumptions are violated. Many other statisticians have also contributed significantly to the field, including Martin, Maronna, Masreliez, Ronchetti, Rousseeuw, and Yohai in [78, 80, 82, 83, 84, 85, 118, 122]. Several robust and nonlinear filters that include system model uncertainties, non-Gaussian noise, and outlier rejection in their designs were briefly stated in Chapter 1. In the next section, the strengths and weaknesses of some of these methods are observed.

2.3 Robust Filtering Techniques

It was stated in Section 1.2.2 that many modern techniques for robustness have limitations for practical use. For example, Bucy's [18] filter requires perfect knowledge of the noise density and is computationally unattractive with increasing order of states. The separate filters of Masreliez and Martin involved convolution operations [84] and a linear transformation [85] that does not exist in general. Other papers [22, 26, 30, 31, 154] do not iterate in the solution, effectively assuming the predictions are in the vicinity of the desired signal. These methods suppress only observation outliers, leaving the estimator vulnerable to innovation and structural errors. Some of the methods [30, 31, 85, 84] considered non-Gaussian noise, but these assumed the noise does not occur simultaneously in the observation and system process sources. Papers [31] considering symmetric heavy-tailed contamination are also not completely effective, as outliers do not necessarily occur in a symmetric manner. For example, Willsky's approach in 1978 [154] was another step towards robustness, in which additive observation outliers are discarded based on a statistical threshold. But, predictions affected by structural or innovation outliers can still cause erroneous state estimates. Clearly, none of these methods from the statistical community can handle all three types of outliers occurring simultaneously. In the following section, we review a technique that has gained a lot of interest and momentum in the robust control community.

2.3.1 The H_∞-Filter

In contrast to methods dealing with arbitrarily large outliers, the H_∞-filter, an approach stemming from robust control [29, 50, 51, 52, 131, 133, 132, 152], treats system modeling errors and noise uncertainties restricted to unknown-but-bounded type of noise [130, 133]. The following description of the filter follows the work of Simon [133]. Research on the technique was first introduced in the frequency domain by Mike Grimble at the University of Strathclyde in 1987 [133]. Early tutorials on the subject are available in [51, 131] and the reader is also referred to [29, 49, 52, 132, 152] for further details. While the KF minimizes the mean-squared estimation error, or gives the smallest possible sample variance for the estimation error, making it the minimum variance estimator for Gaussian noise and linear minimum variance estimator for non-Gaussian noise terms, the basic premise underlying the H_∞-filtering technique is to minimize the worst-case estimation error. Let a system be given by the the following equations:

$$\mathbf{x}_{k+1} = \mathbf{F}_k\mathbf{x}_k + \mathbf{w}_k, \tag{2.42}$$

$$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{e}_k, \tag{2.43}$$

$$\mathbf{c}_k = \mathbf{L}_k\mathbf{x}_k, \tag{2.44}$$

where $\mathbf{w}_k$ and $\mathbf{e}_k$ are the noise vectors and the filter will estimate the vector $\mathbf{c}_k$. The method can be understood conceptually by observing it as a multiple-input-multiple-output linear time-invariant (LTI) filter that is characterized by an m x n matrix $\mathbf{P}$ of transfer functions, where the component $\mathbf{P}_{ij}$ is a transfer function relating the i-th output to the j-th input [11]. Using this transfer function matrix, the LTI system's input and output power spectra $\mathbf{S}$ are related as

$$\mathbf{S}(f)_{output} = |\mathbf{P}(f)|^2\,\mathbf{S}(f)_{input}, \tag{2.45}$$

where $|\mathbf{P}(f)|^2$ means that each element of the matrix $\mathbf{P}$ is squared. Considering the processes $\mathbf{w}_k$ and $\mathbf{e}_k$ as inputs driving the model given by (2.42) - (2.44), the objective function for the H_∞-filter,

given by

$$J_H = \frac{\sum_{k=0}^{N-1} \parallel \mathbf{c}_k - \hat{\mathbf{c}}_k \parallel^2_{\mathbf{B}_k}}{\parallel \mathbf{x}_0 - \hat{\mathbf{x}}_0 \parallel^2_{\mathbf{\Sigma}_0^{-1}} + \sum_{k=0}^{N-1} (\parallel \mathbf{w}_k \parallel^2_{\mathbf{W}_k^{-1}} + \parallel \mathbf{v}_k \parallel^2_{\mathbf{R}_k^{-1}})}, \tag{2.46}$$

can then be seen as limiting the transfer function norm induced from the exogenous signals at the input to the estimation error at the output. In other words, the filter provides a way to limit the frequency response of the estimator [133] through the transfer function matrix. Note that $\mathbf{\Sigma}_0$, $\mathbf{W}_k$, $\mathbf{R}_k$, and $\mathbf{B}_k$ in (2.46) are symmetric, positive definite matrices chosen by the designer given a priori knowledge of the problem [133].

The filter is designed by minimizing the cost function with respect to $\mathbf{c}_k$ after it is maximized with respect to the initial state $\mathbf{x}_0$ and the noise vectors $\mathbf{w}_k$ and $\mathbf{v}_k$, yielding the following criterion:

$$min_{\mathbf{c}_k} max_{\mathbf{w}_k, \mathbf{v}_k, \mathbf{x}_0} J_H. \tag{2.47}$$

The objective function J_H is developed with the intent to minimize the estimation error as given in the numerator; at the same time, the noise introduced by nature into the system cannot be infinitely large in a brute force manner. If so, the objective function as defined would naturally reduce to 0. Thus, the noise can be arbitrary but bounded.

It turns out that direct minimization of J_H is intractable; the estimator is instead designed to guarantee that the H_∞-norm given by (2.46) is less than a predetermined positive value $\frac{1}{\gamma}$, which can be considered a performance bound or a level of noise attenuation imposed on the signal by the filter, yielding

$$J_H < \frac{1}{\gamma}. \tag{2.48}$$

A game theoretic strategy along with the method of dynamic constrained optimization using La-Grange multipliers is used to design the filter [133]. The first step is to find the optimizing values of $\mathbf{w}_k$ and $\mathbf{x}_0$ that maximize the objective function, subject to the constraint of the model equations.

These values are given by

$$\mathbf{w}_k^* = \mathbf{W}_k \boldsymbol{\lambda}_k^*, \tag{2.49}$$

$$\mathbf{x}_0^* = \hat{\mathbf{x}}_0 + p_0 \boldsymbol{\lambda}_0^*. \tag{2.50}$$

The next step is to find the stationary points with respect to $\hat{\mathbf{x}}_k$ and $\mathbf{z}_k$. Simon [133] has provided the details of this step of the optimization, yielding the following solution:

$$\begin{aligned}
\bar{\mathbf{B}}_k &= \mathbf{L}_k^T \mathbf{B}_k \mathbf{L}_k, && (2.51)\\
\mathbf{K}_k &= \boldsymbol{\Sigma}_k [\mathbf{I} - \gamma \bar{\mathbf{B}}_k \boldsymbol{\Sigma}_k + \mathbf{H}_k^T \mathbf{R}_k^{-1} \mathbf{H}_k \boldsymbol{\Sigma}_k]^{-1} \mathbf{H}_k^T \mathbf{R}_k^{-1}, && (2.52)\\
\hat{\mathbf{x}}_{k+1|k+1} &= \mathbf{F}_k \hat{\mathbf{x}}_{k|k} + \mathbf{F}_k \mathbf{K}_k (\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k}), && (2.53)\\
\boldsymbol{\Sigma}_{k+1|k+1} &= \mathbf{F}_k \boldsymbol{\Sigma}_{k|k} [\mathbf{I} - \gamma \bar{\mathbf{B}}_k \boldsymbol{\Sigma}_{k|k} + \mathbf{H}_k^T \mathbf{R}_k^{-1} \mathbf{H}_k \boldsymbol{\Sigma}_{k|k}]^{-1} \mathbf{F}_k^T + \mathbf{W}_k. && (2.54)
\end{aligned}$$

The following condition must hold true at each time step k for the solution to exist:

$$\boldsymbol{\Sigma}_{k|k}^{-1} - \gamma \bar{\mathbf{B}}_k + \mathbf{H}_k^T \mathbf{R}_k^{-1} \mathbf{H}_k > 0. \tag{2.55}$$

Let us consider some interesting aspects of the method. First, an advantage of the filter lies in its design to handle unknown-but-bounded system model uncertainty. It has been shown [133] that the classical Kalman filter performs better than the H_∞-filter when the noise follows assumptions; however, the latter performs better, for example, when a constant bias exists in the mean of the noise (i.e. standard Kalman filter assumptions are violated), as shown in Figure 2.3.

Second, if $\mathbf{L}_k = \mathbf{I}$ in the H_∞-filter, each element of the state vector will be estimated individually, just like the Kalman filter, instead of a linear combination of the state elements. And if the objective bound $\gamma = 0$, then the H_∞-filter actually reduces to the classical Kalman filter solution. But, while the objective function in the case of Kalman filter is not guaranteed to be bounded, it can be for the H_∞-filter.

Third, if the noise covariance matrices are known, the design matrices $\mathbf{W}_k$ and $\mathbf{R}_k$ are analogous

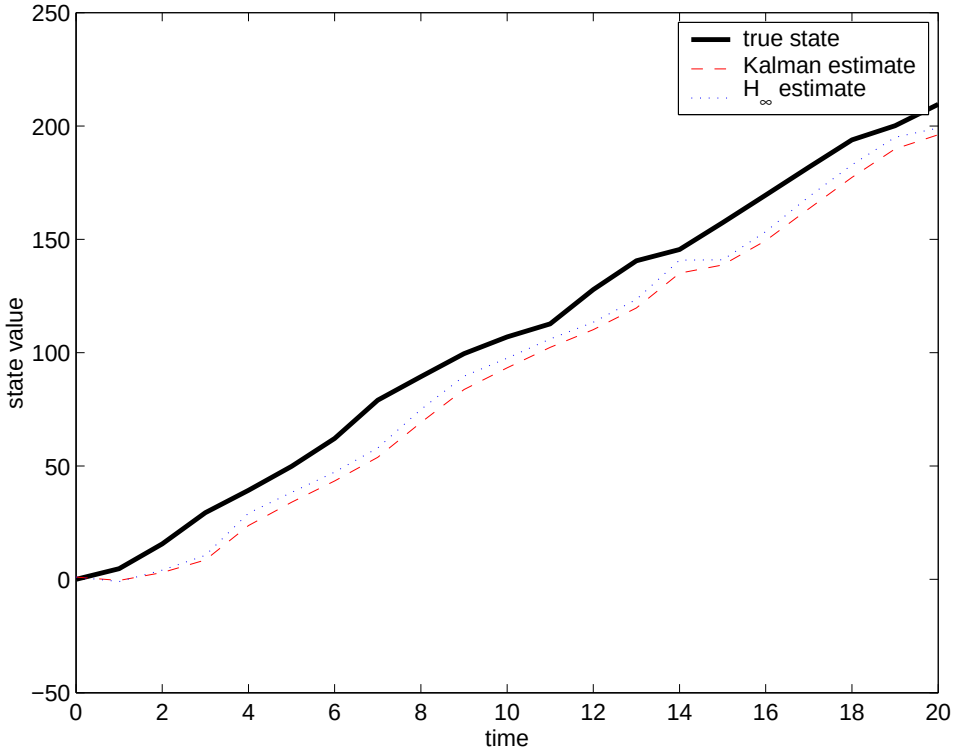


Figure 2.3: The Kalman filter performs better under classical assumptions, but the H_∞-filter has better performance under a constant bias in the noise mean. (Example reproduced following [133] with fair use).

to those in the Kalman filter, as they can be set equal to the noise matrices. If these matrices are unknown, then *a priori* knowledge of the magnitude of $\mathbf{w}_k$, $\mathbf{e}_k$, and the initial estimation error must be used by the designer to assign particular weight values and fine-tune the design matrices for the application [133]. For example, if it is known that a particular state or sensor element has a large disturbance, then that corresponding element in the matrix would be chosen to be large relative to the other values. Or, for example, to obtain a very accurate estimation of a particular element in $\mathbf{c}_k$, the corresponding element in the matrix $\mathbf{B}_k$ would be given a large value compared to the other values in the matrix. The tuning of these design matrices based on the application is an important aspect of the filter, as it clearly leads to the optimizing solution in the first step of the estimation strategy. The filter's sensitivity to the choice of the design matrices $\mathbf{P}_0$, $\mathbf{W}_k$, $\mathbf{R}_k$, and $\mathbf{B}_k$, and the potential need for *a priori* information can be a limitation in its use.

Fourth, Equation 2.55 implies that the quantity $\theta\bar{\mathbf{B}}_k$ be small such that the entire term is positive definite (note that each term is positive definite by itself) [133]. To achieve a small quantity, one of the following must be small: θ, $\mathbf{L}_k$, or $\mathbf{B}_k$. Keeping θ small has the most interesting implication, as that is equivalent to loosening the performance requirement of the filter, or the filter solution may not exist. For large deviations such as those caused by outliers, one may not be able to design a filter strong enough to suppress the effects due to this condition, and certainly not without a priori information to inflate the appropriate elements of the design matrices at the time step corresponding to the occurrence of outliers.

Finally, the filter does not minimize the errors from randomly occurring outliers of any type, only the average worst-case estimation error. Particularly, the covariance of the traditional Kalman filter and the design matrices of the H_∞-filter do not handle well outliers that reside in the thick tails of a noise distribution. As demonstrated in Chapter 6, the H_∞-filter performs poorly in the presence of just one observation or innovation outlier. Thus, the robustness aspects of the H_∞-filter and the proposed GM-KF are complementary, since the former minimizes the maximum estimation error averaged over all samples while the latter is capable of directly suppressing outliers that occur sporadically.

2.4 Other Nonlinear Filtering Techniques

2.4.1 Extended Kalman filter

For the linear system described in Section 2.2, the KF yields the ML-estimates for the states. The more general problem allows nonlinear equations that model practical systems better [88]. The state dynamics and observation equations of the system are given by

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t) + \mathbf{w}_t + \mathbf{u}_t, \tag{2.56}$$

$$\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t) + \mathbf{e}_t, \tag{2.57}$$

where $\mathbf{f}$ and $\mathbf{h}$ are functions of the state vector and are assumed to be continuous and continuously differentiable with respect to all elements of the state vector. Also, the observations in (2.57) are treated as being obtained at discrete intervals. The system process and observation noise covariance matrices are given as

$$E[\mathbf{w}_t \mathbf{w}_\tau^T] = \mathbf{W}_t \ \delta(t-\tau) \tag{2.58}$$

$$E[\mathbf{e}_{t_i} \mathbf{e}_{t_j}^T] = \mathbf{R}_{t_i} \ \delta_{ij}, \tag{2.59}$$

where δ_{ij} is defined as the Kronecker delta function. Thus, the noise processes are assumed to have the same characteristics as the classical Kalman filter: zero mean and uncorrelated with themselves and each other in time. It is important to note that the quantities $\mathbf{W}_t \ \delta(t-\tau)$ and $\mathbf{R}_{t_i} \ \delta_{ij}$ are the covariance matrices while $\mathbf{W}_t$ can be referred to as the intensity matrix [88].

The extended Kalman filter (EKF) is used widely to solve for the state estimates under this system setup [4, 65, 88, 104]. The nonlinear state transition function given in (2.56) is directly applied in the prediction stage. In the correction stage, a linearization is performed on the nonlinear model around the previous corrected estimate to obtain a set of linear perturbation equations [4]. The classical Kalman filter correction equations are then used as a basis for the EKF correction equations. Particularly, let $\mathbf{F}_x$ and $\mathbf{H}_x$ denote the Jacobian matrices of $\mathbf{f}(\mathbf{x}_t)$ and $\mathbf{h}(\mathbf{x}_t)$, respectively.

Formally, we have

$$\mathbf{F}_x = \frac{\partial \mathbf{f}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\hat{\mathbf{x}}_{k-1|k-1}}, \tag{2.60}$$

$$\mathbf{H}_x = \frac{\partial \mathbf{h}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\hat{\mathbf{x}}_{k-1|k-1}}. \tag{2.61}$$

Using these approximations to linearize and discretize the nonlinear continuous-time model yields the following matrices:

$$\mathbf{F}_d = e^{\mathbf{F}_x T_s}, \tag{2.62}$$

$$\mathbf{H}_d = \mathbf{H}_x, \tag{2.63}$$

where $\mathbf{F}_d$ is often called the fundamental matrix and T_s is the integration time step. The following recursive equations can then be written similarly to the classical KF:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_d\hat{\mathbf{x}}_{k-1|k-1} + \int_{t_{k-1}}^{t_k} \mathbf{f}(\hat{\mathbf{x}}_{k-1|k-1})\ dt + \mathbf{B}_d\mathbf{u}_k, \tag{2.64}$$

$$\mathbf{\Sigma}_{k|k-1} = \mathbf{F}_d\mathbf{\Sigma}_{k-1|k-1}\mathbf{F}_d^T + \mathbf{W}_k, \tag{2.65}$$

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1}\mathbf{H}_d^T\ [\mathbf{H}_d\mathbf{\Sigma}_{k|k-1}\mathbf{H}_d^T + \mathbf{R}_k]^{-1}, \tag{2.66}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k[\mathbf{z}_k - \mathbf{h}(\hat{\mathbf{x}}_{k|k-1})], \tag{2.67}$$

$$\mathbf{\Sigma}_{k|k} = \mathbf{\Sigma}_{k|k-1} - \mathbf{K}_k\mathbf{H}_d\mathbf{\Sigma}_{k|k-1}. \tag{2.68}$$

Note that the computations in the EKF cannot be performed off-line due to the linearization performed around the estimates. Also, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge because of the linearization. So, the EKF is considered to be stable if the system is "linear enough" and the filter is initialized well; the more nonlinear a system's behavior, the more accurately the filter initialization needs to be. Besides these heuristic arguments though, rigorous conditions on the boundedness of the errors of the EKF have been analyzed and discussed in [65, 88, 127]. Of particular interest in our work is the underlying least-squares nature of the recursions, which indicates that this filter will also lead to divergence

in the presence of outliers. Therefore, we will revisit these equations in more detail in Chapter 7, where the method is modified to develop a GM-extended Kalman filter.

2.4.2 Hidden Markov Models

Remark: the notation in this section follows the literature for hidden Markov models and is not to be confused with that in the rest of this dissertation.

Many variations have been developed from the basic EKF, such as the second and higher order Taylor series approximations for the nonlinear system, Gaussian sum approach, and Monte-Carlo simulation techniques. Another tool for estimating the desired state PDF is the hidden Markov model (HMM). Rabiner's [113] tutorial on the subject is recommended for the interested reader. Like the KF, the HMM has been applied widely to dynamic systems, with most common use found in the areas of speech recognition and computational linguistic analysis [75, 114]. This model involves an underlying stochastic process that generates a sequence of states Q, with each state emitting an observation according to a second stochastic process O. It is treated as a first-order Markov chain such that the probability is truncated to just the current state and the predecessor state $P(\mathbf{q}_k|\mathbf{q}_{k-1}, \mathbf{q}_{k-2}, \ldots, \mathbf{q}_0) = P(\mathbf{q}_t|\mathbf{q}_{t-1})$. Stationarity is also assumed in the sequence, giving a time-invariant, $n \times n$ matrix $\mathbf{A}$ of transition probabilities. An HMM in completely specified by the following:

- n is the number of states and

  $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ is the set of state vectors;

  $Q = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T\}$ is a sample state vector sequence where $\mathbf{q}_k$ is the state vector at time k.

- m is the number of observation symbols and

  $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$ is the set of observation vectors;

  $O = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ is the sample observation vector sequence where $\mathbf{o}_k$ is the observa-

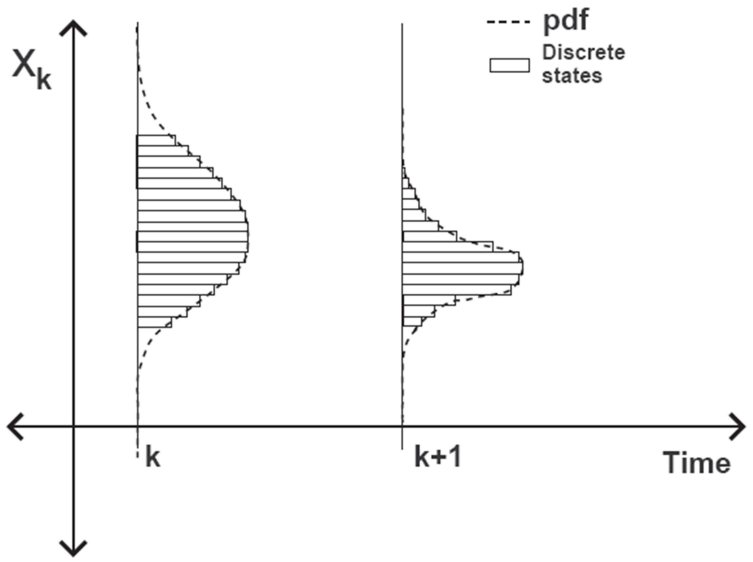


Figure 2.4: Graphical form of a hidden Markov model, representing time-varying probability mass functions for regions of the space estimating the underlying PDF. (Used from [39] with fair use).

tion at time k.

- $\mathbf{A}$ is the state transition probability matrix with elements defined as

  $$A_{ij} = P(\mathbf{q}_{t+1} = \mathbf{s}_j | \mathbf{q}_t = \mathbf{s}_i), \forall\ 1 \leq i, j \leq n \text{ and } A_{ij} \geq 0.$$

- B is the observation probability distribution with elements defined as

  $$b_j(k) = P(\mathbf{o}_t = \mathbf{v}_k | \mathbf{q}_t = \mathbf{s}_j), \forall\ i \leq k \leq m \text{ and } i \leq j \leq n.$$

- π_i is the initial state distribution, defined as

  $$\pi_i = P(\mathbf{q}_1 = \mathbf{s}_i).$$

Using this specification, the tool can be used to solve three types of problems: (1) evaluation, (2) inference, and (3) learning. More information on Problems (1) and (3) can be found in the literature [113, 114]. Problem 2 is of primary interest in the context of this research, as we are interested in determining the optimal state sequence given the observation sequence O and an HMM model $\lambda = (\mathbf{A}, B, \pi_i)$. Solved using the Viterbi algorithm, this method provides optimal

estimates in a conditional mean sense through a discretization of the state space. Particularly, it divides a naturally bounded state space into finite regions with a specific likelihood of being in each of them; furthermore, all states in a given region are represented by a single point. Thus, the PDF is approximated by probability mass functions that give a probability for being in regions of the space, and the entire space is covered by a grid of points that may evolve according to the state equations, as shown in Figure 2.4. Naturally, this approach is only as good as the discrete representation of the actual system. Clearly, the key drawback is that a very large number of hidden states may be needed to represent the state space at the resolution desired or required by a system. In addition, the HMM is not an online algorithm, meaning the most likely state estimate may not be computed recursively at each time step. Indeed, the computations may become quite intractable without the use of efficient algorithms.

2.4.3 Particle Filtering

The Kalman filter, the extended Kalman filter, and other nonlinear methods proposed in the literature estimate the state vector of a dynamic system based on some statistical information summarized by time-dependent PDFs. The integrals describing the evolution of these probability density functions may not have analytic solutions in general. An alternative approach, known as particle filtering, considers the evolution of the whole posterior PDF directly. It was first introduced in the field of automated control theory by Handschin and Mayne [56], and has re-appeared in recent years in part due to the availability of faster computing. Particle filter methods include the condensation algorithm, Bayesian bootstrap, sampling importance resampling filters and other related Monte-Carlo simulation techniques. The reader is referred to many sources in the literature [4, 39, 48, 104] for complete details. Similar to the extended Kalman filter, the particle filtering method requires knowledge of the initial PDF very precisely. More importantly, the algorithm uses an importance function which makes the method very sensitive to outliers [28, 39, 109]. Conceptually, the state vector PDFs in a particle filter are not parameterized. Instead, the entire PDF of the state is approximated by discrete points or particles which evolve according to the

dynamic system equations. It can be shown that as the number of particles increases to infinity, the approximation approaches the true PDF under some well-defined assumptions [28]. This approach can therefore work under non-Gaussian noise, but a very large number of particles may be required for a sufficient representation of the discrete system resulting in intensive computational complexity. Actually, one of the reasons the theory was overlooked until the 1990s was the lack of computational resources.

Even though some advantages can be had in the H_∞-filtering, extended Kalman filtering, hidden Markov modeling, and particle filtering methods, it is clear that a dynamic filter robust to all three types of outliers is not readily available. In Chapter 4, the new GM-Kalman filter is developed that combines the classical Kalman filter with robust estimation concepts, described in the next chapter.

Chapter 3

Properties of Classical and Robust Estimators

In this chapter, concepts from parametric estimation theory, founded by Fisher in the early 1900s [1, 38, 105], are reviewed, beginning with the estimators of location, scale, and scatter. The goodness of such estimators is discussed from classical and robustness perspectives. Finally, classical [38, 55, 62] estimation concepts including the class of ML-estimators are reviewed. These concepts from classical and robust statistics form the basis for the development of the GM-Kalman filter in Chapter 4.

3.1 Basic Estimators

3.1.1 Estimators of Location

Let $Z \equiv \{z_1, z_2, \ldots, z_m\}$ denote a set containing m i.i.d. samples from a random variable, satisfying a univariate model $z = x + e$. The first quantity of interest for this sample set is its location on the real line. One solution for this parameter is given by the sample mean, a least squares estimator

which minimizes

$$J(x) = \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} (z_i - x)^2, \tag{3.1}$$

and is a maximum likelihood and asymptotically efficient estimate at the Gaussian distribution [38]. Also known as the average value, it is the center of gravity of a set of numbers or a distribution. Statistically, it estimates the expected value of the random process, as follows:

$$\mu = E[z] = \int_{-\infty}^{\infty} z f(z)\ dz. \tag{3.2}$$

In the discrete case, it is given by

$$\mu = E[z] = \sum_{i=1}^{m} z_i\ p(z_i), \tag{3.3}$$

and the sample mean is given as

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z_i. \tag{3.4}$$

However, the sample mean is not a robust estimator since a single arbitrarily placed outlier can lead to an arbitrarily biased estimate, yielding a breakdown point of zero. Recall that the breakdown point is a measure that gives the maximum number of outliers under which an estimator gives a finite bounded bias, and is discussed in further detail in Section 3.4.

Another estimator for the location parameter is the estimator of the distribution's median, or its center of *probability*, given as

$$\int_{-\infty}^{med} f(z)\ dz = \int_{med}^{\infty} f(z)\ dz = \frac{1}{2}. \tag{3.5}$$

The sample median is an order statistic. First, the set of numbers is ordered by increasing values. Second, let $\nu = [m/2] + 1$ where $[\cdot]$ represents the integer part. Then, the sample median is given

by

$$\hat{z}_{med} = \begin{cases} z_\nu, & \text{for } m \text{ odd} \\ (z_{\nu-1} + z_\nu)/2, & \text{for } m \text{ even} \end{cases} \tag{3.6}$$

The sample median minimizes the L_1 norm criterion and yields the maximum likelihood estimate at the Laplacian distribution [89]. This estimator is extremely robust, capable of reaching a theoretical maximum breakdown point of 1/2 [78]. Hence, we use it in a projection method in the GM-Kalman filter to derive appropriate weights for the data.

3.1.2 Estimators of Scale

Estimators of scale provide a measure of spread around the location of a sample. There are also location-free estimators of scale. In the one-dimensional case, the classical measure of scale is the standard deviation, denoted by σ and given by the square root of the variance σ^2:

$$\sigma^2 = E[(z-\mu)^2] = \int_{-\infty}^{\infty} (z-\mu)^2 f(z)\, dz. \tag{3.7}$$

In the discrete case, the variance is given by

$$\sigma^2 = \sum_{i=1}^{m} (z_i - \mu)^2\, p(z_i). \tag{3.8}$$

A maximum likelihood estimator of σ^2 at the Gaussian distribution is the sample variance, given by

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^{m} (z_i - \hat{\mu})^2 \tag{3.9}$$

for m i.i.d. samples. It can be shown that this estimator is biased, i.e. $E[\hat{\sigma}^2] \neq \sigma^2$. For unbiasedness, the fractional component of (3.9) should be replaced by $1/(m-1)$ [89]. Though widely used, this estimate of scale is not robust as it incorporates the non-robust sample mean. In fact, the standard deviation does not even exist for some distributions, e.g. Cauchy distribution.

In contrast, the median absolute deviation (MAD), introduced by Gauss around 1800, is a robust estimate of scale and is defined [55] as

$$MAD = 1.4826\ \kappa\ \underset{i}{\text{med}}|z_i - \underset{j}{\text{med}}\ (z_j)|. \tag{3.10}$$

where κ is tabulated for $m \leq 9$, and given by

$$\kappa = \frac{m}{m - 0.8} \tag{3.11}$$

for $m > 9$. Thanks to this correction factor, the estimator is unbiased and Fisher consistent at the Gaussian distribution [120], meaning the MAD reaches the true σ asymptotically. We will use this estimator to standardize the residuals and obtain appropriate weights in the GM-Kalman filter.

3.1.3 Estimators of Scatter

For multivariate samples, the scatter of the samples around the location is measured by a covariance matrix that contains correlation information about the random vector. The actual covariance matrix of a random vector can only be obtained asymptotically. In practice, an $n \times n$ sample covariance matrix given m observations of a vector $\mathbf{z}$ of length $n \times 1$ with zero mean is given by

$$\mathbf{R} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i^T. \tag{3.12}$$

Similarly to the sample mean and sample variance in the one-dimensional case, this estimate is prone to breakdown in the presence of outliers; consequently, a method to calculate the sample covariance matrix robustly has been developed and applied in the GM-Kalman filter.

3.2 Maximum Likelihood Estimation

Parametric estimation theory was introduced by Fisher in his key paper in 1925 [38]. Classical methods estimate the parameters of a model given the fundamental assumption that the cumulative probability distribution is known *a priori*. The estimator is then said to be optimal when the sample set or the system noise follows the assumed distribution exactly. In maximum likelihood estimation, maximizing the conditional probability $p(\mathbf{x}|\mathbf{z})$ is equivalent to maximizing the joint probability $p(\mathbf{z}, \mathbf{x})$. Effectively, the likelihood function, defined as a constant c times the joint density $f(\mathbf{z}; \mathbf{x})$, i.e.

$$l(\mathbf{x}; \mathbf{z}) = c \times f(\mathbf{z}; \mathbf{x}), \tag{3.13}$$

is maximized. Equivalently, for i.i.d. variables, the sum of the negative log of the likelihood function is minimized, as follows:

$$\hat{\mathbf{x}}_{ML} = \min_{\mathbf{x}} \sum_{i=1}^{m} -\ln f(z_i, x_i). \tag{3.14}$$

For the location case, the residual is given by

$$r_i = z_i - \hat{x}_i. \tag{3.15}$$

So, the PDF of z_i can be written as a function of the residuals r_i, as follows:

$$f(z_i; x_i) = f(z_i - \hat{x}_i) = f(r_i). \tag{3.16}$$

It follows that the ML-estimator for the location case is given by

$$\hat{\mathbf{x}}_{ML} = \min_{\mathbf{x}} \sum_{i=1}^{m} -\ln f(r_i). \tag{3.17}$$

In general, the ML-estimator is defined as the minimization of the objective function

$$J(\mathbf{x}) = \sum_{i=1}^{m} -\ln f(r_i) = \sum_{i=1}^{m} \rho(r_i), \tag{3.18}$$

where $\rho(r_i)$ is the general form of the function's kernel. The minimum of $J(\mathbf{x})$ in (3.18) is found by taking the derivative and setting it equal to zero, yielding

$$\frac{dJ(\mathbf{x})}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}\left(\sum_{i=1}^{m}\rho(r_i)\right) \tag{3.19}$$

$$= \sum_{i=1}^{m}\frac{d\rho(r_i)}{dr_i}\frac{dr_i}{dx_i} \tag{3.20}$$

$$= \sum_{i=1}^{m}-\frac{d\rho(r_i)}{dr_i} = \sum_{i=1}^{m}\frac{f'(r_i)}{f(r_i)} = 0. \tag{3.21}$$

Consequently, the ML-estimator is the solution of the implicit equation given by

$$\sum_{i=1}^{m}-\psi(r_i) = 0, \tag{3.22}$$

where $-\psi(r_i) = -d\rho(r_i)/dr_i$ is called the score function, and once again, the ρ-function is specifically $\rho(r_i) = -\ln f(r_i)$. Various ML-estimators corresponding to the most common distributions have been derived in the literature. For example, for the Laplacian distribution expressed as

$$f(r) = \frac{1}{2a}\, e^{-\frac{|r|}{a}}, \tag{3.23}$$

we get the following least absolute value criterion as the objective function to be minimized:

$$J(\mathbf{x}) = \sum_{i=1}^{m}|r_i|. \tag{3.24}$$

Differentiation of this equation yields

$$\frac{dJ(\mathbf{x})}{d\mathbf{x}} = -\sum_{i=1}^{m} sign(r_i) = 0. \tag{3.25}$$

So, the solution must be such that the number of measurements that are smaller is equal to the number of measurements that are larger (than itself). Thus, the ML-estimator for the one-dimensional location parameter at the Laplacian distribution is the sample median. Similarly, it

can be shown that the L_2-norm objective function yields the sample mean as the ML-estimator at the Gaussian distribution.

From the statistical perspective, the ML-estimators have an interesting asymptotic behavior. First, given the first and second derivatives of the log-likelihood function are defined, an ML-estimator is known to be asymptotically normal [96]; as the number of samples increases, its distribution tends to the Gaussian distribution. Second, recall that an ML-estimator is optimal for a given distribution, meaning the bias tends to zero as the number of samples tends to infinity, and asymptotically, no unbiased estimator has a lower MSE than ML-estimators. But, because the distribution H of the outliers (and effectively G) is unknown, we cannot apply this method to obtain optimal state solutions. Instead, we will incorporate robust statistical methods to provide reliability in the presence of outliers. Particularly, robustness theory adds value to parametric estimation by attempting to identify and accommodate data that deviate from the assumptions, as will be seen in the development of the GM-Kalman filter in Chapter 4. Before that, we study various properties from classical and robust statistics that may be used to characterize and assess an estimator's performance.

3.3 Goodness of Estimators from a Classical Perspective

An important question when designing an estimation solution is what makes the estimator good or bad after all? To answer this, we briefly highlight a set of desirable properties [38, 62, 89] from classical and robust perspectives, beginning with the following from the former: consistency, unbiasedness, statistical efficiency, and rate of convergence.

3.3.1 Consistency

For a set of multivariate samples, an estimator is said to be *Fisher consistent* if it tends to the true value of the parameter vector $\mathbf{x}_t$ given infinite number of samples in its support. Formally, this

property is expressed as

$$\lim_{m\to\infty} \mathbf{T}_m = \lim_{m\to\infty} \mathbf{T}(F_m) = \mathbf{T}(F) = \mathbf{x}_t, \tag{3.26}$$

where m is the number of samples used in the estimate, F_m is the empirical distribution with m observations, F is the true distribution, and $\mathbf{T}(F_m)$ is the statistical functional form of the estimator. As an example, the sample mean and sample median are considered to be Fisher consistent if the desired true parameter is the mean or median of the distribution, respectively.

3.3.2 Unbiasedness

The conventional definition of bias is given by

$$b_m = E[\hat{\mathbf{x}}_m] - \mathbf{x}_t. \tag{3.27}$$

Given a sample of m observations, an estimator $\hat{\mathbf{x}}$ is said to be unbiased if

$$E[\hat{\mathbf{x}}_m] = \mathbf{x}_t, \tag{3.28}$$

implying that the sample mean of the estimator be equal to its true mean for any sample size. Suppose a large number n of sets Z, each containing m samples, are drawn and an estimate calculated for each of them; then, the sample mean of all these estimates is taken:

$$\hat{\mathbf{x}}_m = \frac{1}{m}\sum_{i=1}^{m} \hat{\mathbf{x}}_i \tag{3.29}$$

If the estimator is unbiased, then this sample mean will yield the true value and $b_m = 0$. It is well known that the sample mean and the sample median are unbiased estimators [55, 62].

3.3.3 Asymptotic Efficiency of an Estimator

In general, it is desired that the estimator has as small a variance as possible. The minimum possible value is given by the Crámer-Rao lower bound, which is used a benchmark against which the performance of all unbiased estimators is compared. The efficiency of an unbiased estimator is expressed as

$$\xi = \frac{[I_f(x)]^{-1}}{\mathrm{VAR}(\sqrt{m}\hat{x})}, \tag{3.30}$$

where the denominator is a normalized sample covariance [74] and $I_f(x)$ is the Fisher information [38], defined as the amount of information that can be obtained for the parameter vector x from the sample set. It is given by

$$I_f(x) = E\left\{\left[\frac{\partial}{\partial x}\ln f(z;x)\right]^2\right\}. \tag{3.31}$$

A larger Fisher information value at a given parameter vector means that it is easier to distinguish that vector from neighboring ones. Therefore, it can be estimated more accurately [89]. Vice versa, a small value for the Fisher information indicates that the problem being studied is fundamentally difficult.

In the asymptotic case, a consistent estimator is efficient and converges to the minimum possible variance when

$$\lim_{m\to\infty} \mathrm{VAR}\left(\sqrt{m}\hat{x}\right) = [I_f(x)]^{-1}, \tag{3.32}$$

yielding $\xi = 1$. The concept of efficiency can also be used to compare two consistent estimators by computing the ratio of their asymptotic variances. The asymptotic relative efficiency (ARE) is formally stated as

$$\xi_{\hat{\mathbf{x}}_1,\hat{\mathbf{x}}_2} = \lim_{m\to\infty} \frac{\mathrm{VAR}(\sqrt{m}\hat{x}_1)}{\mathrm{VAR}(\sqrt{m}\hat{x}_2)} \tag{3.33}$$

Practically, good efficiency per (3.30) indicates that the estimator is well-suited to estimate a parameter at the assumed distribution. For example, consider the estimators of location. It can be shown that for all m, the sample mean is 36.3% more efficient than the sample median at the Gaussian distribution; but, for the case of a Gaussian mixture distribution, Figure 3.1 shows that

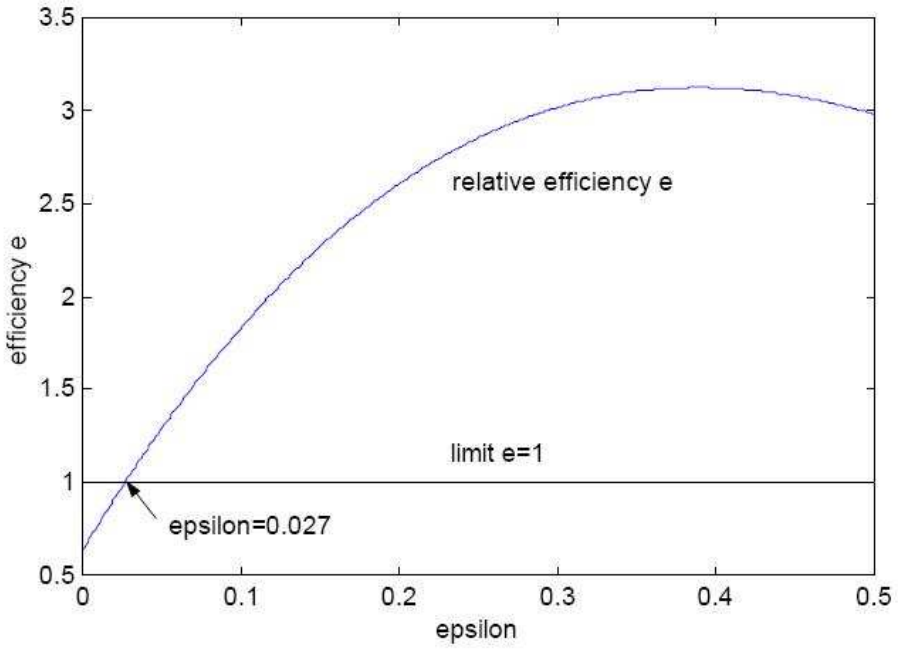


Figure 3.1: Relative efficiency of the sample mean with respect to the sample median at a Gaussian mixture distribution. With increased contamination, the sample median quickly becomes more efficient for $\epsilon \geq 0.027$

the sample median quickly becomes much more efficient that the sample mean. At the Laplacian distribution, the sample median is 50% more efficient than the sample mean.

3.3.4 Rate of Convergence

Another important property for estimators is its rate of convergence to the parameter's true value. For example, the rate for the sample mean is $1/\sqrt{m}$. Some estimators have a much lower rate of convergence, such as the least median of squares (LMS) estimator's rate of $1/m^{1/3}$ [119]. Effectively, this estimator [89] would need 10^6 observations to have an accuracy of 1% whereas the sample mean needs 10^4 observations. Clearly, a fast rate of convergence is desired and will be revisited in the context of Iteratively Reweighted Least Squares algorithm used in the GM-Kalman filter.

3.4 Goodness of Estimators from a Robustness Perspective

While satisfying the preceding properties is a desirable goal, they require knowledge of the probability distribution of the sample in one form or another. Robust statistics allows one to develop techniques that are resistant to deviations from the assumptions, particularly in the case of outliers, such that the estimators' degradation is graceful and bounded. While robustness has been desired for a long time in such situations, it is considered to have been formally introduced by Huber in 1964 and further expanded by Hampel [55, 54, 53] in 1968 by initiating the concepts of infinitesimal robustness, influence function, and asymptotic breakdown point. We now discuss these properties that measure the robustness of an estimator.

3.4.1 Qualitative Robustness

Qualitative robustness measures the effects of small perturbations in the assumptions on the estimator. Let there be a sample set $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ with m observations following a distribution F, and let L_F be the resulting distribution of the estimator. Suppose the data actually follows G, with the resulting distribution of the estimator L_G. An estimator $\hat{\mathbf{x}}$ is said to be qualitatively robust if a small deviation between F and G yields a small deviation between L_F and L_G. In other words, if G is in the close neighborhood of F, then L_G remains in the close neighborhood of L_F. Formally, the estimator [55] is said to be qualitatively robust at G if

$$\forall \delta > 0, \;\; \ni \beta > 0, \text{ such that } \forall m, \; d(F, G) < \beta \rightarrow d(L_G, L_F) < \delta, \tag{3.34}$$

where the function $d(\cdot)$ is a distance measure between two distribution functions. This method requires one to define appropriate metric spaces and distance measures to understand the occurrence of outliers, which can become very complicated. Instead, we use the simpler ϵ-contaminated model to induce a topological neighborhood around F such that the data follows G, as follows [38, 62]:

$$G = (1 - \epsilon)F + \epsilon H. \tag{3.35}$$

where H is an unknown distribution for the outliers.

3.4.2 Local Robustness: Influence Functions

The influence curve, which was later converted into the influence function (IF), provides insight into the local robustness of an estimator. It measures the effects of an infinitesimal contamination in the input data on the estimator $\mathbf{T}$. First, we review the finite sample influence function of the location estimator $\hat{\mathbf{x}}$ at the distribution F. Hampel [54] has defined two versions, one by addition of an observation and one by replacement. Following Tukey [149], suppose m observations from distribution F are observed, $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m-1}, \mathbf{z}\}$, with $\mathbf{z}$ a contaminating point. The finite sample influence function is given by

$$\mathbf{IF}(\mathbf{z}; F) = m\left[\hat{\mathbf{x}}_m(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m-1}, \mathbf{z}) - \hat{\mathbf{x}}_{m-1}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m-1})\right], \tag{3.36}$$

where $\hat{\mathbf{x}}_{m-1}$ is the estimator given $m-1$ points and $\hat{\mathbf{x}}_m$ includes the additional outlying point $\mathbf{z}$. Clearly, the influence function is the effect of that point on the estimator. Tukey's sensitivity curve can also be interpreted as this finite sample influence function.

For asymptotic influence functions [54, 55, 62], it is said that a statistical functional $\mathbf{T}$ is Gâteaux differentiable at F if there exists a linear functional L_F such that for all H,

$$L_F(H - F) = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}(G) - \mathbf{T}(F)}{\epsilon}, \tag{3.37}$$

where $G = (1-\epsilon)F + \epsilon H$ and L_F is the Gâteaux derivative of $\mathbf{T}$ at F. Then, the asymptotic influence function is (3.37) evaluated with the distribution H equal to a point probability unit mass located at $\mathbf{z}$, i.e. $H = \Delta_{\mathbf{z}}$. Under regularity conditions given in [55], the influence function is then given by

$$\mathbf{IF}(\mathbf{z}; F) = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}((1-\epsilon)F + \epsilon\Delta_{\mathbf{z}}) - \mathbf{T}(F)}{\epsilon}, \tag{3.38}$$

which reduces to

$$\mathbf{IF}(\mathbf{z}; F) = \frac{\partial \mathbf{T}(G)}{\partial \epsilon}\Big|_{\epsilon=0}, \tag{3.39}$$

where

$$G = (1-\epsilon)F + \epsilon\Delta_{\mathbf{z}}. \tag{3.40}$$

It describes the local behavior of the estimator in an arbitrarily close neighborhood of F, with the contamination bias expressed as

$$b_c \approx \epsilon|\mathbf{IF}(\mathbf{z})|. \tag{3.41}$$

3.4.3 Gross Error Sensitivity

The influence function can be used to study the gross-error sensitivity (GES) of an estimator $\mathbf{T}$ at a distribution F. This quantity measures the worst possible influence on an estimator by an arbitrary infinitesimal contaminant. Formally, it is the supremum over all $\mathbf{z}$ for which the IF exists and expressed as

$$\gamma^* = sup|\mathbf{IF}(\mathbf{z}; F)|. \tag{3.42}$$

Using this definition, it is easy to relate the contamination bias to the maximum bias for small ϵ, as follows:

$$b_{max} \simeq \epsilon\gamma^*. \tag{3.43}$$

Clearly, it is desirable for an estimator to have a bounded GES, and equivalently, a bounded maximum bias. Such estimators are generally termed B-robust estimators (Bias-robust). On the other hand, an unbounded GES with $\gamma^* = \infty$ means the estimator is completely intolerant to outliers, i.e. a single outlier can ruin the estimator. The sample median has been shown to have the minimum maximum bias curve [62], and thereby, is the most B-robust estimator of location with the smallest possible γ^* [55, 78]. The concept of maximum bias is discussed next.

3.4.4 Global Robustness: Maximum Bias Curve

Over and above the influence function, which assesses the effects of an infinitesimal fraction of contamination [55], the maximum bias curve indicates the upper bound of the bias of an estimator for varying levels of contamination, $0 \leq \epsilon < \epsilon^*$. In other words, the estimator's global robustness to some amount of contamination ϵ in a target distribution F can be analyzed through this curve. For the ϵ-contaminated model given in (3.35), the asymptotic maximum bias of the estimator $\hat{\mathbf{x}}$ in its functional form $\mathbf{T}$ at any such G is defined as

$$b_{max}(G, \mathbf{x}) = max \parallel \mathbf{T}(G) - \mathbf{T}(F) \parallel, \tag{3.44}$$

which reduces to

$$b_{max}(G, \mathbf{x}) = max \parallel \mathbf{T}(G) - \mathbf{x} \parallel, \tag{3.45}$$

assuming the estimator $\mathbf{T}$ is Fisher consistent at F, that is, $\mathbf{T}(F) = \mathbf{x}$ asymptotically [83, 120]. It is desired for an estimator's maximum bias under contamination to be not much larger than the minimum possible maximum bias curve. For example, Huber [62] showed that the sample median attains the minimax bias in the location case. We follow the work of Rousseeuw and Croux [120] to derive the maximum bias curve of an M-estimator of location with a given ρ-function in the one-dimensional case. Consider again the ϵ-contaminated distribution in (3.35) and the model $z = x + e$. Asymptotically, the objective function is given by

$$J(x) = \int_{-\infty}^{\infty} \rho(z - x)\ dG(z), \tag{3.46}$$

whose minimum with respect to x is a root of the following implicit equation:

$$\frac{\partial J(x)}{\partial x} = -\int_{-\infty}^{\infty} \psi(z - x)\ dG(z) = 0. \tag{3.47}$$

To solve for the bias curve, we expand the implicit equation as

$$\begin{aligned}
\frac{\partial J(x)}{\partial x} &= \int_{-\infty}^{\infty} \psi(z-x)\ d[(1-\epsilon)F(z)+\epsilon H(z)] && (3.48)\\
&= \int_{-\infty}^{\infty} \psi(z-x)(1-\epsilon)\ dF(z) + \int_{-\infty}^{\infty} \epsilon\psi(z-x)\ dH(z) && (3.49)\\
&= (1-\epsilon)\int_{-\infty}^{\infty} \psi(z-x)\ dF(z) + \epsilon\int_{-\infty}^{\infty} \psi(z-x)\ dH(z), && (3.50)
\end{aligned}$$

such that the bias curve is a root of the equation

$$(1-\epsilon)E_F[\psi(z-x)] + \epsilon\psi(\infty) = 0. \tag{3.51}$$

To solve for the bias $b(\epsilon, x^k, F)) = x^k - x$, we can apply Newton's method given by

$$\begin{aligned}
x^k &= x^{k-1} + \frac{\partial J(x^{k-1})}{\partial x^{k-1}} \Bigg/ \frac{\partial^2 J(x^{k-1})}{\partial^2 x^{k-1}} && (3.52)\\
&= x^{k-1} + \frac{J'(x^{k-1})}{J''(x^{k-1})}. && (3.53)
\end{aligned}$$

Next, we solve for the component derivatives as follows:

$$\begin{aligned}
\frac{\partial J(x^{k-1})}{\partial x^{k-1}} &= (1-\epsilon)\int_{-\infty}^{\infty} \psi(z-x^{k-1})\ dF(z) + \epsilon\psi(\infty) && (3.54)\\
&= (1-\epsilon)\int_{-\infty}^{\infty} \psi(z-x^{k-1})\ \Phi(z)dz + \epsilon\psi(\infty), && (3.55)
\end{aligned}$$

where the last equation assumes the target distribution $F(z)$ is a Gaussian $\Phi(z)$. Continuing with the second derivative,

$$\begin{aligned}
\frac{\partial^2 J(x^{k-1})}{\partial^2 x^{k-1}} &= \frac{\partial}{\partial x^{k-1}}\left[(1-\epsilon)\int_{-\infty}^{\infty} \psi(z-x^{k-1})\ dF(z) + \epsilon\psi(\infty)\right] && (3.56)\\
&= (1-\epsilon)\int_{-\infty}^{\infty} \frac{\partial}{\partial x^{k-1}}\left[\psi(z-x^{k-1})\right]\ dF(z) + 0 && (3.57)\\
&= -(1-\epsilon)\int_{-\infty}^{\infty} \psi'(z-x^{k-1})\ dF(z). && (3.58)
\end{aligned}$$

Then, using these derivatives in Newton's method, the maximum bias for the M-estimator is given by

$$b(\epsilon; x^k, F) = b(\epsilon; x^{k-1}, F) + \frac{E_G[\psi(x - b(\epsilon, x^{k-1}, F)]}{E_F[\psi']} \tag{3.59}$$

For example, for the sample median next, consider the contamination model given in (3.35) with $H(z) = \Delta_z$. Then, it is straightforward to determine that for $\epsilon = 1/2$, $x_{med} = z$ with $b_{max} = \infty$. Assuming $\epsilon < 1/2$, we then have $G(x_{med}) = (1 - \epsilon)F(x_{med}) = 1/2$. Thus, the maximum bias curve of the sample median is given by

$$b_{max} = |x_{med}| = F^{-1}\left(\frac{1}{2(1-\epsilon)}\right). \tag{3.60}$$

Using (3.59), the asymptotic maximum bias curve of the Huber M-estimate and the sample median in the one-dimensional case is shown in Figure 3.2. This plot makes the connection between many of the estimator's robustness measures, including its qualitative robustness, gross error sensitivity, and breakdown point. Recall that an estimator is qualitatively robust if the maximum possible bias b_{max} is bounded when the sample is contaminated by at least one outlier. It can be determined by the continuity of the curve at $\epsilon = 0$; particularly, the slope of the tangent of the curve at $\epsilon = 0$ is the GES γ^* described previously. Finally, the measure known as breakdown point of the estimator is visible on the curve as the contaminating amount ϵ at which the maximum bias curve has a vertical asymptote. We now discuss the breakdown point in detail.

3.4.5 Global Robustness: Breakdown Point

Besides the influence function and maximum bias curve, the breakdown point, ϵ^*, is another measure to assess the robustness of an estimator. It is defined as the maximum fraction of outliers to which an estimator yields a finite maximum bias under contamination, and gives a measure of the *global* robustness of an estimator. Thus, the breakdown point can be treated as an upper bound on the number of outliers for which the estimator can be considered reliable [55, 121]. Formally, it is given

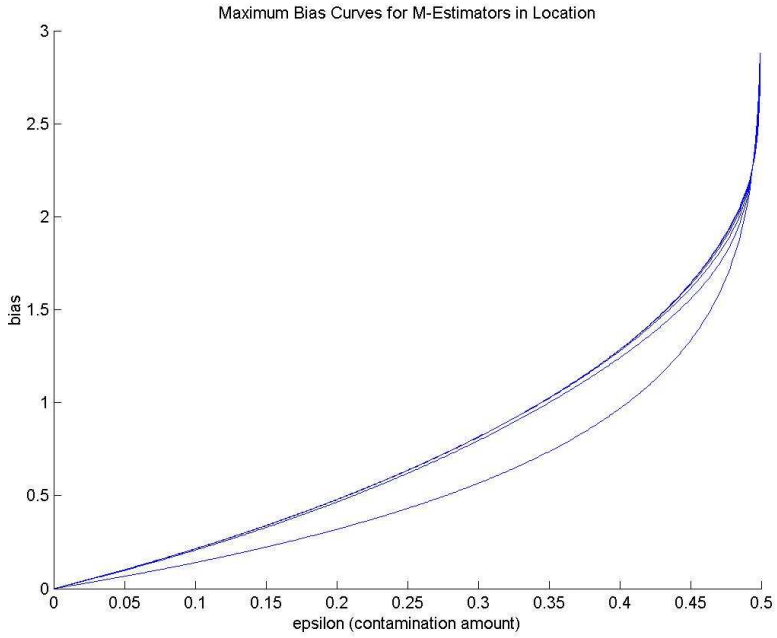


Figure 3.2: Asymptotic maximum bias curves of the k-step M-estimators in location using the Huber function, with vertical asymptote of the curve representing the breakdown point.

by

$$\epsilon^* = max\left\{\epsilon = \frac{f}{m};\ b_{max}\ \text{finite}\right\}, \tag{3.61}$$

where f is the number of contaminant points, m the total number points in the sample $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$, and b_{max} the maximum bias.

Following the work of Rousseeuw and Leroy [121], Mili and Coakley [91] showed that the maximum breakdown point of any regression equivariant estimator $\mathbf{T}$ with m observations and n variables is equal to

$$\epsilon^*_{max} = [(m-n)/2]/m, \tag{3.62}$$

where $[\cdot]$ is the integer part and the estimator satisfies the general position assumption. This property requires that, for a linear regression given by

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{3.63}$$

any set of n row vectors of $\mathbf{H}$ are linearly independent, thus ensuring that any set of n vectors yields a unique solution. The regression is considered to be in reduced position if there exists at least one set of n row vectors of $\mathbf{H}$ that are linearly dependent. Let M be the maximum number of row vectors of $\mathbf{H}$ that lie in an $n-1$ dimensional vector space passing through the origin. Then, the maximum breakdown point of any regression equivariant estimator under this reduced position is no larger than

$$\epsilon^*_{max} \leq [(m - M - 1)/2]/m. \tag{3.64}$$

Under general position, $M = n-1$; thus, (3.64) reduces to (3.62). Note that a location equivariant estimator's maximum breakdown point with m observations and $n = 1$ is given by

$$\epsilon^*_{max} = [(m - 1)/2]/m. \tag{3.65}$$

In general, both bounded influence under contamination and a positive breakdown point are very important for an estimators bias stability. The sample mean can be shown to have a breakdown point $\epsilon^* = 0$ since a single arbitrarily placed outlier can lead to an arbitrarily biased estimate. In contrast, the highest possible asymptotic value for ϵ^* is 0.5 and is reached by the sample median in the location case. In regression, it is attained by the Least Median of Squares and Least Trimmed Squares estimator [119]. In practice, estimators may have a breakdown point much less than 0.5, yet are considered robust because a positive breakdown point is of value to handle at least some outliers. This is the case for GM-estimators used in the proposed GM-Kalman filter in Chapter 4.

Chapter 4

Development of the GM-Kalman Filter

We begin this chapter with a discussion on what to do with undesired outliers in a signal. Then, we study the trade-off between the breakdown point and statistical efficiency of estimators designed to accommodate these outliers. After that, the GM-Kalman filter is developed. To that effect, we motivate the need for observation redundancy and express the filter in a batch-mode regression form to get such redundancy. A new pre-whitening procedure to robustly decorrelate the data in the presence of outliers is then presented. Finally, we solve for the state estimates using the GM-estimator. As will be seen in Section 4.8.3, we use this estimator instead of the M-estimator proposed by Durovic [31] because the latter is robust to observation and innovation outliers only, and not structural ones.

4.1 What To Do with Outliers?

Non-homogeneous observations in a system may be handled using either the diagnostic or accommodation approaches [59, 78]. The former type of methods identify and discard an outlying point from all computations. These methods are widely used [90, 124, 154] because they are easy to ap-

ply in an estimator without changing or adding complexity to the algorithm itself. But, discarding a series of corrupted points may not be a suitable option, especially in critical applications. In addition, these estimators may suffer performance degradation even after discarding the obvious outliers, if only *approximately* Gaussian data remains for processing [78]. In contrast, the accommodation approach systematically down-weights outliers, instead of deleting them [5, 9]. By doing so, one can maintain some level of statistical efficiency while providing robustness. This is especially true for boundary outliers, i.e. points that are just beyond the outlier detection threshold but still close to the bulk of the points. The framework proposed in this work yields filters that follow this accommodation approach.

4.2 Breakdown Point versus Statistical Efficiency

A trade-off exists between the modern robustness concept of breakdown point and the classical Fisherian concept of statistical efficiency of an estimator. Typically, a robust estimator suffers a loss of efficiency at the Gaussian distribution while attaining a positive breakdown point. For example, the sample mean is not robust with a breakdown point of zero, but it is 100% efficient at the Gaussian distribution and its variance attains the Crámer-Rao lower bound. On the other hand, the sample median attains the highest possible breakdown point, but the price is a reduction in statistical efficiency at the Gaussian distribution to 64% as shown by Fisher [38]. Because the classical KF uses the least squares estimator, it suffers from this trade-off also, i.e. it is very efficient but non-robust at the Gaussian distribution.

This robustness-efficiency trade-off needs to be properly addressed when designing the solution to an estimation problem. This is exactly what we have done with the design of the GM-KF. First of all, the broad class of nonlinear GM-estimators with the Huber ρ-function has a positive breakdown point to begin with. Assuming known noise variances and using the maximum bias curve from [78], the maximum possible breakdown point for the estimator can be inferred to be 35%, i.e. $\epsilon^* = 0.35$. Secondly, redundancy is leveraged in the filter by combining the predictions with

the current observations to form the larger data set, which further increases the filter's breakdown point. Finally, unlike the Mallows-type estimators that down-weight both good and bad leverage points, we have used the Schweppe-type method that down-weights only the latter, and therefore, asymptotically attains a 95% statistical efficiency for the Huber ρ-function. In fact, use of the Huber ρ-function yields L_2-type properties for non-outliers, giving good statistical efficiency at the Gaussian distribution, and L_1-type properties for the vertical outliers, bounding their influence on the estimator. Besides, we use the Projection Statistics to derive appropriate weight factors to bound the influence of bad leverage points. Overall, it is clear that the the GM-KF will produce more reliable state estimates in the presence of outliers while also providing statistical efficiency at the Gaussian distribution given the underlying estimator design.

4.3 Need for Redundancy for Positive Breakdown Point

We now develop the GM-KF in complete detail. The first step is to understand the need and benefit of a batch-mode regression form for the Kalman filter in achieving a positive breakdown point. Recall from Chapter 3 that any regression equivariant estimator's maximum breakdown point, under the assumption of general position, is given by

$$\epsilon^*_{max} = [(m-n)/2]/m, \tag{4.1}$$

where there are m observations and n state variables. In the classical recursive Kalman filter, we have $m = n+1$ total observations at each time step k; this is because there are n predictions (one per state variable) and 1 observation collected at each time. Thus, the maximum breakdown point is given by

$$\begin{aligned} \epsilon^*_{max} &= [(n+1-n)/2]/m & (4.2)\\ &= [1/2]/m = 0/m = 0, & (4.3) \end{aligned}$$

i.e. zero breakdown. To achieve a positive breakdown and resilience to outliers, more total observations need to be processed simultaneously. For example, a single redundant measurement gives $m = n + 2$ total observations, with which the maximum breakdown point becomes

$$\begin{aligned} \epsilon^*_{max} &= [(n + 2 - n)/2]/m \quad (4.4) \\ &= [2/2]/m = 1/m, \quad (4.5) \end{aligned}$$

meaning the filter can now handle up to one outlier! Clearly, the estimator is able to handle one more outlier for every 2 observations available; thus, $m_r = m - 1$ redundant observations will give the estimator the ability to handle $m/2$ outliers. For example, to get robustness to 2 simultaneous outliers, i.e. $m/2 = 2$, the estimator requires $m = 4$ total observations, and equivalently, $m_r = 3$ additional measurements. With this redundancy, it follows that

$$\begin{aligned} \epsilon^*_{max} &= [(n + 4 - n)/2]/m \quad (4.6) \\ &= [4/2]/m = 2/m. \quad (4.7) \end{aligned}$$

We need to use a batch-mode linear regression model for the Kalman filter to get this type of observation redundancy, as discussed next.

4.4 Linear Regression Model with Redundancy

To convert the discrete dynamic model given in (2.1) - (2.2) into a batch-mode regression form, we combine the observation equation,

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{e}_k, \quad (4.8)$$

and the following relation between the true state and its prediction,

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{x}_k + \boldsymbol{\delta}_{k|k-1}, \quad (4.9)$$

where $\boldsymbol{\delta}_{k|k-1}$ is the error between the true state and its prediction, to obtain

$$\begin{bmatrix} \mathbf{z}_k \\ \hat{\mathbf{x}}_{k|k-1} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_k \\ \mathbf{I} \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} \mathbf{e}_k \\ \boldsymbol{\delta}_{k|k-1} \end{bmatrix}, \tag{4.10}$$

where $\mathbf{I}$ is the identity matrix. Equation 4.10 is expressed compactly as

$$\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}\mathbf{x}_k + \tilde{\mathbf{e}}_k, \tag{4.11}$$

where

$$\tilde{\mathbf{H}}_k = \begin{bmatrix} \mathbf{H}_k \\ \mathbf{I} \end{bmatrix}, \tag{4.12}$$

and

$$\tilde{\mathbf{e}}_k = \begin{bmatrix} \mathbf{e}_k \\ \boldsymbol{\delta}_{k|k-1} \end{bmatrix}. \tag{4.13}$$

The covariance matrix of the error $\tilde{\mathbf{e}}_k$ is given by

$$\tilde{\mathbf{R}}_k = \begin{bmatrix} \mathbf{R}_k & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{k|k-1} \end{bmatrix}, \tag{4.14}$$

where $\mathbf{R}_k$ is assumed to be the known noise covariance of $\mathbf{e}_k$ and $\boldsymbol{\Sigma}_{k|k-1}$ is the propagated filter error covariance matrix after prediction, given by

$$\boldsymbol{\Sigma}_{k|k-1} = \mathbf{F}_k \boldsymbol{\Sigma}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{W}_k, \tag{4.15}$$

where $\mathbf{W}_k$ is the known noise covariance of $\mathbf{w}_k$.

Thus, we can now process the predictions and all the available observations together via the redundant observation vector $\tilde{\mathbf{z}}$. But, we must be careful to satisfy the assumptions underlying the filter. Particularly, even if the observed data is generated by a process with the target distribution F, small correlation may still exist in the regression of (4.11) given the finite number of samples.

And when outliers are present, the covariance computed in (4.14), using known values of $\mathbf{R}_k$ and $\mathbf{W}_k$, would not represent the actual data well. Therefore, we need to decorrelate the data before solving for the state estimates by pre-whitening it.

4.5 Effects of Classical Pre-Whitening on Outliers

Various pre-whitening methods have been presented in the literature [32, 34]; however, outliers must first be identified and handled, or else, the data would be subject to unintended negative effects. We study next the effects of classical pre-whitening on outliers. For a linear regression model

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{4.16}$$

where

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ h_{m1} & h_{m2} & \dots & h_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_m^T \end{bmatrix}, \tag{4.17}$$

let each $\mathbf{h}_i$ define a point in an n-dimensional space called the design space or factor space, and let $[\mathbf{h}_1, \dots, \mathbf{h}_m]$ be m realizations of a random vector $\mathbf{h}$ following a normal distribution, i.e. $\sim N(\bar{\mathbf{h}}, \mathbf{R})$. The unbiased sample covariance matrix of this finite set of vectors is given by

$$\hat{\mathbf{R}} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{h}_i - \bar{\mathbf{h}})(\mathbf{h}_i - \bar{\mathbf{h}})^T, \tag{4.18}$$

with sample mean,

$$\bar{\mathbf{h}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{h}_i. \tag{4.19}$$

Now, let $m = 50$ and $n = 2$. The ideal covariance matrix for such decorrelated two-dimensional

Gaussian vectors with zero mean is given by

$$\mathbf{R}_{ideal} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.20}$$

However, real data is rarely perfectly decorrelated. For example, consider the following finite set of vectors $\mathbf{h}$:

$$\mathbf{h}_1, \ldots, \mathbf{h}_{50} = \begin{bmatrix} 0.441 & 1.281 & -0.498 & \cdots & 0.808 & 0.041 & -0.756 & -0.089 \\ -0.981 & -0.689 & 1.339 & \cdots & 0.413 & -0.506 & 1.620 & 0.081 \end{bmatrix}. \tag{4.21}$$

The sample covariance matrix is given as

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.0791 & -0.2017 \\ -0.2017 & 0.9820 \end{bmatrix}, \tag{4.22}$$

which yields the 97.5%-confidence ellipse depicted in Figure 4.1, indicating a small amount of correlation and linear dependence in the data. The equation for this error ellipse is given by

$$(\mathbf{h} - \bar{\mathbf{h}})^T \hat{\mathbf{R}}^{-1} (\mathbf{h} - \bar{\mathbf{h}}) = \chi^2_{2,0.975\%}. \tag{4.23}$$

Through pre-whitening, the correlated data can be transformed to produce decorrelated points that resemble the ideal case as closely as possible. In the frequency domain, the spectral components are flattened as linear dependencies are reduced (or removed in perfect decorrelation). The classical pre-whitening method is given as

$$\mathbf{h}_{W_i} = \hat{\mathbf{R}}^{-\frac{1}{2}} (\mathbf{h}_i - \bar{\mathbf{h}}), \tag{4.24}$$

where $\mathbf{h}_{W_i}$ are the desired decorrelated sample vectors. This method gives satisfactory results on data without outliers, as shown in figure 4.2.

Now, consider the same set of vectors, now with 5 outliers (i.e., 10% contamination) that may

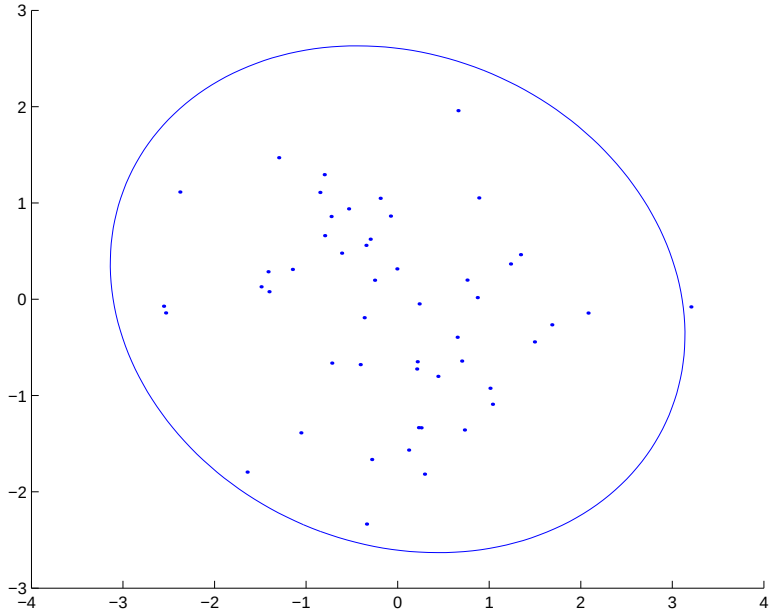


Figure 4.1: 97.5%-confidence ellipse indicating small correlation in the Gaussian data without outliers.

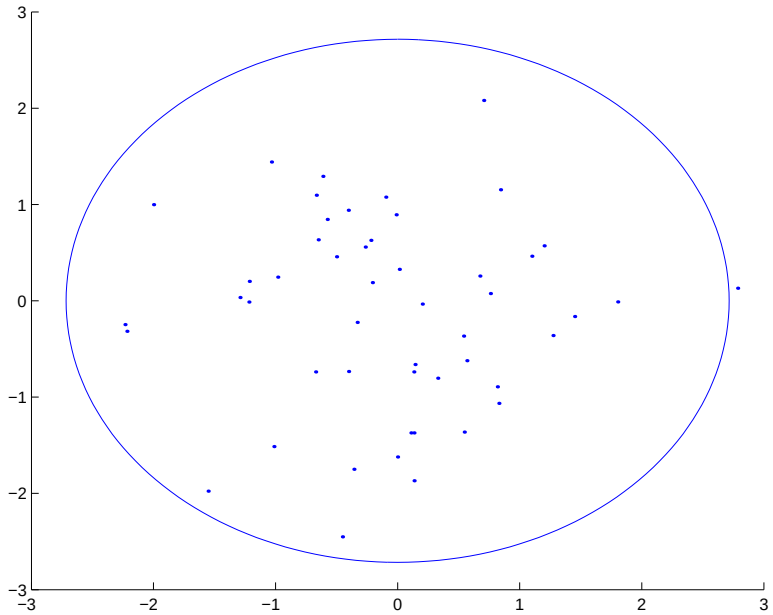


Figure 4.2: 97.5% confidence ellipse after classical pre-whitening of data containing small correlation and no outliers

be induced by faulty sensor observations, hardware faults in demodulation, etc:

$$\mathbf{h}_1, \ldots, \mathbf{h}_{50} = \begin{bmatrix} 0.441 & 1.281 & -0.498 & \cdots & \mathbf{10.808} & \mathbf{10.041} & -0.756 & -0.089 \\ -0.981 & -0.689 & 1.339 & \cdots & \mathbf{9.587} & \mathbf{9.494} & 1.620 & 0.081 \end{bmatrix}. \quad (4.25)$$

The associated covariance matrix is given by

$$\hat{\mathbf{R}} = \begin{bmatrix} 21.1059 & 18.2713 \\ 18.2713 & 17.9013 \end{bmatrix}, \quad (4.26)$$

with the resulting 97.5%-confidence ellipse depicted in Figure 4.3, which shows a significant bias due to the outliers. Clearly, it is not desirable to use such an inaccurately representation of the bulk of the data in the pre-whitening procedure.

So, next, suppose that the original covariance given in (4.22) is known, as assumed in the Kalman filter and the GM-KF. Using this matrix to decorrelate the outlier-contaminated data through classical pre-whitening still does not work properly, as it shifts the outliers artificially closer to the bulk of the points, as shown in Figure 4.4. Thus, we lose information on the true position of the outliers in the original data set, subsequently resulting in an inaccurate down-weighting using any of the distances. Clearly, simply decorrelating outlier-contaminated data using classical pre-whitening is not wise. We need to identify and handle the outliers first. In the next section, we study some techniques to detect the outliers.

4.6 Outlier Detection using Statistical Distance Measures

A wide variety of tools have been developed in the literature for outlier diagnosis, each with its own strengths and weaknesses [7, 25, 33, 111]. Consider again m n-dimensional vectors, $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m\}$. We define an outlier as a point that deviates from a target Gaussian distribution F, using the formal context stated in Chapter 1. To identify such an outlier, one method is

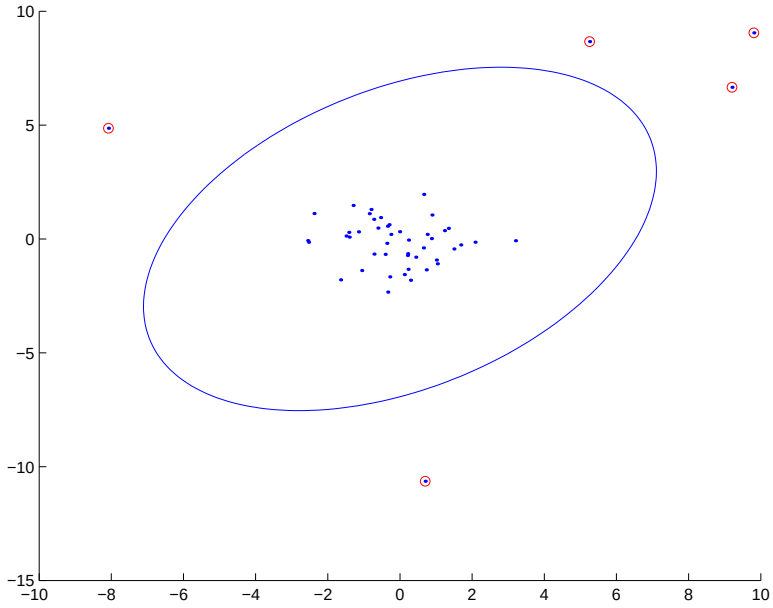


Figure 4.3: 97.5% confidence ellipse for correlated Gaussian data with outliers.

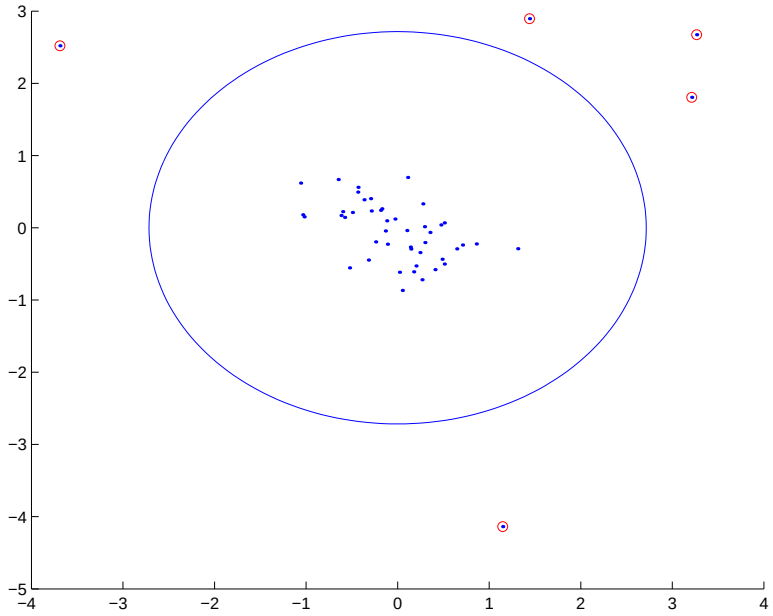


Figure 4.4: 97.5% confidence ellipse for the MD and PS measures after classical data pre-whitening is applied on correlated data with outliers. Note that pre-whitening has artificially brought the outliers closer to the bulk of the data.

to apply a statistical test to the residual vector,

$$\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}. \tag{4.27}$$

However, the value of the residual does not necessarily indicate how outlying the data is, making it hard to use in marking outliers. For instance, a residual of 2 is large when the $\sqrt{MSE}$ (square root of the mean squared error) is 4 but it may be negligible when the $\sqrt{MSE}$ is 10000.

Instead, detection of outliers should be based on statistical distance tests that are robust (high breakdown), computationally feasible, and efficient in high dimensional data. Three statistical tests are discussed next: the Mahalanobis distances (MD), the Projection Statistics (PS), and the Minimum Covariance Determinant (MCD). The PS has been shown to be more robust than the MD [90]; here, we compare the MCD method also.

4.6.1 Mahalanobis Distances

In the Mahalanobis distance method, we diagnose an outlying data point by computing the center of the point cloud and the standardized distances of all points in the design space to that center. The distance measure in its general form is defined as

$$D_i = \sqrt{(\mathbf{h}_i - \bar{\mathbf{h}})^T \hat{\mathbf{R}}^{-1} (\mathbf{h}_i - \bar{\mathbf{h}})}. \tag{4.28}$$

The Mahalanobis distance is then defined as

$$MD_i = \sqrt{(\mathbf{h}_i - \bar{\mathbf{h}}_{MD})^T \hat{\mathbf{R}}_{MD}^{-1} (\mathbf{h}_i - \bar{\mathbf{h}}_{MD})}, \tag{4.29}$$

where the estimate of location, $\bar{\mathbf{h}}_{MD}$, is the sample mean given as

$$\bar{\mathbf{h}}_{MD} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{h}_i \tag{4.30}$$

and the estimate of scatter, $\hat{\mathbf{R}}_{MD}$, is the sample covariance matrix given as

$$\hat{\mathbf{R}}_{MD} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{h}_i - \bar{\mathbf{h}}_{MD})(\mathbf{h}_i - \bar{\mathbf{h}}_{MD})^T. \tag{4.31}$$

For Gaussian distributed $\mathbf{h}_i$, it has been shown [135] that the squared Mahalanobis distances are distributed according to $\chi^2_{d,\alpha}$, where χ^2_d is the chi-squared distribution with d degrees of freedom and α, known as the confidence coefficient, is the probability that the normalized error magnitude is contained within the ellipsoid region defined by the covariance matrix. Using this rationale, points are marked as outliers if $MD_i^2 > \chi^2_{d,0.975}$ for a 97.5% containment probability. But, the MD suffers from masking effects and has a breakdown point of zero. because it uses the non-robust sample mean and sample covariance matrix. For example, in the case of multiple outliers occurring in the same vicinity of the factor space, the sample mean is attracted towards this cluster and the sample covariance matrix is inflated. Furthermore, if one of the $\mathbf{h}_i$ moves to infinity, both the mean vector $\bar{\mathbf{h}}_{MD}$ and the covariance matrix $\hat{\mathbf{R}}_{MD}$ will be unbounded causing a breakdown of the estimator.

4.6.2 Projection Statistics

It is clear that the distance of a point from the cloud should be calculated by means of consistent and high-breakdown multivariate estimators of location and scale. The first affine equivariant multivariate estimator with a high breakdown point was initiated independently by Stahel in 1981 [136] and Donoho in 1982 [27]. It was motivated through the following equivalent expression of the Mahalanobis distances:

$$MD_i = \max_{\|\mathbf{u}\|=1} \frac{\mathbf{h}_i^T\mathbf{u} - \boldsymbol{\mu_p}}{\boldsymbol{\sigma_p}}, \tag{4.32}$$

where $\boldsymbol{\mu_p}$ and $\boldsymbol{\sigma_p}$ are the sample mean and sample standard deviation of the projections of the points $\mathbf{h}_i$ on the direction of the vector $\mathbf{u}$. The equality in this expression holds when all possible directions are considered [90, 145]. To robustify this equation, it was suggested in [42, 43] to use the sample median and median absolute deviation for location and scale estimates. Thus, the PS

estimator is formally expressed as

$$PS_i = \max_{\|\mathbf{u}\|=1} \frac{\left| \mathbf{h}_i^T \mathbf{u} - \underset{j}{\text{med}}(\mathbf{h}_j^T \mathbf{u}) \right|}{1.4826 \underset{i}{\text{med}} \left| \mathbf{h}_i^T \mathbf{u} - \underset{j}{\text{med}}(\mathbf{h}_j^T \mathbf{u}) \right|}. \tag{4.33}$$

While the concept is interesting, it is not practical to pursue the projections in every single direction. Donoho and Gasko therefore advocated the use of a projection algorithm in which only those vectors $\mathbf{u}$ that originate at the coordinate-wise median $\mathbf{m}$ and pass through one of the data points $\mathbf{h}_i$ are investigated [42]. A Projection Statistic is then defined as the maximum of standardized projections of the point cloud on these particular directions for the data point, and represents the worst one-dimensional projection of that data point [43, 129]. Following is the exact procedure of the Projection Statistics algorithm used in this work:

1. Given m data vectors $\mathbf{h}_j$ for $j = 1, \ldots, m$, compute the coordinate-wise median given by

$$\mathbf{m} = \{med(\mathbf{h}_{j1}), med(\mathbf{h}_{j2}), \ldots, med(\mathbf{h}_{jn})\}. \tag{4.34}$$

2. Calculate the j normalized directions:

$$\mathbf{u}_j = \frac{\mathbf{h}_j - \mathbf{m}}{\| \mathbf{h}_j - \mathbf{m} \|} \text{ for } j = 1, \ldots, m \tag{4.35}$$

3. For each direction vector $\mathbf{u}_j$:

   (a) Project the data vectors $\mathbf{h}$ onto each one of these directions $\mathbf{u}_j$, given by

$$z_{1j} = \mathbf{h}_1^T \mathbf{u}_j;\ z_{2j} = \mathbf{h}_2^T \mathbf{u}_j; \ldots;\ z_{mj} = \mathbf{h}_m^T \mathbf{u}_j.$$

   (b) Calculate the median of $\{z_{1j}, \ldots, z_{mj}\} = z_{med,j}$ for each direction j

(c) Calculate the median absolute deviation (as defined in Section 3.1):

$$MAD = 1.4826 \underset{i}{\text{med}} |z_{ij} - z_{med,j}|$$

(d) Calculate the standardized projections:

$$P_{ij} = \frac{|z_{ij} - z_{med,j}|}{MAD} \text{ for } i = 1, \ldots, m$$

4. Using the recursion in Step 3, obtain the standardized projections

$$\{P_{i1}, P_{i2}, \ldots, P_{im}\} \text{ for } i = 1, \ldots, m$$

5. Calculate the Projection Statistics as

$$PS_i = max\{P_{i1}, P_{i2}, \ldots, P_{im}\} \text{ for } i = 1, \ldots, m$$

The PS value computed using this algorithm indicates how far the associated point is from the cloud. Therefore, when applied to the rows of $\mathbf{H}$, a data point $\mathbf{h}_i$ can then be declared an outlier if PS_i exceeds a specified threshold. To determine this threshold, a characterization of the output statistical distribution is desired. But, due to the method's non-parametric nature, a tractable proof of such a distribution is difficult to derive; instead, monte-carlo analyses have been performed by Rousseeuw and Van Zomeren [123, 124]. Similar to the Mahalanobis distances, it has been shown that the PS^2 follow the χ^2_d when the bulk of the data points are normally distributed and redundancy in the design space is larger than 4, i.e. $\zeta = m/n \geq 4$. Therefore, points are marked as outliers using a threshold cutoff value of 97.5% confidence from the χ^2_d distribution.

A large fraction of outliers can be handled due to the robust estimates in this algorithm. Indeed, given that the general position assumption is satisfied, the breakdown point [79] of Projection

Statistics is given as

$$\epsilon^* = \left[\frac{m-n-1}{2}\right] \Big/ m, \tag{4.36}$$

where $[\bullet]$ is the integer part. Computing the distances using this algorithm is also very fast when compared to other alternatives, even in high dimensions. However, this comes at a price, which is a loss of affine equivariance [89]. Yet, this is not a shortcoming in our application because we use the distances as a diagnostic tool to identify the vertical outliers and leverage points before any transformation is applied to the data. In general, the algorithm has been applied in many areas, such as structured and sparse non-linear regression in high dimensional applications for power system state estimation problems [90] and circle-fitting applications [145].

4.6.3 Minimum Covariance Determinant

Other methods to determine the distance of a point from the cloud include the Minimum Covariance Determinant (MCD) and the Minimum Volume Ellipsoid (MVE). We focus on the MCD as it has better statistical properties than the MVE, i.e. MCD is normally distributed asymptotically [20, 24, 106]. For example, the asymptotic efficiency of the reweighted scatter matrix with weights obtained from an MCD estimator is 83% whereas the MVE yields 0% for a problem with only $n = 10$ dimensions [20].

Like the Projection Statistics, the MCD is a robust estimator of location and scatter under a contaminated multivariate distribution. The first step is to find a subset k of the m data points which have the lowest determinant based on the classical covariance matrix. The MCD estimate of location and scatter are then defined as the average and sample covariance matrix of these k points. This estimator is robust to outliers because the subset associated with the computation is from the bulk of the original data set. Specifically, the MCD estimates can reject $(m-k)$ outliers. Clearly, the value of k determines the robustness of the estimator, and the highest rejection of outliers can be achieved by setting $k = [(m+n+1)/2]$. The drawback of MCD is that its computation is involves a combinatorial optimization problem. Various approximations have been proposed to

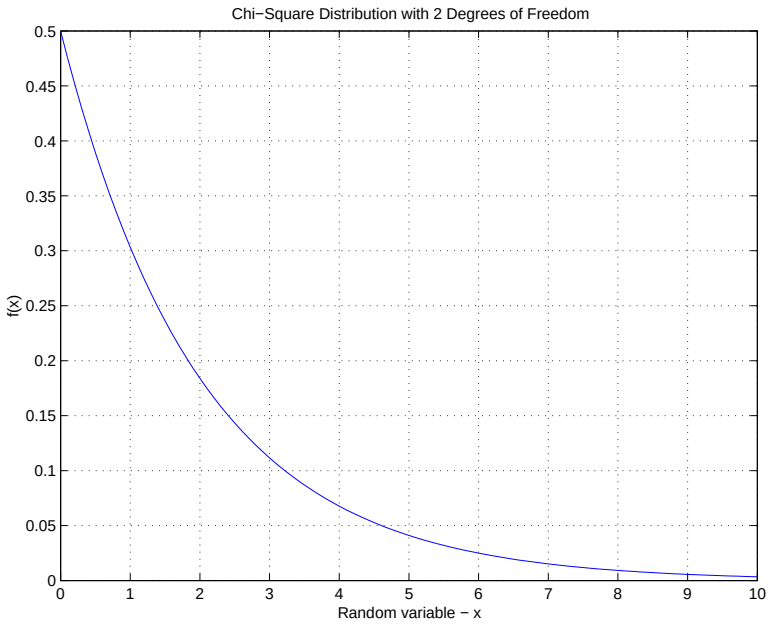
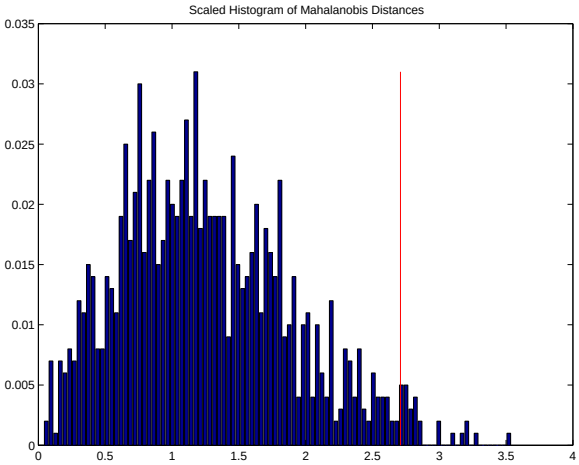


Figure 4.5: Plot of the chi-square PDF with 2 degrees of freedom. Notice the similarity of the histograms of the distances in Figure 4.6 to the PDF.

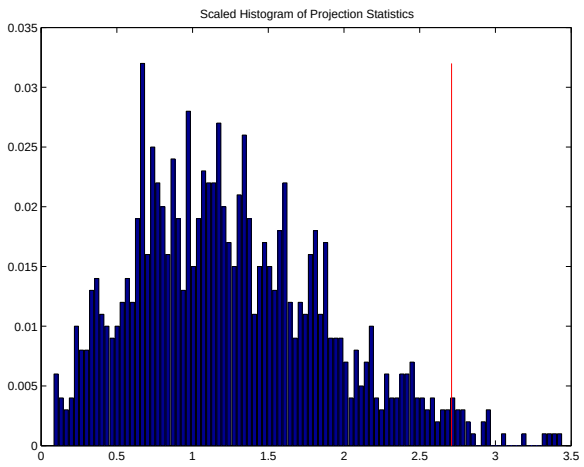
overcome some of this computational complexity, such as the Feasible Subset Algorithm (FSA) of Hawkins [57], Hawkins and Olive [58] and the FAST-MCD proposed by Rousseeuw and Van Driessen [125]. Based on a resampling scheme, a given number of random subsamples are initially drawn and improved iteratively in these procedures. We have used the FAST-MCD algorithm to compute the MCD in the comparisons that follow.

4.6.4 Comparisons between Distance Measures

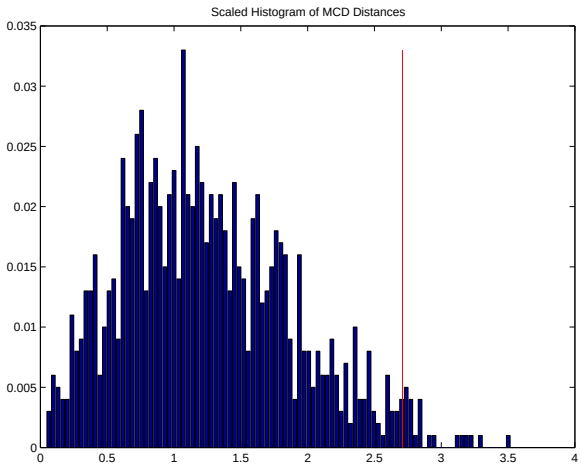
Using simulated data, we study next the use of these measures in diagnosing outliers. Consider again the regression model $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$ with $m = 50$, $n = 2$; i.e. $\mathbf{H}$ consists of 50 two-dimensional Gaussian-distributed data points with zero mean and unit variance. With a redundancy of $\zeta = m/n = 25$, the distances (or normalized errors) are expected to follow the χ^2_2 distribution given in Figure 4.5, as confirmed by a good fit of the relative scaled frequency histograms of MD, PS, and MCD in Figure 4.6. A measure of the estimators' performance can be attained through the covariance matrix. Particularly, the square of the distance measure in (4.28) defines an n-



Mahalanobis Distances Histogram



Projection Statistics Histogram



MCD Distances Histogram

Figure 4.6: Scaled frequency histograms for the distance measures without any outliers follow the χ^2_d distribution well as seen in Figure 4.5.

dimensional error ellipsoid centered at $\bar{\mathbf{h}}$ on which the PDF has a constant value [3, 135]. This ellipsoid is given as

$$D^2 = (\mathbf{h} - \bar{\mathbf{h}})^T \hat{\mathbf{R}}^{-1} (\mathbf{h} - \bar{\mathbf{h}}) = \chi^2_{d,\alpha} = c, \tag{4.37}$$

where c is the threshold corresponding to the confidence coefficient α. As an example, the threshold c corresponding to the 97.5% containment probability is marked in Figure 4.6. The principle axes of this ellipsoid are determined by the eigenvectors of $\hat{\mathbf{R}}$, with magnitudes corresponding to the estimator error in the appropriate directions and equal to $\sqrt{4c\lambda_j}$ and λ_j corresponds to the eigenvalues of the matrix $\hat{\mathbf{R}}$. The volume of this ellipsoid is equal to the determinant of the sample covariance $\hat{\mathbf{R}}$.

Now, we analyze the performance of the distance measures. Figure 4.7 shows $\chi^2_{2,0.975}$ confidence ellipses corresponding to MD, PS, and MCD on the two-dimensional data. Without any outliers, most of the data points are contained within the 97.5% confidence ellipse; for the zero-mean data, the associated threshold c is given by

$$\mathbf{h}^T \hat{\mathbf{R}}^{-1} \mathbf{h} = \chi^2_{2,0.975} = 2.71. \tag{4.38}$$

To test the robustness of these distances, outliers are introduced into the $\mathbf{H}$ matrix. Figure 4.8 shows the relative scaled frequency histograms of the MD, PS, and MCD with 20% outliers in the data. Observing these distributions, the MD fails to identify much of the contaminating points as they are under the threshold, whereas PS and MCD appropriately identify these. Figure 4.9 shows the corresponding effects of the outliers on the confidence ellipse of MD, while the PS and MCD resist the attraction. Note that the data is not necessarily distributed in a bimodal fashion nor are the outliers necessarily grouped together as shown in Figure 4.9. Such a placement of outliers in this example only illustrates the *masking effect*, whereby the MD is undermined and attracted by the outlying group of points. In general, outliers can occur randomly via an unknown underlying distribution in the ϵ-contaminated model presented in Chapter 1.

These results can also be visualized using the Distance-Distance plot that was introduced by

Table 4.1: Results of MD, MCD, and PS under varying outlier contamination amounts.

% Outlier	*MD*	*PS*	*MCD*
1	85.7	25.5	14.5
5	19.3	24.6	13.0
10	9.75	22.1	12.5
15	6.54	20.6	11.9
20	**2.54**	**19.8**	**11.2**

Rousseeuw [125]. In Figure 4.10, the bottom two plots comparing MD versus PS and MCD show points that are correctly marked as outliers using PS and MCD but not by MD. The top plot in the figure depicts the PS versus MCD, both performing equally well in identifying the contamination. Table 4.1 provides a summary of the distance measures' performance under different levels of contamination for grouped outliers; the PS and MCD continue to maintain resistance to the contamination and identify the outliers properly. Once again, the PS measure is applied for outlier diagnosis and down-weighting in the new pre-whitening procedure of the GM-KF; equivalently, the MCD measure could also be used.

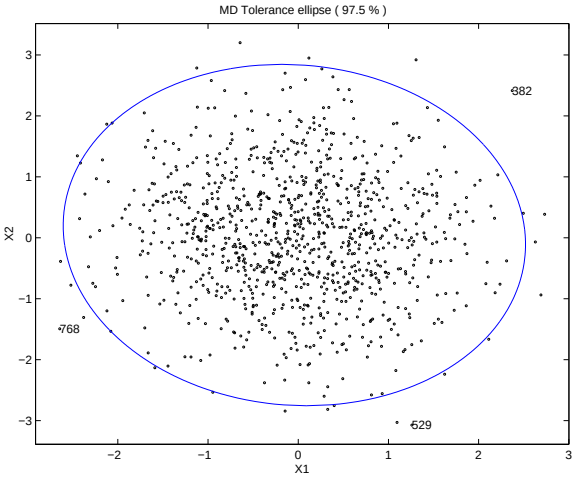
4.7 Robust Pre-Whitening using Projection Statistics

With an understanding of outlier identification methods, we return to the batch-mode regression in (4.11) and its associated covariance matrix in (4.14), where we concluded that a robust pre-whitener is needed to properly decorrelate the data. We propose the following robust pre-whitening procedure in this context:

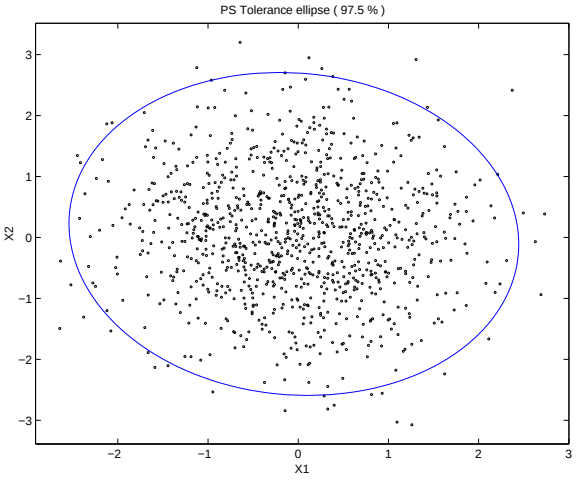
1. Identify the outliers using the Projection Statistics algorithm given in Section 4.6.2. Alternatively, one may use the Minimum Covariance Determinant as discussed in Sections 4.6.3 and 4.6.4.

2. Remove the outliers temporarily from the data vector $\tilde{\mathbf{z}}_k$.

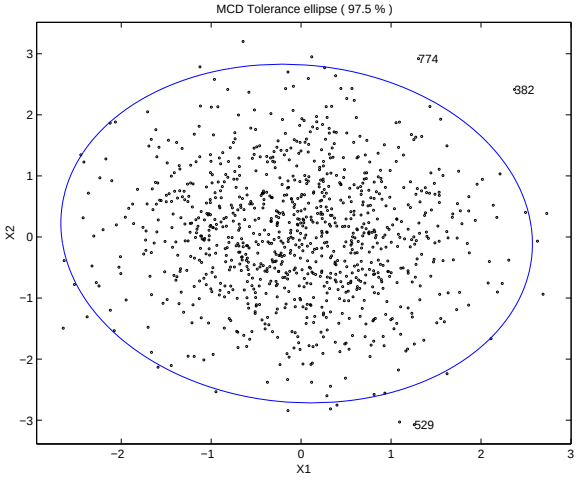
3. Compute the sample covariance matrix $\tilde{\mathbf{R}}_k$ given by (4.14). In the GM-KF framework, this is computed using the known observation and process noise covariances $\mathbf{R}_k$ and $\mathbf{W}_k$,



Mahalanobis Distances

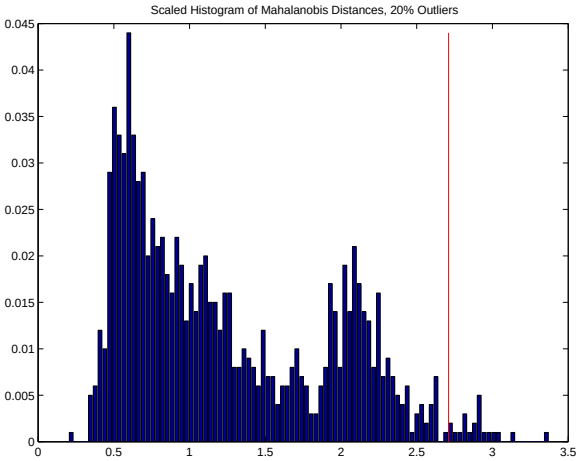


Projection Statistics

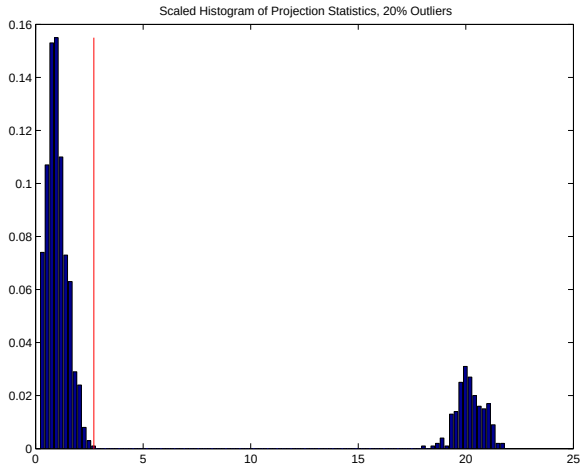


MCD Distances

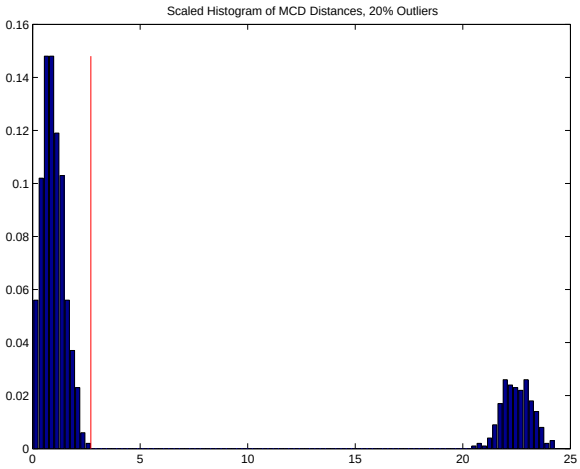
Figure 4.7: Confidence ellipses corresponding to MD, PS, and MCD results, indicating a good containment of the points $\mathbf{h}_i^T$ with no outliers.



Mahalanobis Distances Histogram

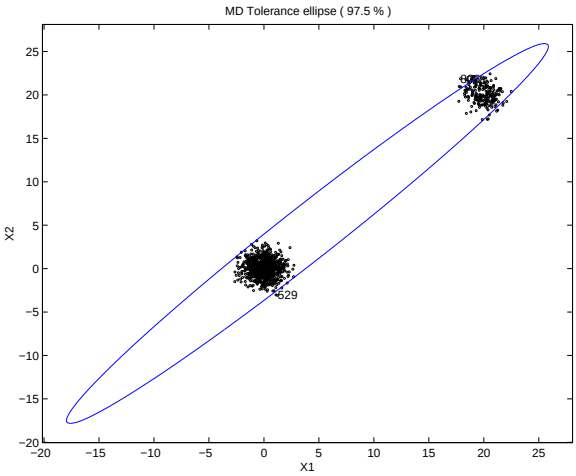


Projection Statistics Histogram

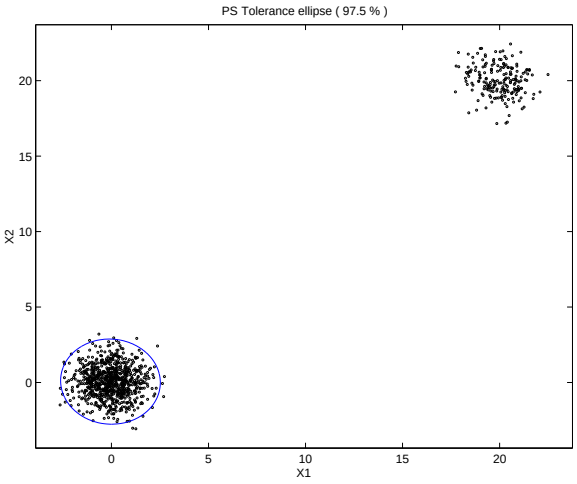


MCD Distances Histogram

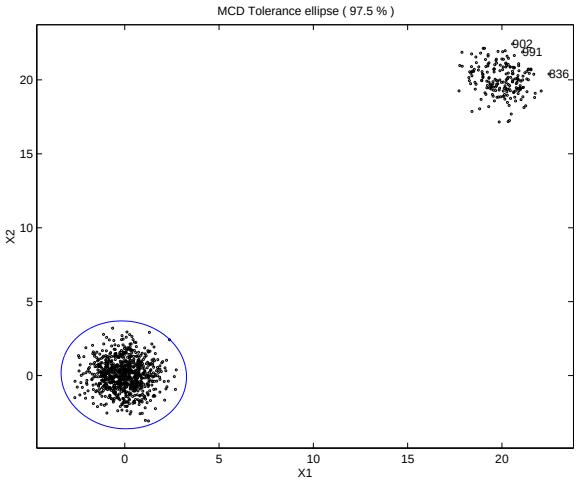
Figure 4.8: Scaled frequency histograms for the distance measures with 20% outliers, indicating failure of MD measure.



Mahalanobis Distances

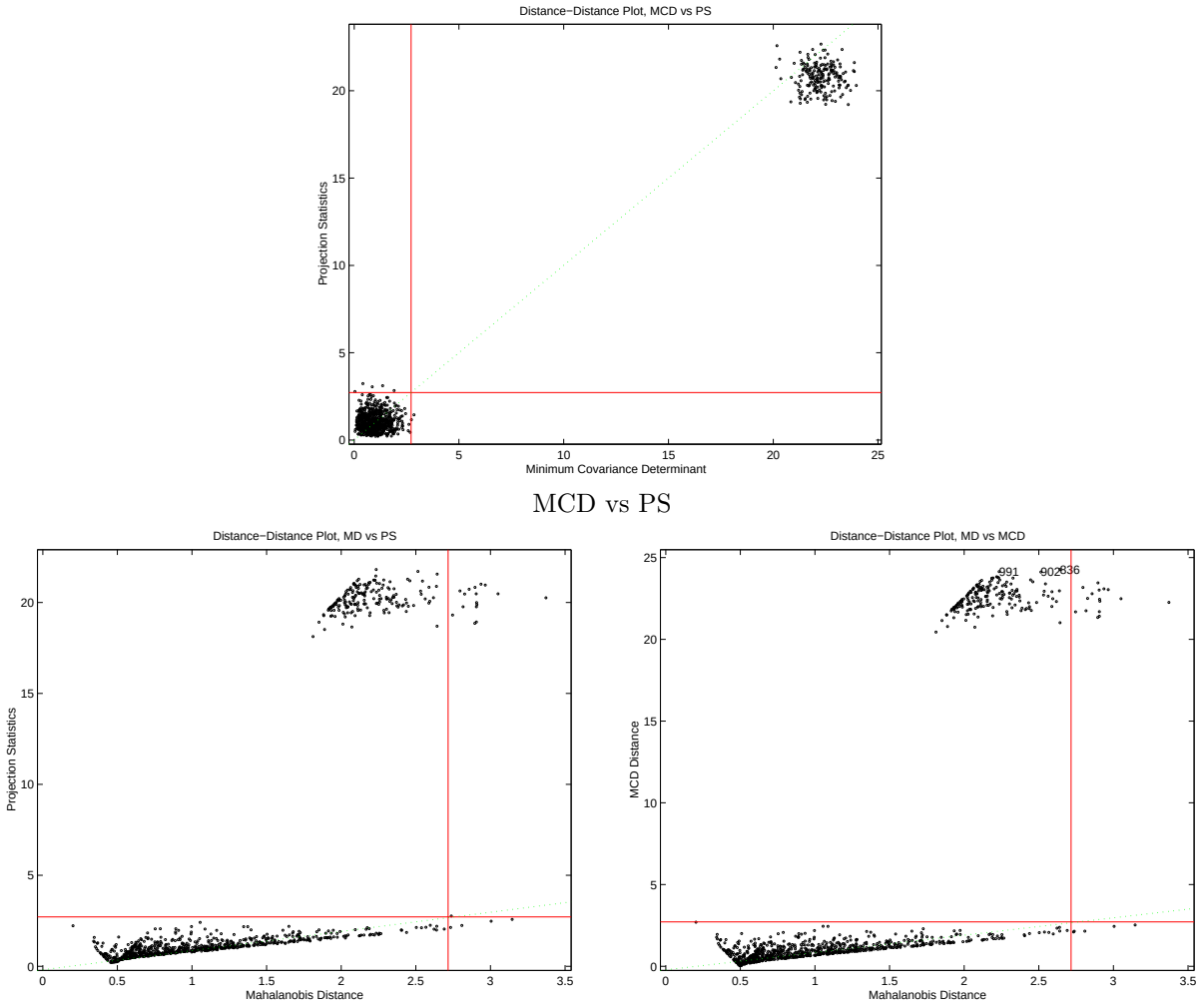


Projection Statistics



MCD Distances

Figure 4.9: Confidence ellipses corresponding to MD, PS, and MCD results, indicating exclusion of outliers by PS and MCD, but not MD.



MCD vs PS

MD vs PS

MD vs MCD

Figure 4.10: Distance-Distance plots comparing MD, MCD, and PS distance measures, indicating good performance of PS and MCD but not MD.

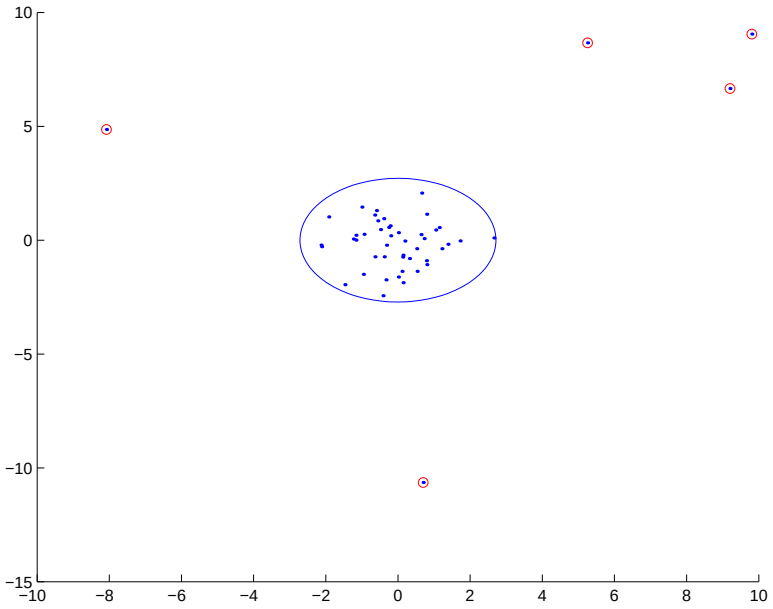


Figure 4.11: 97.5% confidence ellipse for the MD and PS measures after robust data pre-whitening is applied on correlated data with outliers.

respectively.

4. Using either upper diagonal factorization or Cholesky decomposition, obtain the matrix $\mathbf{S}_k$ such that $\tilde{\mathbf{R}}_k = \mathbf{S}_k\mathbf{S}_k^T$. Equivalently, use the square-root method to obtain $\sqrt{\tilde{\mathbf{R}}_k}$ such that $\tilde{\mathbf{R}}_k = \sqrt{\tilde{\mathbf{R}}_k}\sqrt{\tilde{\mathbf{R}}_k}$.

5. Multiply the linear regression model $\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}_k\mathbf{x}_k + \tilde{\mathbf{e}}_k$ on the left-hand side by $(\mathbf{S}_k)^{-1}$ or $(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}$ to perform pre-whitening; for example,

$$(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{z}_k} = (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{H}_k}\mathbf{x}_k + (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{e}_k}. \tag{4.39}$$

6. Re-insert the outliers into the pre-whitened data vector $\tilde{\mathbf{z}}_k$, yielding the final form of the regression as

$$\mathbf{y}_k = \mathbf{A}_k\mathbf{x}_k + \boldsymbol{\eta}_k. \tag{4.40}$$

Figure 4.11, to be compared with Figure 4.4, shows the results of applying this procedure on the data given in (4.25). Clearly, there is a more distinct separation between the point cloud and

the outliers. But, note that some boundary points, i.e. those that were almost included in the confidence ellipse before pre-whitening, are now farther also. The result is a loss of statistical efficiency when solving for the state estimates. Particularly, applying PS on the modified data set will cause an unnecessarily strong down-weighting of the observations in the GM-estimator, and therefore, loss of statistical efficiency. Therefore, to maintain a balance between robustness and efficiency, we compute the weights and down-weight the outliers *directly* in a modified procedure, as follows:

1. Identify the outliers using the Projection Statistics algorithm given in Section 4.6.2. Alternatively, one may use the Minimum Covariance Determinant as discussed in Sections 4.6.3 and 4.6.4.

2. Obtain weights for the elements of the data vector $\tilde{\mathbf{z}}_k$ using the PS values, as follows:

$$\bar{\omega}_i = min\left(1, \frac{d^2}{PS_i^2}\right), \tag{4.41}$$

   where we pick $d = 1.5$ to yield good statistical efficiency at the Gaussian distribution without increasing the bias too much under contamination [55, 62]. Down-weight the vector $\tilde{\mathbf{z}}_k$ using these $\bar{\omega}_i$.

3. Compute the sample covariance matrix $\tilde{\mathbf{R}}_k$ given by (4.14). Again, in the GM-KF, this is computed using the known covariances $\mathbf{R}_k$ and $\mathbf{W}_k$.

4. Using either upper diagonal factorization or Cholesky decomposition, obtain the matrix $\mathbf{S}_k$ such that $\tilde{\mathbf{R}}_k = \mathbf{S}_k\mathbf{S}_k^T$. Equivalently, use the square-root method to obtain $\sqrt{\tilde{\mathbf{R}}_k}$ such that $\tilde{\mathbf{R}}_k = \sqrt{\tilde{\mathbf{R}}_k}\sqrt{\tilde{\mathbf{R}}_k}$.

5. Multiply the linear regression model $\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}_k\mathbf{x}_k + \tilde{\mathbf{e}}_k$ on the left-hand side by $(\mathbf{S}_k)^{-1}$ or $(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}$ to perform pre-whitening; for example,

$$(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{z}_k} = (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{H}}_k\mathbf{x}_k + (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{e}_k}, \tag{4.42}$$

yielding the final form of the regression as

$$\mathbf{y}_k = \mathbf{A}_k \mathbf{x}_k + \boldsymbol{\eta}_k. \tag{4.43}$$

4.8 Solving the Linear Regression Model

4.8.1 Least Squares Solution

Equation 4.43 is the final linear regression form to be used in the GM-Kalman filter scheme. If it is solved with the least squares estimator, the solution [26] is given by

$$\hat{\mathbf{x}}_{k|k} = (\mathbf{A}_k^T \mathbf{A}_k)^{-1} \mathbf{A}_k^T \mathbf{y}_k. \tag{4.44}$$

Using the following Matrix Inversion Lemma,

$$(\mathbf{E} - \mathbf{C}\mathbf{B}^{-1}\mathbf{D})^{-1} = \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{C}(\mathbf{B} - \mathbf{D}\mathbf{E}^{-1}\mathbf{C})^{-1}\mathbf{D}\mathbf{E}^{-1}, \tag{4.45}$$

it can be shown that this batch solution yields the same Kalman filter recursion as described in Section 2.2.

4.8.2 Solution using M-Estimators

The linear regression can also be solved using the robust class of maximum likelihood-type estimators (M-estimators) proposed by Huber in 1972 [61]. For the linear regression

$$\mathbf{y}_k = \mathbf{A}_k \mathbf{x}_k + \boldsymbol{\eta}_k, \tag{4.46}$$

where

$$\mathbf{A}_k = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}, \tag{4.47}$$

an M-estimator is defined as that which minimizes an objective function expressed as

$$J(\mathbf{x}) = \sum_{i=1}^{m} \rho\left(\frac{r_i}{s}\right), \tag{4.48}$$

where $\rho(\cdot)$ is a nonlinear function, r_i the residuals, and s a robust scale estimate. Three subclasses of M-estimators are defined by the choice of the $\rho(\cdot)$ functions. The first subclass is comprised of convex $\rho(\cdot)$ functions, such as the popular Huber function given by

$$\rho\left(\frac{r_i}{s}\right) = \begin{cases} \frac{1}{2}\left|\frac{r_i}{s}\right|^2, & \text{for } \left|\frac{r_i}{s}\right| < b \\ b\left|\frac{r_i}{s}\right| - \frac{b^2}{2}, & \text{elsewhere,} \end{cases} \tag{4.49}$$

where $b = 1.5$ is a cutoff threshold that yields good statistical efficiency at the Gaussian distribution without increasing the bias too much under contamination [55, 62], i.e. high efficiency at Gaussian noise but a bounded and continuous influence function for outliers. This function reduces to L_2-norm for small residuals and to L_1-norm for large residuals. If the noise PDF f is known, recall that $\rho = -\ln f(r_i)$ yields the maximum likelihood solution. The other two classes of M-estimators are comprised of non-convex $\rho(r)$ functions. Figure 4.12 provides examples of the most popular M-estimators in practice from all three classes. Returning to the objective function in (4.48), the residual vector is defined as

$$\mathbf{r} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}} \tag{4.50}$$

where the i^{th} component of $\mathbf{r}$ is given by

$$r_i = y_i - \mathbf{a}_i^T \hat{\mathbf{x}}. \tag{4.51}$$

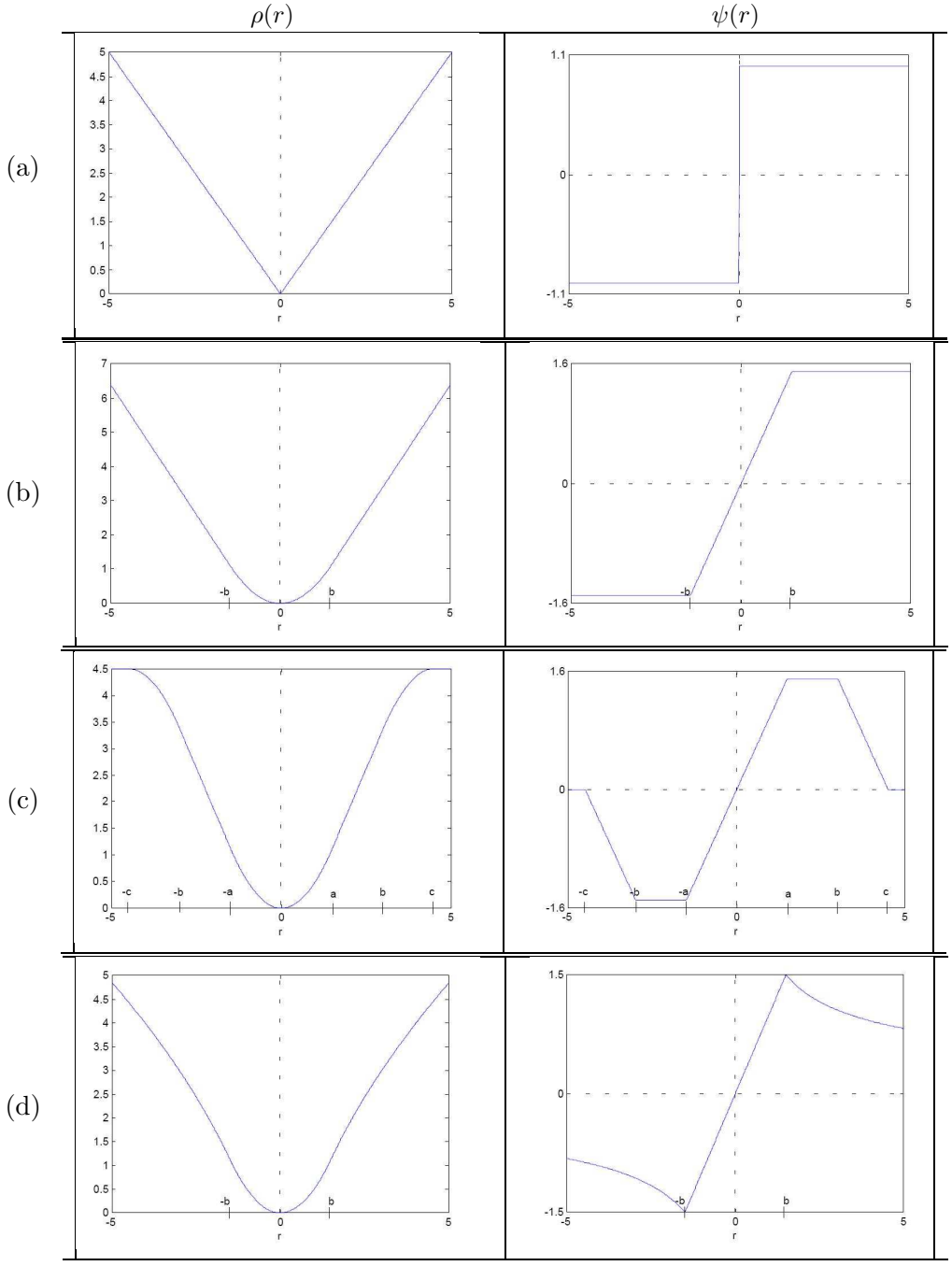


Figure 4.12: ρ-functions for M-estimators yielding different performance properties to deviating data points: (a) L_1 Norm (b) Huber (c) Hampel (d) Merrill Schweppe

.

The Median Absolute Deviation (MAD) is chosen for the robust scale s in (4.48), and is defined as

$$s = MAD = 1.4826\ median_i\ |r_i|. \tag{4.52}$$

The solution is obtained by setting the partial derivatives of the objective function to zero, yielding

$$\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^{m} \frac{1}{s} \frac{\partial \rho\left(\frac{r_i}{s}\right)}{\partial\left(\frac{r_i}{s}\right)} \left(\frac{\partial r_i}{\partial \mathbf{x}}\right)^T = \mathbf{0} \tag{4.53}$$

$$= \sum_{i=1}^{m} -\frac{\mathbf{a}_i}{s} \frac{\partial \rho\left(\frac{r_i}{s}\right)}{\partial\left(\frac{r_i}{s}\right)} = \mathbf{0}, \tag{4.54}$$

where the ρ-function is subject to well-behaved properties such as differentiability. Using the score function $-\psi(u)$, defined as $-\psi(u) = -\partial\rho(u)/\partial u$, we obtain

$$\sum_{i=1}^{m} \frac{\mathbf{a}_i}{s} \psi\left(\frac{r_i}{s}\right) = \mathbf{0}. \tag{4.55}$$

For the Huber function, the score function turns out to be

$$\psi\left(\frac{r_i}{s}\right) = \begin{cases} \left(\frac{r_i}{s}\right), & \text{for } \left|\frac{r_i}{s}\right| < b \\ b \times sign\left(\frac{r_i}{s}\right), & \text{elsewhere.} \end{cases} \tag{4.56}$$

The system of nonlinear equations defined by (4.55) is solved using the Iteratively Re-weighted Least Squares (IRLS) algorithm, as follows:

$$\sum_{i=1}^{m} \left(\frac{\mathbf{a}_i}{s}\right) \left(\frac{r_i}{s}\right) \frac{\psi\left(\frac{r_i}{s}\right)}{\left(\frac{r_i}{s}\right)} = \mathbf{0}, \tag{4.57}$$

$$\sum_{i=1}^{m} \left(\frac{\mathbf{a}_i}{s}\right) \left(\frac{y_i - (\mathbf{a}_i)^T \hat{\mathbf{x}}}{s}\right) q\left(\frac{r_i}{s}\right) = \mathbf{0}, \tag{4.58}$$

where

$$q\left(\frac{r_i}{s}\right) = \frac{\psi\left(\frac{r_i}{s}\right)}{\left(\frac{r_i}{s}\right)}. \tag{4.59}$$

Now, putting (4.58) in matrix form, assuming deterministic $\mathbf{A}$, we obtain

$$\mathbf{A}^T\mathbf{Q}(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0}, \tag{4.60}$$

$$\mathbf{A}^T\mathbf{Q}\mathbf{y} - \mathbf{A}^T\mathbf{Q}\mathbf{A}\hat{\mathbf{x}} = \mathbf{0}, \tag{4.61}$$

where

$$\mathbf{Q} = diag\left\{q\left(\frac{r_i}{s}\right)\right\}. \tag{4.62}$$

The M-Kalman filter solution for the linear regression in (4.43) is given by

$$\hat{\mathbf{x}}_{k|k}^{\nu+1} = \left(\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{y}_k. \tag{4.63}$$

This solution is effective against vertical outliers induced via the observation or system process noise, but not bad leverage points arising from structural outliers in the discrete model. We study this concept next.

4.8.3 Leverage Points and the Need for GM-Estimators

In linear regression, the distribution of the points in the design space is an important factor in the estimator's performance. In this space, robust distance measures computed from explanatory variables allow us to detect a leverage point [125], defined as one whose projection on the design space is an outlier compared to other data points. Figure 4.13 shows examples of good and bad leverage points, where the latter are not consistent with the pattern of the bulk of the data. Mathematically, the influence of these points on the estimator can be factored into the influence of residuals (IR) and the influence of position (IP). The former measures the influence of observation and innovation outliers, while the latter assesses the influence of structural outliers. We study the performance of the least squares, M-, and GM-estimators in a simple regression. Figure 4.14 shows that the weighted least squares does not perform well in the presence of vertical outliers, while the M-estimator is robust to such influence of residuals. However, the M-estimator breaks down due

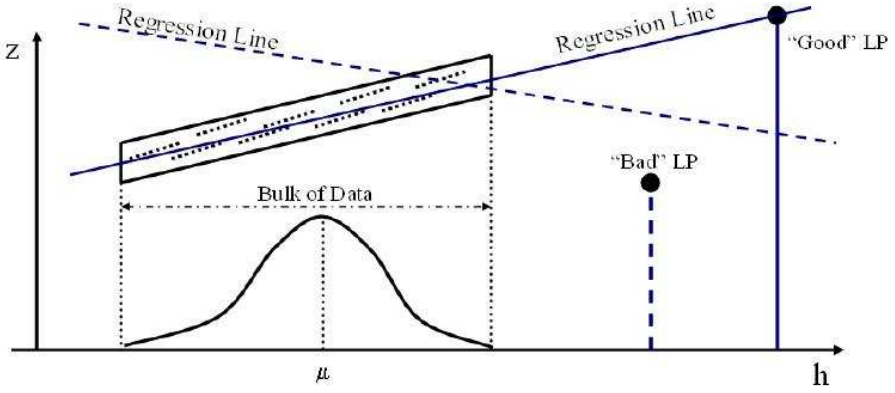


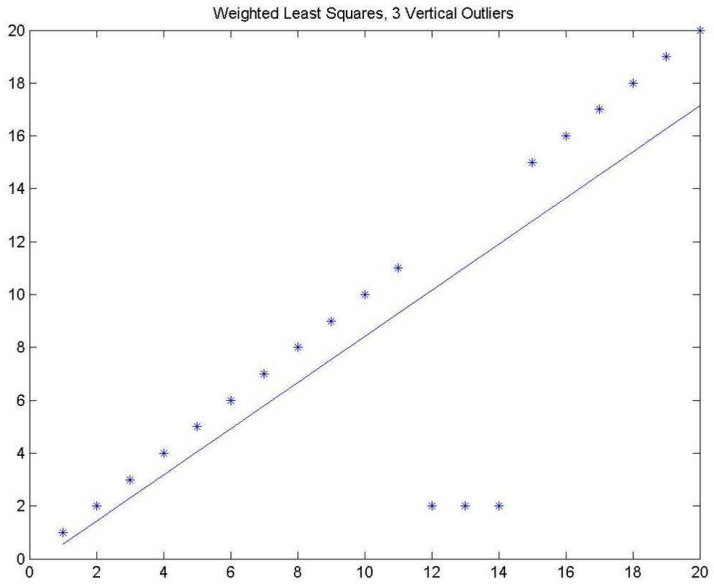
Figure 4.13: Good and bad leverage points can have a strong positive or negative effect on the estimator's performance through the influence of position [129].

to bad leverage points in Figure 4.15, whereas the GM-estimator yields a robust result. This is because the influence of residual is bounded in M-estimators whereas the influence of position is not.

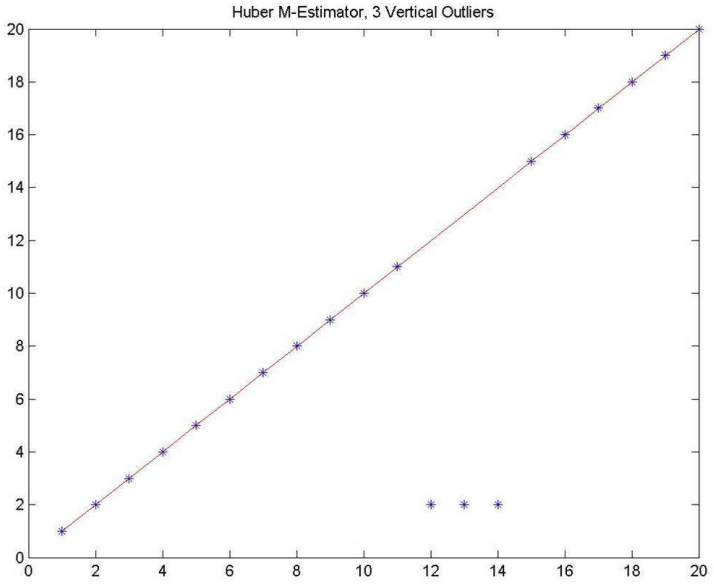
4.8.4 The GM-Estimator Solution

We use the GM-estimators as they have bounded influence to both residuals and position. However, Mallows' proposal did not bound both of them simultaneously. It was Schweppe's proposal, also known as the "Hampel-Krasker-Welsch" estimator [72], to multiply $\mathbf{a}^i$ by a weight function $\bar{\omega}_i$ and divide the residuals by $\bar{\omega}_i$ such that the product of the two, or the *total* influence, is bounded. Schweppe's proposal has been shown to have a positive breakdown point by Maronna [78]. Indeed, assuming known noise variances and using the maximum bias curve from [78], the breakdown point of the GM-estimator can be inferred to be as high as 35%, i.e. $\epsilon^* = 0.35$. Formally, the objective function of the M-estimator is modified in the following manner for the GM-estimator:

$$J(\mathbf{x}) = \sum_{i=1}^{m} \bar{\omega}_i^2 \rho\left(\frac{r_i}{s\bar{\omega}_i}\right). \tag{4.64}$$

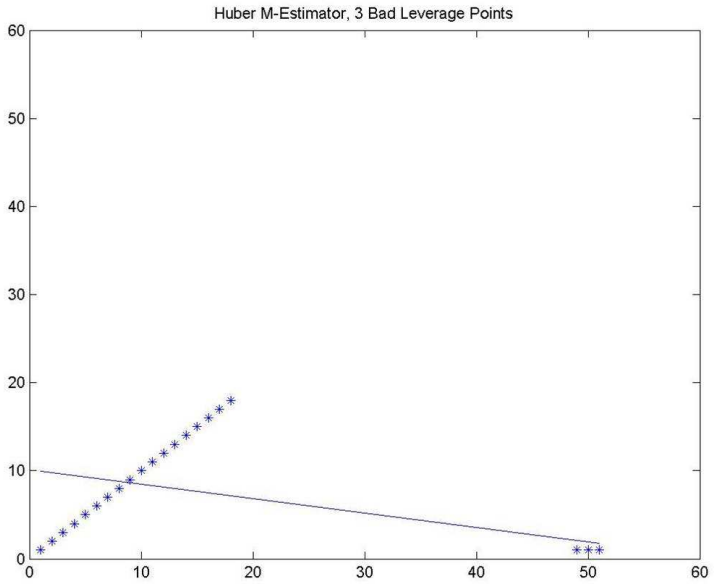


Weighted Least Squares Solution with Vertical Outliers

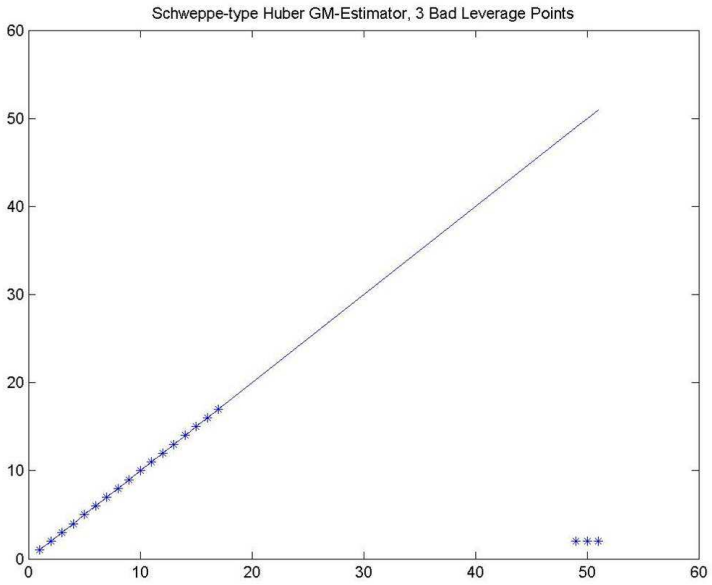


M-Estimator Solution with Vertical Outliers

Figure 4.14: Influence of residuals from vertical outliers causing erroneous results in the weighted least squares solution, but not the M-estimator solution.



M-Estimator Solution with Bad Leverage Points



GM-Estimator Solution with Bad Leverage Points

Figure 4.15: Influence of residuals and position from bad leverage points causing erroneous results in the M-estimator solution, but not the GM-estimator.

Rederiving the solution, we obtain

$$\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^{m} \frac{\bar{\omega}_i}{s} \frac{\partial \rho\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\partial(\frac{r_i}{s\bar{\omega}_i})} \left(\frac{\partial r_i}{\partial \mathbf{x}}\right)^T = 0 \tag{4.65}$$

$$= \sum_{i=1}^{m} -\frac{\bar{\omega}_i \mathbf{a}_i}{s} \frac{\partial \rho\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\partial(\frac{r_i}{s\bar{\omega}_i})} = 0, \tag{4.66}$$

where the ρ function is subject to well-behaved properties such as differentiability. Using the same function $\psi(u)$ defined as $\psi(u) = \partial\rho(u)/\partial u$, we obtain

$$\sum_{i=1}^{m} \frac{\bar{\omega}_i \mathbf{a}_i}{s} \psi\left(\frac{r_i}{s\bar{\omega}_i}\right) = 0. \tag{4.67}$$

This system of nonlinear equations is again solved using the IRLS algorithm. Rewriting (4.67) gives

$$\sum_{i=1}^{m} \bar{\omega}_i \mathbf{a}_i \left(\frac{r_i}{s\bar{\omega}_i}\right) \frac{\psi\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left(\frac{r_i}{s\bar{\omega}_i}\right)} = \mathbf{0}. \tag{4.68}$$

Again, defining the scalar weight function

$$q\left(\frac{r_i}{s\bar{\omega}_i}\right) = \frac{\psi\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left(\frac{r_i}{s\bar{\omega}_i}\right)}, \tag{4.69}$$

Equation 4.68 can be expressed as

$$\sum_{i=1}^{m} \mathbf{a}_i q\left(\frac{r_i}{s\bar{\omega}_i}\right)\left(y_i - \mathbf{a}_i^T \hat{\mathbf{x}}\right) = \mathbf{0}. \tag{4.70}$$

Defining the weight matrix $\mathbf{Q}$ again as

$$\mathbf{Q} = diag\left\{q\left(\frac{r_i}{s\bar{\omega}_i}\right)\right\} \tag{4.71}$$

reduces (4.70) to the following, assuming deterministic $\mathbf{A}$:

$$\mathbf{A}^T\mathbf{Q}(\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0}. \tag{4.72}$$

Solving for the state estimates yields

$$\mathbf{A}^T\mathbf{Q}\mathbf{y} - \mathbf{A}^T\mathbf{Q}\mathbf{A}\hat{\mathbf{x}} = \mathbf{0}. \tag{4.73}$$

The GM-Kalman filter solution is then given by

$$\hat{\mathbf{x}}_{k|k}^{\nu+1} = \left(\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{y}. \tag{4.74}$$

The robust scale s and ρ-function are chosen similarly to the M-estimator, given in (4.49) and (4.52); the key addition in this case are the weights $\bar{\omega}_i$, given by

$$\bar{\omega}_i = min\left(1, \frac{d^2}{PS_i^2}\right), \tag{4.75}$$

where $d = 1.5$. Note that one cannot simply re-compute the Projection Statistics using the vector $\tilde{\mathbf{z}}$ to obtain $\bar{\omega}_i$. Particularly, because the outliers in $\tilde{\mathbf{z}}$ have already been down-weighted as discussed in Section 4.7, corresponding PS values would not yield accurate weights for the remaining structural outliers in the weight matrix $\mathbf{Q}$. To overcome this, we just use the values computed in the pre-whitening procedure described earlier, thereby achieving computational efficiency also. Finally, note that for $\bar{\omega}_i = 1$, a GM-Kalman filter reduces to an M-Kalman filter.

4.9 Summary of the GM-Kalman Filter Scheme

The GM-Kalman filter scheme presented in this chapter is given as a block diagram in Figure 4.16 and summarized as follows:

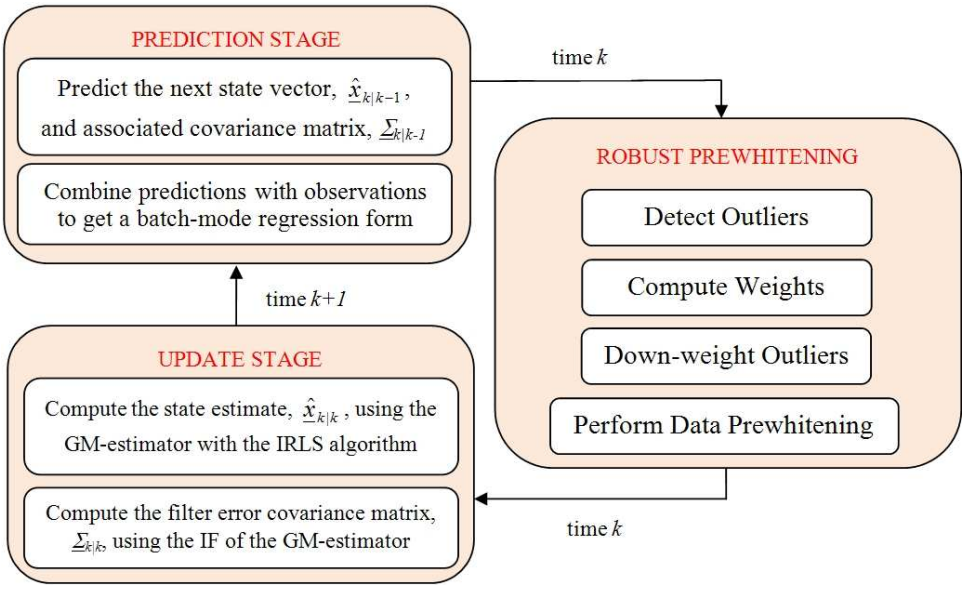


Figure 4.16: Block diagram of the robust GM-Kalman filter and robust data pre-whitening is shown.

1. Begin with the original model:

$$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{w}_k + \mathbf{u}_k \tag{4.76}$$

$$\mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{e}_k \tag{4.77}$$

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{x}_k + \boldsymbol{\delta}_{k|k-1} \tag{4.78}$$

2. Predict the next state using

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{u}_k, \tag{4.79}$$

and as discussed in Section 4.4, combine these predictions with the observations to form a redundant observation vector in the linear regression:

$$\begin{bmatrix} \mathbf{z}_k \\ \hat{\mathbf{x}}_{k|k-1} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_k \\ \mathbf{I} \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} \mathbf{e}_k \\ \boldsymbol{\delta}_{k|k-1} \end{bmatrix} \tag{4.80}$$

$$\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}\mathbf{x}_k + \tilde{\mathbf{e}}_k. \tag{4.81}$$

3. Perform robust data pre-whitening, as discussed in Section 4.7:

   - Identify the outliers using the Projection Statistics algorithm given in Section 4.6.2. Alternatively, one may use the Minimum Covariance Determinant as discussed in Sections 4.6.3 and 4.6.4.
   - Using (4.41), obtain the weights for the elements of $\tilde{\mathbf{z}}_k$ and apply them to the latter.
   - Compute the sample covariance matrix $\tilde{\mathbf{R}}_k$ given by (4.14). For the KF framework, this is computed using the known covariances $\mathbf{R}_k$ and $\mathbf{W}_k$.
   - Using either upper diagonal factorization or Cholesky decomposition, obtain the matrix $\mathbf{S}_k$ such that $\tilde{\mathbf{R}}_k = \mathbf{S}_k\mathbf{S}_k^T$. Equivalently, use the square-root method to obtain $\sqrt{\tilde{\mathbf{R}}_k}$ such that $\tilde{\mathbf{R}}_k = \sqrt{\tilde{\mathbf{R}}_k}\sqrt{\tilde{\mathbf{R}}_k}$.
   - Multiply the linear regression model $\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}_k\mathbf{x}_k + \tilde{\mathbf{e}}_k$ on the left-hand side by $(\mathbf{S}_k)^{-1}$ or $(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}$ to perform pre-whitening; for example,

$$(\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{z}_k} = (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{H}}_k\mathbf{x}_k + (\tilde{\mathbf{R}_k})^{-\frac{1}{2}}\tilde{\mathbf{e}_k}, \tag{4.82}$$

     yielding the final form of the regression as

$$\mathbf{y}_k = \mathbf{A}_k\mathbf{x}_k + \boldsymbol{\eta}_k. \tag{4.83}$$

4. Finally, as discussed in Section 4.8.4, solve for the state estimate iteratively using the following equations:

$$\hat{\mathbf{x}}_{k|k}^{(\nu+1)} = \left(\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{y}_k, \tag{4.84}$$

   where

$$\mathbf{Q} = diag\left\{q\left(\frac{r_i}{s\bar{\omega}_i}\right)\right\} \tag{4.85}$$

   and

$$q\left(\frac{r_i}{s\bar{\omega}_i}\right) = \frac{\psi\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left(\frac{r_i}{s\bar{\omega}_i}\right)} \tag{4.86}$$

and

$$\bar{\omega}_i = min\left(1, \frac{d^2}{PS_i^2}\right). \tag{4.87}$$

All aspects except the error covariance matrix of the robust GM-Kalman filter are now complete. Particularly, the $\mathbf{\Sigma}_{k|k-1}$ remains the same as the classical KF in the prediction stage, but for the correction step, $\mathbf{\Sigma}_{k|k}$ must be derived using the influence function of the GM-estimator.

Chapter 5

Statistical and Numerical Analysis of the GM-Kalman Filter

In this chapter, a statistical and numerical analysis of the robust GM-Kalman filter is carried out. First, the influence functions of the M-estimator and GM-estimator in linear regression are derived. Based on the work of Fernholz [37], we establish a relationship between the influence functions and the asymptotic covariance matrices of these estimators. Using these results, the asymptotic covariance matrix for the correction step of the robust GM-Kalman filter is derived. The convergence rate of the IRLS algorithm used in the filter is also presented.

5.1 Influence Functions of M- and GM-Estimators

The influence functions of the M- and GM-estimators are derived next for the regression given by (4.43). Recall that $\mathbf{y}$ and $\boldsymbol{\eta}$ in this model are i.i.d. vectors following Gaussian distributions with outlier contamination. Assuming the system matrices $\mathbf{F}_k$ and $\mathbf{H}_k$ are deterministic and independent of the residuals, the cumulative probability distribution function of the residual error vector $\mathbf{r}$, expressed as

$$\mathbf{r} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}, \tag{5.1}$$

is denoted by $\Phi(\mathbf{r})$, with each element of the residual vector given as

$$r_i = y_i - \mathbf{a}_i^T \hat{\mathbf{x}}, \tag{5.2}$$

and $\mathbf{a}_i$ is the i^{th} column vector of the matrix $\mathbf{A}^T$.

5.1.1 Influence Function of M-Estimators in Linear Regression

Recall that the M-estimator in regression provides an estimate for $\mathbf{x}$ by processing the redundant observation vector $\mathbf{y}$ and solving the implicit equation given by

$$\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^{m} -\frac{\mathbf{a}_i}{s} \psi\left(\frac{r_i}{s}\right) = \mathbf{0}. \tag{5.3}$$

This equation is written in compact form as

$$\sum_{i=1}^{m} \boldsymbol{\lambda}_i \left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) = \mathbf{0}, \tag{5.4}$$

where, given that the derivative of ρ-function exists,

$$\begin{aligned} \boldsymbol{\lambda}_i \left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) &= \frac{\partial}{\partial \mathbf{x}} \left\{ \rho\left(\frac{r_i}{s}\right) \right\}, & (5.5) \\ &= \frac{\partial \left(\frac{r_i}{s}\right)}{\partial \mathbf{x}} \frac{\partial \rho \left(\frac{r_i}{s}\right)}{\partial \left(\frac{r_i}{s}\right)}, & (5.6) \\ &= \mathbf{a}_i \ \psi\left(\frac{r_i}{s}\right). & (5.7) \end{aligned}$$

Given the empirical cumulative probability distribution function F_m, the functional form of the estimator, where $\mathbf{x}$ is replaced by $\mathbf{T}$, is given by the vector-valued functional

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}\right) dF_m = \mathbf{0}. \tag{5.8}$$

Asymptotically, $F_m \xrightarrow{P=1} G$ by virtue of the Glivenko-Cantelli Theorem [115] and (5.8) becomes

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}(G)\right) dG = \mathbf{0}. \tag{5.9}$$

Following [98, 134, 145], we begin the derivation of the asymptotic influence function, given by

$$\mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi) = \frac{\partial \mathbf{T}(G)}{\partial \epsilon}\Big|_{\epsilon=0} = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}((1-\epsilon)\Phi + \epsilon H) - \mathbf{T}(\Phi)}{\epsilon}, \tag{5.10}$$

by substituting $G = (1-\epsilon)\Phi + \epsilon H$ into (5.9), yielding

$$\begin{aligned}
\int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ dG &= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d[(1-\epsilon)\Phi + \epsilon H] && (5.11)\\
&= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d[\Phi + \epsilon(H - \Phi)] && (5.12)\\
&= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \epsilon \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) = \mathbf{0}. && (5.13)
\end{aligned}$$

Differentiating with respect to ϵ and applying the chain rule yields

$$\frac{\partial}{\partial \epsilon} \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) + \epsilon \frac{\partial}{\partial \epsilon} \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) = \mathbf{0}. \tag{5.14}$$

Next, recall the following interchangeability of differentiation and integration:

$$\frac{d}{d\epsilon} \int_S f(\epsilon, x)\mu(dx) = \int_S f'(\epsilon, x)\mu(dx), \tag{5.15}$$

assuming that f is continuous and measurable and f' is measurable on S [137, 141]. Using this result and given $H = \Delta_r$, where Δ_r is the probability mass at r, (5.14) reduces to

$$\int \frac{\partial}{\partial \epsilon} \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(\Delta_r - \Phi) + \epsilon \int \frac{\partial}{\partial \epsilon} \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(\Delta_r - \Phi) = \mathbf{0}. \tag{5.16}$$

Evaluating this expression at $\epsilon = 0$ makes the last term equal to zero. It follows that

$$\int \frac{\partial}{\partial \epsilon} \left[\boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\right]_{\epsilon=0}\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Delta_r = \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi. \tag{5.17}$$

Assuming Fisher consistency at Φ, the right-hand side of (5.17) becomes zero. Then, using the sifting property of Δ_r, we obtain

$$\int \frac{\partial \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} \frac{\partial \mathbf{T}}{\partial \epsilon}\Big|_{\epsilon=0} d\Phi + \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)) = \mathbf{0}. \tag{5.18}$$

This results in the following expression for the influence function:

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = \frac{\partial \mathbf{T}}{\partial \epsilon}\Big|_{\epsilon=0} = -\left[\int \frac{\partial \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} d\Phi\right]^{-1} \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)). \tag{5.19}$$

Substituting (5.7) into (5.19) yields

$$\begin{aligned} \mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) &= -\left[\int \frac{\partial \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} d\Phi\right]^{-1} \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)), & (5.20)\\ &= \left[-\int \frac{\partial}{\partial \mathbf{x}} \left[\mathbf{a}\ \psi(r_s)\right] d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \psi(r_s)\right], & (5.21) \end{aligned}$$

where $r_s = (r/s)$. Expanding the partial derivative, we can also write (5.21) as

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = \left[-\int \mathbf{a}\ \frac{\partial \psi(r_s)}{\partial r_s}\left(\frac{\partial r_s}{\partial \mathbf{x}}\right) d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \psi(r_s)\right], \tag{5.22}$$

which reduces to

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = \left[\int \frac{\partial \psi(r_s)}{\partial r_s}\ \mathbf{a}\mathbf{a}^T\ d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \psi(r_s)\right]. \tag{5.23}$$

The integral term in (5.23) computes the ensemble average of the random vector $\mathbf{r}$, and equivalently, the term $(\partial\psi(r_s)/\partial r_s)$ for every $(\mathbf{a}_i\mathbf{a}_i^T),\ i = 1,\ldots,m$. Therefore, factoring out the mean of $(\partial\psi(r_s)/\partial r_s)$ yields the following:

$$\int \frac{\partial \psi(r_s)}{\partial r_s}\ \mathbf{a}\mathbf{a}^T\ d\Phi = E_\Phi\left[\psi'\left(\frac{r}{s}\right)\right](\mathbf{A}^T\mathbf{A}). \tag{5.24}$$

Finally, substituting (5.24) into (5.23) gives the influence function as

$$\boldsymbol{IF}(\mathbf{r},\mathbf{a};\Phi) = \frac{\psi\left(\frac{r}{s}\right)}{E_\Phi\left[\psi'\left(\frac{r}{s}\right)\right]}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{a}. \tag{5.25}$$

The following regularity conditions were assumed on the objective function in achieving this final result:

- continuity, boundedness, measurability;
- existence of measurable derivatives.

5.1.2 Influence Function of GM-Estimators in Linear Regression

In a very similar manner, we derive next the influence function of a GM-estimator in linear regression. First, the implicit equation that the GM-estimator solves is given as follows:

$$\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^{m} -\frac{\bar{\omega}_i \mathbf{a}_i}{s} \psi\left(\frac{r_i}{s\bar{\omega}_i}\right) = \mathbf{0}, \tag{5.26}$$

where $\bar{\omega}_i$ are the weights obtained in the prewhitening procedure in Section 4.7. This equation is written in compact form as

$$\sum_{i=1}^{m} \boldsymbol{\lambda}_i \left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) = \mathbf{0}. \tag{5.27}$$

The definition of the function $\boldsymbol{\lambda}$, still assuming that the derivative of the ρ-function exists, then becomes

$$\begin{aligned} \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) &= \bar{\omega}_i^2 \, \frac{\partial}{\partial \mathbf{x}} \left\{ \rho\left(\frac{r_i}{s\bar{\omega}_i}\right) \right\} && (5.28) \\ &= \bar{\omega}_i^2 \, \frac{\partial\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\partial \mathbf{x}} \frac{\partial \rho\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\partial\left(\frac{r_i}{s\bar{\omega}_i}\right)} && (5.29) \\ &= \bar{\omega}_i \; \mathbf{a}_i \; \psi\left(\frac{r_i}{s\bar{\omega}_i}\right). && (5.30) \end{aligned}$$

Again, given the empirical cumulative probability distribution function F_m, the functional form of the estimator $\mathbf{T}_n$ is given by the vector-valued functional

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}\right) dF_m = \mathbf{0}. \tag{5.31}$$

Asymptotically, $F_m \longrightarrow G$ by virtue of the Glivenko-Cantelli Theorem [115] and (5.31) becomes

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}(G)\right) dG = \mathbf{0}. \tag{5.32}$$

Next, given the preceding derivation in (5.11) - (5.18), we use the result in (5.19) and continue as follows:

$$\begin{aligned} \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi) &= \frac{\partial \mathbf{T}}{\partial \epsilon}\Big|_{\epsilon=0} = -\left[\int \frac{\partial \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} d\Phi\right]^{-1} \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(\Phi)), & (5.33)\\ &= \left[-\int \frac{\partial}{\partial \mathbf{x}} \left[\mathbf{a}\ \bar{\omega}\ \psi(r_{\bar{\omega}})\right] d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \bar{\omega}\ \psi(r_{\bar{\omega}})\right], & (5.34) \end{aligned}$$

where $r_{\bar{\omega}} = (r/s\bar{\omega})$. Expanding the partial derivative yields

$$\begin{aligned} \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi) &= \left[\int \mathbf{a}\ \bar{\omega}\ \frac{\partial \psi(r_{\bar{\omega}})}{\partial r_{\bar{\omega}}}\ \frac{\partial r_{\bar{\omega}}}{\partial \mathbf{x}}\ d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \bar{\omega}\ \psi(r_{\bar{\omega}})\right], & (5.35)\\ &= \left[\int \psi'(r_{\bar{\omega}})\ \mathbf{a}\mathbf{a}^T\ d\Phi\right]^{-1}_{\mathbf{T}(\Phi)} \left[\mathbf{a}\ \bar{\omega}\ \psi(r_{\bar{\omega}})\right]. & (5.36) \end{aligned}$$

Again, the integral term in (5.36) computes the ensemble average of the random vector $\mathbf{r}$, and equivalently, the term $(\partial\psi(r_{\bar{\omega}})/\partial r_{\bar{\omega}})$ for every $(\mathbf{a}_i\mathbf{a}_i^T)$, $i = 1, \ldots, m$. Therefore, factoring out the mean of $(\partial\psi(r_{\bar{\omega}})/\partial r_{\bar{\omega}})$ yields the following:

$$\int \psi'(r_{\bar{\omega}})\ \mathbf{a}\mathbf{a}^T\ d\Phi = E_\Phi\left[\psi'\left(\frac{r}{s\bar{\omega}}\right)\right](\mathbf{A}^T\mathbf{A}). \tag{5.37}$$

Finally, substituting (5.37) into (5.36) gives the influence function as

$$\boldsymbol{IF}(\mathbf{r}, \mathbf{a}; \Phi) = \frac{\psi\left(\frac{r}{s\bar{\omega}}\right)}{E_\Phi\left[\psi'\left(\frac{r}{s\bar{\omega}}\right)\right]}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{a}\ \bar{\omega}. \tag{5.38}$$

5.2 Relationship Between the Estimator's Asymptotic Covariance Matrix and the Influence Function

As stated in Chapter 4, the asymptotic filter error covariance matrix needs to be revised to reflect that of the nonlinear GM-estimator in the robust Kalman filter. The asymptotic covariance matrix of the estimation error vector is related to the estimator's influence function as

$$\mathbf{\Sigma} = E[\mathbf{IF}\ \mathbf{IF}^T]. \tag{5.39}$$

Following the approach based on von Mises functionals developed by Fernholz [37], we derive this relationship next. The statistical functional $\mathbf{T}$, following Taylor expansion, becomes

$$\mathbf{T}(G) = \mathbf{T}(\Phi) + \mathbf{T}'(G - \Phi) + rem(G - \Phi) \tag{5.40}$$

and if the influence function exists, then

$$\mathbf{T}'(G - \Phi) = \int \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi)\ d(G - \Phi), \tag{5.41}$$

yielding the von Mises expansion

$$\mathbf{T}(G) - \mathbf{T}(\Phi) = \int \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi)\ d(G - \Phi) + rem(G - \Phi). \tag{5.42}$$

This equation reduces to

$$\mathbf{T}(G) - \mathbf{T}(\Phi) = \int \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi)\ d(G) + rem(G - \Phi), \tag{5.43}$$

since

$$\int \mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi)\ d\Phi = 0, \tag{5.44}$$

due to Fisher consistency at the distribution Φ. For $G = F_m(\mathbf{r})$, where $F_m(\mathbf{r})$ is the empirical distribution function given by

$$F_m(\mathbf{r}) = \frac{1}{m}\sum_{i=1}^{m} u(\mathbf{r} - \mathbf{r}_i), \tag{5.45}$$

and u is the unit step function, the integral term in (5.43) becomes

$$\int \mathbf{IF}(\mathbf{r}, \mathbf{a}; F_m)\ dF_m = \frac{1}{m}\sum_{i=1}^{m} \mathbf{IF}_i(\mathbf{r}, \mathbf{a}; F_m), \tag{5.46}$$

a sum of independent and identically distributed random vectors that tends to $N(\mathbf{0}, \mathbf{\Sigma})$ by virtue of the central limit theorem. Hence, we have

$$\sqrt{m}\,(\mathbf{T}(F_m) - \mathbf{T}(\Phi)) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m} \mathbf{IF}_i(\mathbf{r}, \mathbf{a}; F_m) + \sqrt{m}\ rem(F_m - \Phi). \tag{5.47}$$

If the remainder term converges in probability to zero, that is,

$$\sqrt{m}\ rem(F_m - \Phi) \xrightarrow{P} \mathbf{0}, \tag{5.48}$$

the error term tends in distribution to a multivariate Gaussian, that is,

$$\sqrt{m}\,(\mathbf{T}(F_m) - \mathbf{T}(\Phi)) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}), \tag{5.49}$$

with covariance matrix $\mathbf{\Sigma}$ expressed as

$$\mathbf{\Sigma} = E\left[\mathbf{IF}\ \mathbf{IF}^T\right]. \tag{5.50}$$

5.3 Asymptotic Error Covariance Matrix of the GM-KF

Finally, using the relationship in (5.50) and the influence function in (5.38), the asymptotic error covariance matrix $\mathbf{\Sigma}_{k|k}$ of the robust GM-Kalman filter is given by

$$\mathbf{\Sigma}_{k|k} = E[\mathbf{IF}\ \mathbf{IF}^T] = \frac{E_\Phi\left[\psi^2\left(\frac{r}{s\bar{\omega}}\right)\right]}{\left\{E_\Phi\left[\psi'\left(\frac{r}{s\bar{\omega}}\right)\right]\right\}^2}(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{Q}_{\bar{\omega}}\mathbf{A})(\mathbf{A}^T\mathbf{A})^{-1}, \tag{5.51}$$

where $\mathbf{Q}_{\bar{\omega}} = diag(\bar{\omega}_i^2)$. In the finite-sample case, following Tukey's proposal for M-estimators of location [81], we propose to approximate (5.51) by

$$\begin{aligned}
\mathbf{\Sigma}_{k|k} &= \frac{E_\Phi\left[\psi^2\left(\frac{r}{s\bar{\omega}}\right)\right]}{\left\{E_\Phi\left[\psi'\left(\frac{r}{s\bar{\omega}}\right)\right]\right\}^2}(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{Q}_{\bar{\omega}}\mathbf{A})(\mathbf{A}^T\mathbf{A})^{-1}, && (5.52)\\
&= \frac{\frac{1}{m}\sum_{i=1}^m \psi^2\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left[\frac{1}{m}\sum_{i=1}^m \psi'\left(\frac{r_i}{s\bar{\omega}_i}\right)\right]\left[\frac{1}{m}\sum_{i=1}^m \psi'\left(\frac{r_i}{s\bar{\omega}_i}\right)-1\right]}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}(\mathbf{A}^T\mathbf{Q}_{\bar{\omega}}\mathbf{A})(\mathbf{A}^T\mathbf{A})^{-1}, && (5.53)\\
&= \frac{m\ \sum_{i=1}^m \psi^2\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left[\sum_{i=1}^m \psi'\left(\frac{r_i}{s\bar{\omega}_i}\right)\right]\left[\sum_{i=1}^m \psi'\left(\frac{r_i}{s\bar{\omega}_i}\right)-1\right]}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}(\mathbf{A}^T\mathbf{Q}_{\bar{\omega}}\mathbf{A})(\mathbf{A}^T\mathbf{A})^{-1}. && (5.54)
\end{aligned}$$

If the M-estimator were used instead of the GM-estimator, then the corresponding M-Kalman filter's error covariance matrix [62] using its influence function from (5.25) is given by

$$\mathbf{\Sigma}_{k|k} = E[\mathbf{IF}\ \mathbf{IF}^T] = \frac{E_\Phi\left[\psi^2\left(\frac{r_i}{s}\right)\right]}{\left\{E_\Phi\left[\psi'\left(\frac{r_i}{s}\right)\right]\right\}^2}(\mathbf{A}^T\mathbf{A})^{-1}. \tag{5.55}$$

This result is also readily available from the GM-KF's covariance matrix by recognizing that the weight matrix $\mathbf{Q}_{\bar{\omega}} = \mathbf{I}$ for the M-estimator. Again, we may follow Tukey's proposal to approximate

(5.55) the covariance in the finite-sample case as

$$\begin{aligned}\mathbf{\Sigma}_{k|k} &= \frac{E_\Phi\left[\psi^2\left(\frac{r_i}{s}\right)\right]}{\left\{E_\Phi\left[\psi'\left(\frac{r_i}{s}\right)\right]\right\}^2}(\mathbf{A}^T\mathbf{A})^{-1}, &(5.56)\\ &= \frac{\frac{1}{m}\sum_{i=1}^m \psi^2\left(\frac{r_i}{s}\right)}{\left[\frac{1}{m}\sum_{i=1}^m \psi'\left(\frac{r_i}{s}\right)\right]\left[\frac{1}{m}\sum_{i=1}^m \psi'\left(\frac{r_i}{s}\right)-1\right]}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}, &(5.57)\\ &= \frac{m\sum_{i=1}^m \psi^2\left(\frac{r_i}{s}\right)}{\left[\sum_{i=1}^m \psi'\left(\frac{r_i}{s}\right)\right]\left[\sum_{i=1}^m \psi'\left(\frac{r_i}{s}\right)-1\right]}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}. &(5.58)\end{aligned}$$

Finally, with $\psi(r) = r$, (5.55) further reduces to the error covariance matrix of the classical Kalman filter (i.e. the least squares estimator), given as

$$\mathbf{\Sigma}_{k|k} = (\mathbf{A}^T\mathbf{A})^{-1}. \tag{5.59}$$

5.4 Convergence Rate of the IRLS Algorithm

In this section, the convergence rate of the Iteratively Reweighted Least Squares algorithm is analyzed in its application to an M-estimator of location. The convergence rate is defined as

$$\lim_{\nu\to\infty}\frac{\left|x - x^{(\nu+1)}\right|}{\left|x - x^{(\nu)}\right|^\beta} = \lim_{\nu\to\infty}\frac{\left|e^{(\nu+1)}\right|}{\left|e^{(\nu)}\right|^\beta} = \alpha, \tag{5.60}$$

where ν as the iteration variable, $\beta > 0$, and $\alpha \neq 0$. For a given function $\psi(r)$, where $r = z - x$, recall the IRLS solution of

$$\sum_{i=1}^m \psi(r_i) = 0 \tag{5.61}$$

is given by

$$\begin{aligned}\sum_{i=1}^m r_i\frac{\psi(r_i)}{r_i} &= 0, &(5.62)\\ \sum_{i=1}^m (z_i - x)q(r_i) &= 0, &(5.63)\end{aligned}$$

yielding

$$x^{(\nu+1)} = \frac{\sum_{i=1}^{m} z_i q\left(r_i^{(\nu)}\right)}{\sum_{i=1}^{m} q\left(r_i^{(\nu)}\right)}, \tag{5.64}$$

where $q(r) = \psi(r)/r$ is termed the weight function. A first order Taylor series expansion around $x^{(\nu)}$ yields

$$\psi(z_i - x) = \psi\left(z_i - x^{(\nu)}\right) + \psi'\left(z_i - x^{(\nu)}\right)\left(x - x^{(\nu)}\right). \tag{5.65}$$

For a simple root at $x^{(\nu)}$, such that $\sum_{i=1}^{m} \psi(z_i - x) = 0$, we have

$$\sum_{i=1}^{m} \psi\left(z_i - x^{(\nu)}\right) + \sum_{i=1}^{m} \psi'\left(z_i - x^{(\nu)}\right)\left(x - x^{(\nu)}\right) = 0. \tag{5.66}$$

Re-arranging terms yields

$$\sum_{i=1}^{m} \frac{\psi\left(z_i - x^{(\nu)}\right)}{\left(z_i - x^{(\nu)}\right)}\left(z_i - x^{(\nu)}\right) + \sum_{i=1}^{m} \psi'\left(z_i - x^{(\nu)}\right)\left(x - x^{(\nu)}\right) = 0. \tag{5.67}$$

Using the definition of $q(r)$,we have

$$-\sum_{i=1}^{m} q(r_i) z_i + \sum_{i=1}^{m} q(r_i) x^{(\nu)} = \sum_{i=1}^{m} \psi'\left(z_i - x^{(\nu)}\right)\left(x - x^{(\nu)}\right). \tag{5.68}$$

Dividing both sides of (5.68), by $\sum_{i=1}^{m} q(r_i)$ and using (5.64), we get

$$x - x^{(\nu+1)} + x^{(\nu)} - x = \frac{\sum_{i=1}^{m} \psi'(r_i)(x - x^{(\nu)})}{\sum_{i=1}^{m} q(r)} \tag{5.69}$$

$$e^{(\nu+1)} = e^{(\nu)} + \frac{\sum_{i=1}^{m} \psi'(r_i) e^{(\nu)}}{\sum_{i=1}^{m} q(r_i)} \tag{5.70}$$

$$e^{(\nu+1)} = \left(1 + \frac{\sum_{i=1}^{m} \psi'(r_i)}{\sum_{i=1}^{m} q(r_i)}\right) e^{(\nu)} \tag{5.71}$$

$$|e^{(\nu+1)}| = \left|1 + \frac{\sum_{i=1}^{m} \psi'(z_i - x^{(\nu)})}{\sum_{i=1}^{m} q(r_i)}\right| |e^{(\nu)}| \tag{5.72}$$

Thus, we see that the IRLS algorithm has a linear convergence rate since $\nu = 1$. An alternative to this algorithm is Newton's Method, which has been shown to have a faster quadratic convergence rate for a simple root. However, this method requires convexity in the underlying ρ-function [86], and for multiple roots of order m, it has a linear convergence rate [86].

Chapter 6

Applications of the GM-Kalman Filter

In this chapter, we apply the GM-Kalman filter for improved state estimation and outlier suppression in discrete dynamic models for autonomous vehicles tracking and speech processing applications. In particular, simulations have been carried out to evaluate the performance of the GM-KF from both the efficiency and the robustness view point. Specifically, we first assess the impact that observation redundancy has on the state estimate mean-square error (MSE). Then, we investigate the robustness of our filter when applied to three discrete dynamic models, namely a vehicle and aircraft tracking model, both based on the global positioning system (GPS), and a helicopter's dynamic model. Comparison to the H_∞-filter is performed for the first model. The speech processing problem requires detecting and reconstructing speech segments corrupted by outliers, and as seen in Section 6.7, also involves structural outliers due to the autoregressive model used to represent the speech signal.

6.1 Mean-Square Error of the GM-KF State Estimates

To see the benefits of observation redundancy, we consider a GPS-based vehicle tracking controller that is governed by a dynamic model with the following transition and observation matrices:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.1}$$

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2}$$

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}^T \tag{6.3}$$

This model is provided with 4 or 8 observations when $\mathbf{H}_1$ or $\mathbf{H}_2$ are applied, respectively, and is characterized by a state vector $\mathbf{x} = [x_1\ x_2\ x_3\ x_4]^T$, which contains the horizontal position, x_1, the vertical position, x_2, and the associated velocities, x_3 and x_4. A 10 Hz sampling rate is assumed, giving $T = 0.1s$ as the sampling period.

We have simulated the classical Kalman filter using this model with two scenarios: single observation available at each time step, and multiple observations available at each time step, but *still processed one at a time*. For the latter case, the prediction-update computation is performed multiple times for each observation. Table 6.1 shows the MSE performance when the observation matrix is $\mathbf{H}_1$. The MSE when the filter estimate is derived using a single observation with $\sigma_e^2 = 1$

Table 6.1: The Kalman filter has improved results for different observation noise covariances when multiple observations are available at a given time step, but still processed one at a time.

System process noise variance	**Noise variance of single observation**			**Noise variance of each of the 3 observations per time step**			
	1	10	100	1, 1, 1	1, 10, 100	10, 10, 100	10, 100, 100
10	15.2	20.0	35.7	15.0	15.3	17.9	20.1
100	45.3	48.1	63.2	45.4	45.6	47.6	48.7

is 15.2. Next, let there be 3 observations from noise processes with variances $\sigma_e^2 = 1$, $\sigma_e^2 = 10$, and $\sigma_e^2 = 100$, respectively. When these observations are used to derive the filter estimate, the associated MSE is 15.3 (driven by the lowest observation noise variance). Clearly, multiple observations lead to better filter performance since the one with the lowest noise variance drives the solution.

We now evaluate the benefits of redundancy for the GM-KF. The resulting MSE values are displayed in Tables 6.2 and 6.3 for $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively. Again, we notice that the filter's MSE is mainly determined by the observation with the lowest variance. For example, the MSE equals 19.3 when a single observation with noise variance $\sigma_e^2 = 10$ is processed, whereas the MSE reduces to 5.4 when multiple observations are processed and at least one of them has a variance $\sigma_e^2 < 10$. Note that in both cases the system process noise variance is $\sigma_w^2 = 10$. Thus, we may say that from an efficiency view point, multiple observations are advantageous in that the one with the lowest noise variance drives the MSE value.

Furthermore, the relative efficiencies of the GM-KF and the KF can be viewed with respect to the number of redundant observations processed by the filter. Particularly, the average ratio of the MSE values of the KF and GM-KF in Table 6.3 gives a relative efficiency of 84%. Another example can be seen by comparing the MSE value for both the KF and GM-KF when the observation noise variance is 1; if two observations are processed, the MSE value is lowered for both filters. Similarly, it is observed in Figure 6.1 that the relative efficiency of the GM-KF approaches 95% as the number of redundant observations is increased.

Table 6.2: KF and GM-KF MSE for a single observation per state variable.

System process noise variance	**Noise variance of the single-observation case**	**KF MSE**	**GM-KF MSE**
10	1	4.94	9.46
	10	13.3	19.3
	100	33.7	58.9
100	1	5.19	13.6
	10	15.9	35
	100	61.4	73.5

Table 6.3: KF and GM-KF MSE for two observations per state variable.

System process noise variance	**Noise variance of the single-observation case**	**KF MSE**	**GM-KF MSE**
10	1	3.52	3.96
	10	4.65	5.39
	100	12.6	14.1
100	1	3.54	4.38
	10	4.89	6.1
	100	14.9	17.9

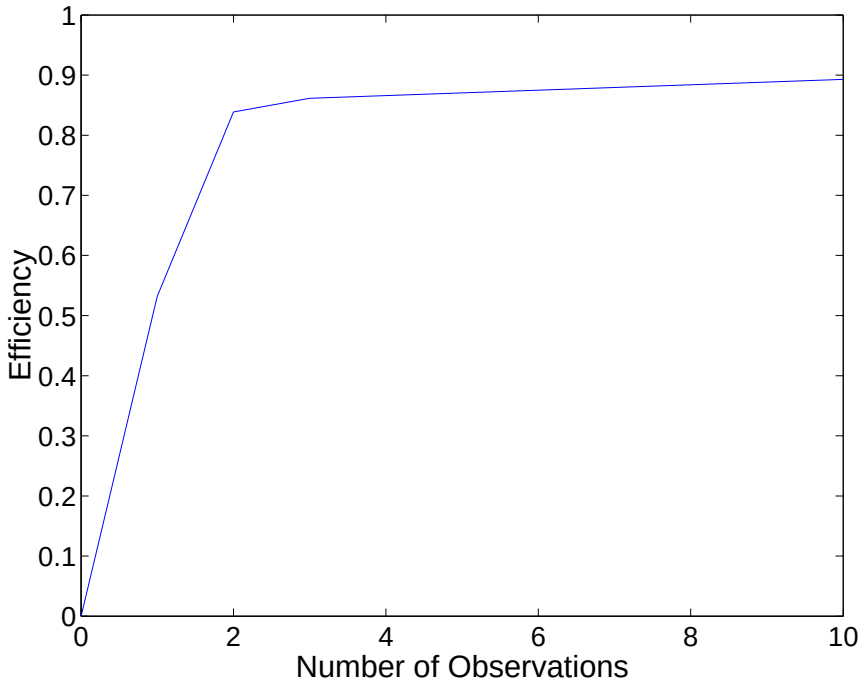


Figure 6.1: Efficiency of the GM-KF approaches 90% with 10 redundant observations, and is expected to reach 95% asymptotically.

6.2 Performance with Uncertainty in Noise Covariance

We now compare and evaluate the performance of the GM-Kalman filter with the classical Kalman filter and H_∞-filter in the presence of noise covariance uncertainty. Recall that the classical KF does not perform well when the covariance matrices of the process and observation noise sources are unknown or different from the assumptions. The GM-Kalman filter inherits this weakness of the KF when no observation redundancy is present. The classical KF's performance in the presence of noise uncertainty has been studied mathematically by Kosanam and Simon [70]; here, we recognize the shortcoming via a simulation. In particular, it is seen in Figure 6.2 that the uncertainty ellipses (i.e. covariance matrix of the estimation error) inflate significantly when the actual noise covariance matrix is different from the assumed one. When there is uncertainty in the noise variance, or there is a fixed bias in the noise that is not considered in the filter design, the H_∞-filter performs by about 23% better than the Kalman filter as depicted in Figure 6.3 and should be the filter of choice. But, the H_∞-filter does not perform well in the presence of outliers such as those induced by modeling mismatch or time-varying bias. Next, we will see the GM-KF performs well against these.

6.3 GPS-based Vehicle Tracking Controller

In the subsequent sections, we study the robustness and sensitivity of the GM-Kalman filter to non-homogeneous observations or predictions. We discuss the performance of the filter in terms of its breakdown point and capability to suppress asymmetrically located outliers placed in the least favorable position for the estimator. The strength of the robust GM-Kalman filter lies in its capability to reject such outliers. First, let us consider a model with the following transition and observation matrices,

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6.4}$$

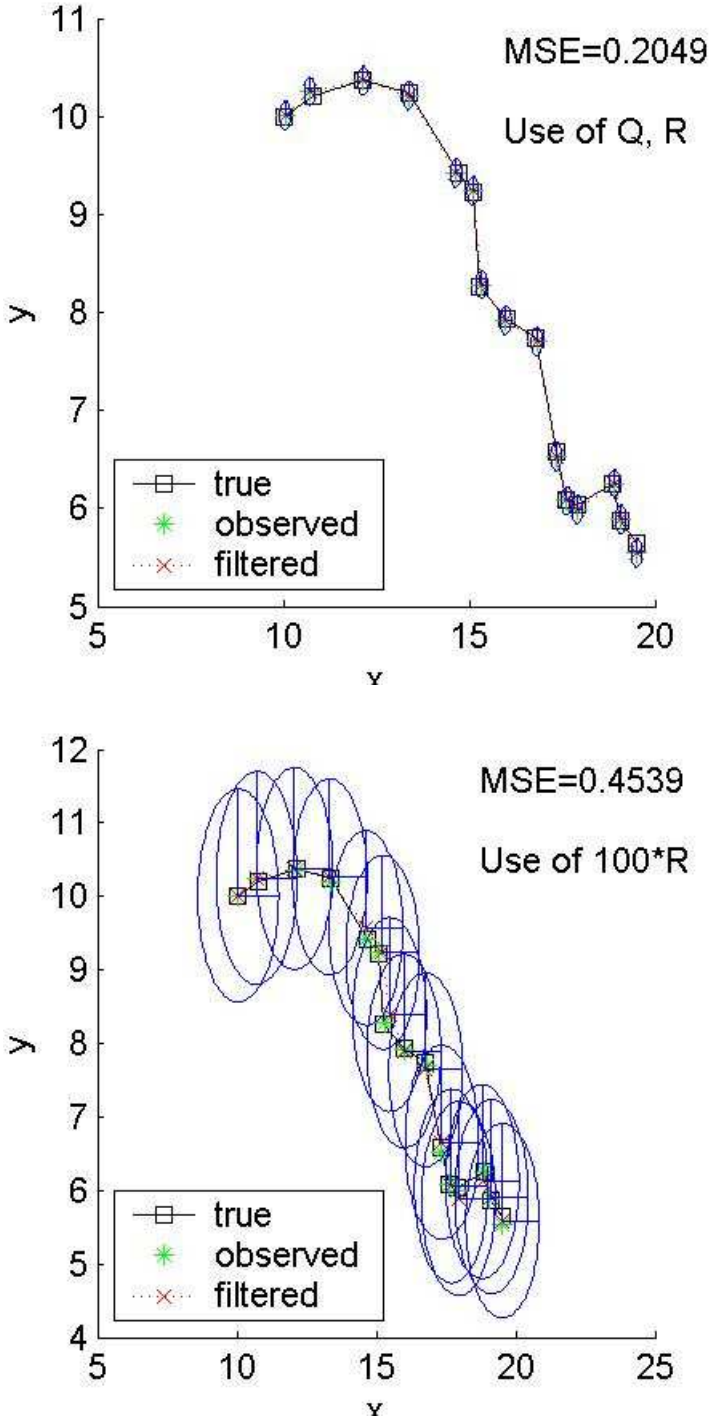


Figure 6.2: The GM-KF, like the KF, shows an increase in the size of the uncertainty ellipses when the actual noise covariances are different from the assumed ones.

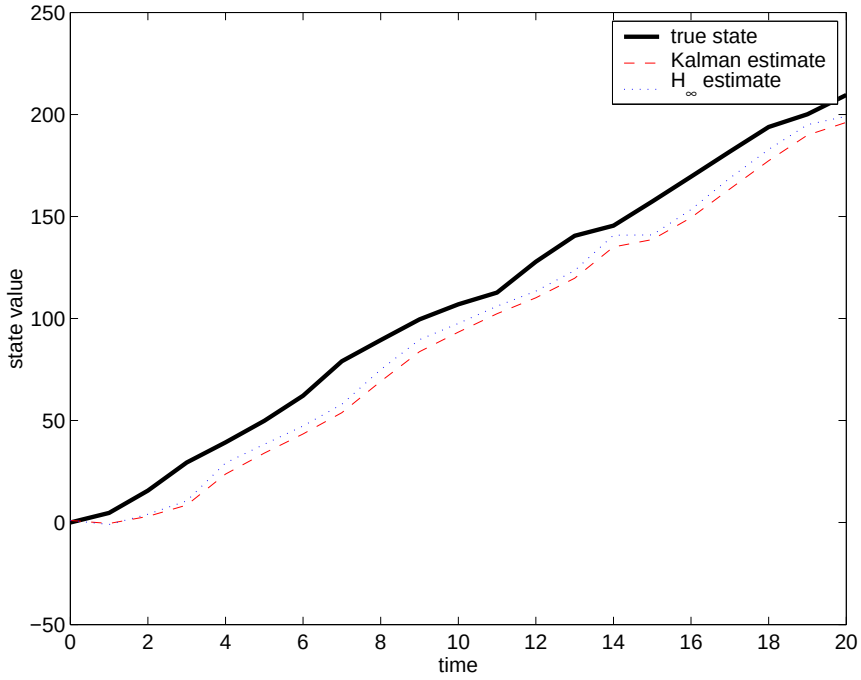


Figure 6.3: The H_∞-filter performs by about 23% better than the classical Kalman filter in the presence of bias in the noise, i.e. when the noise does not follow exactly the assumptions in the model [133].

and

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}^T, \tag{6.5}$$

yielding $m = 8$ and $n = 4$. The model assumptions are as follows: 10 Hz sampling rate, giving $T = 0.1s$ as the sampling period; Gaussian observation noise with a known 4×4 diagonal covariance matrix $\mathbf{R}$ with elements equal to 0.1; and Gaussian system noise with a known 4×4 covariance matrix $\mathbf{W}$ equal to the identity matrix.

To investigate the resistance of our filter and of the H_∞-filter in this model, a single observation outlier with a value of -400 is placed in z_{30}^2, i.e. the second element of $\mathbf{z}$ at $t = 30$ s, when the vehicle is located at position of $x_1 = 325m$ and $x_2 = 100m$. As observed in Figure 6.4, our GM-KF can withstand this outlier whereas the H_∞-filter's estimate is pulled far away from the true position. For the H_∞-filter, the parameter γ is set to 0.025, B_k is a 4×4 diagonal matrix with elements equal to 0.025, and $\mathbf{R}$ and $\mathbf{W}$ are the same as above.

From a statistical viewpoint, while the H_∞-filter performs better than the Kalman filter for noise covariance matrix uncertainty or noise bias, it does not perform well here because the covariance matrix does not capture the tails of the distribution and hence is not effective in capturing the effects of the outliers. On the other hand, the key driver in the performance of the GM-Kalman filter is the use of robust weights in the weight matrix $\mathbf{Q}$ of the GM-estimation solution. To determine the robust weights better, we combined the multiple measurements with the predictions, providing a high level of observation redundancy, $\zeta = m/n >> 1$. By contrast, the recursive procedure of the classical KF and of the H_∞-filter provides nearly no observation redundancy since only one measurement is processed at each step, yielding a redundancy of $\zeta = (n+1)/n \simeq 1$.

Finally, a comparison of the proposed GM-KF with the H_∞-filter should also encompass the computational costs of the algorithms that solve them. Using the same model, we have provided in Table 6.4 the amount of time required to compute the estimates using these two filters. Table

Table 6.4: Computational cost of GM-KF vs. H_∞-filter for different observation redundancies.

Filter Name	**Time to Process 60 steps (ms)**	**Average Time Per Step (ms)**	**# of Redundant Observations**
GM-KF	68.0	1.13	2
	70.5	1.18	4
	74.2	1.24	6
	79.5	1.33	8
	84.4	1.41	10
H_∞	6.40	0.11	0

6.4 also presents the trade-off between the number of redundant observations and computational costs for the GM-KF. Clearly, the GM-KF takes more computational effort than the H_∞-filter. However, the average time per step is on the order of a 1-2 milliseconds, a delay which may be acceptable in many applications. While such increases in computational costs are often a price to be paid for robustness, they are to be minimized wherever possible. It should be noted that these computing time estimates are derived using un-optimized, experimental code implemented in MATLAB; hence, we only expect that this delay would be less when implemented in a more suitable, non-interpretive, programming language.

In all, the proposed robust Kalman filter and the H_∞-filter are very complementary to each other to handle outliers using the former and uncertainties in the covariance matrices of the observation or system modeling errors via the latter. An approach in the literature has been to combine the classical Kalman filter with the H_∞-filter to find the best state estimator in the Kalman filter sense (high statistical efficiency at the Gaussian distribution) but subject to the constraint that the maximum estimation error is bounded. Given this complementary nature, it would be an even better approach to combine the robust Kalman filter with the H_∞-filter in a combined framework, a potential subject of future research.

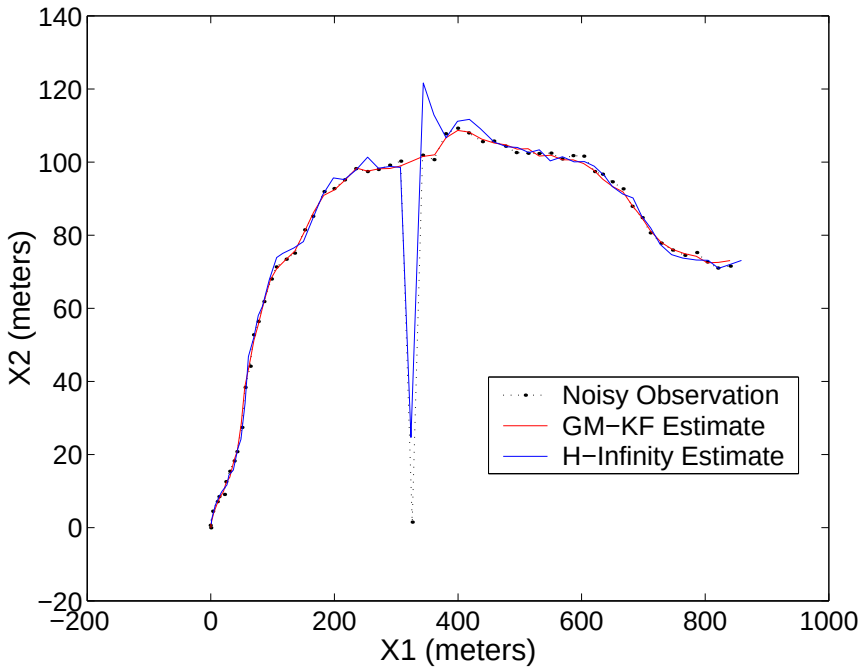


Figure 6.4: Metered position (dotted black line) vs. GM-KF estimate (red line) and an H_∞-filter estimate (blue line) of a vehicles position with one observation outlier. The H_∞-filter performs poorly with just a single outlier.

6.4 GPS-based Aircraft Tracking Model

Next, we consider a tracking problem using GPS data for an aircraft in a circular maneuvering exercise. The state transition matrix, $\mathbf{F}$, is given by

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0.1 \\ 0 & 0 & -0.9 & 1 \end{bmatrix}. \tag{6.6}$$

The observation matrix is once again equal to

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}^T . \tag{6.7}$$

For this model, the maximum number of outliers that any equivariant estimator can handle is $[(m+n)-n)/2] = [m/2] = 8/2 = 4$, yielding a maximum breakdown point of $\epsilon^* = 4/12 = 0.33$. Extensive simulations were carried out in which observation, innovation, and structural outliers were introduced simultaneously at a given time step and successively over several time steps. Through these simulations, it is found that the GM-KF is able to suppress up to 3 concurrent outliers but not 4, implying that it has a breakdown point of $\epsilon^* = 3/12 = 0.25$, which is a little bit less than the maximum.

As an example, consider a scenario in which $\mathbf{W}$ and $\mathbf{\Sigma}_{0|0}$, are set equal to the identity matrix; $\mathbf{R}$ contains 0.75 on the diagonal elements and zero otherwise; and a total of 3 outliers occur from $t = 15s$ to $t = 18s$, one innovation and two observation. Specifically, the first element of the predicted vector $\hat{\mathbf{x}}^1_{k|k-1}$, i.e.

$$[\hat{\mathbf{x}}^1_{15|14}, \hat{\mathbf{x}}^1_{16|15}] = [-16.6, -20.7], \tag{6.8}$$

is replaced by

$$[\hat{\mathbf{x}}^1_{15|14}, \hat{\mathbf{x}}^1_{16|15}] = [-1.61, -3.2]. \tag{6.9}$$

Furthermore, the observation $\mathbf{z}$ is corrupted by outliers such that the first two elements of the original vectors from $t = 15s$ to $t = 18s$, given by

$$\begin{bmatrix} \mathbf{z}^1_{15}, \mathbf{z}^1_{16}, \mathbf{z}^1_{17}, \mathbf{z}^1_{18} \\ \mathbf{z}^2_{15}, \mathbf{z}^2_{16}, \mathbf{z}^2_{17}, \mathbf{z}^2_{18} \end{bmatrix} = \begin{bmatrix} -1.7 & -3.1 & -3.9 & -5.5 \\ -24.6 & -21.3 & -16.0 & -10.8 \end{bmatrix}, \tag{6.10}$$

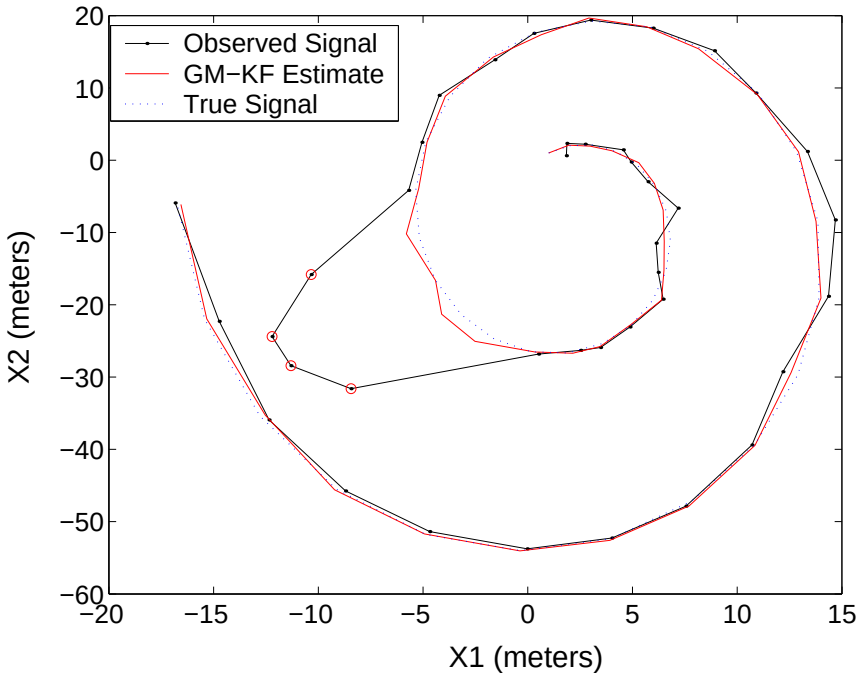


Figure 6.5: True (blue dotted line) and metered (black line) aircraft position vs. GM-KF estimate (red line) in presence of 3 simultaneous and successive outliers (one innovation and two observations).

are replaced by

$$\begin{bmatrix} \mathbf{z}_{15}^1, \mathbf{z}_{16}^1, \mathbf{z}_{17}^1, \mathbf{z}_{18}^1 \\ \mathbf{z}_{15}^2, \mathbf{z}_{16}^2, \mathbf{z}_{17}^2, \mathbf{z}_{18}^2 \end{bmatrix} = \begin{bmatrix} -8.4 & -11.3 & -12.2 & -10.3 \\ -31.6 & -28.4 & -24.4 & -15.8 \end{bmatrix}. \tag{6.11}$$

Figure 6.5 depicts the recorded signal, true state sequence, and the GM-KF estimated values of this example. As observed in the figure, the GM-KF suppresses these three outliers, providing a robust solution near the true state sequence. On average, the filter converges within 4 iterations of the IRLS algorithm in this scenario.

6.5 GPS-based Aircraft Dynamic Model

Now, we consider a model that represents the dynamic behavior of a helicopter under typical loading and flight conditions at airspeed of 135 knots [97, 146, 156]. In the discrete time, the transition

matrix $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{bmatrix} 0.9964 & 0.002579 & -0.0004258 & -0.04597 \\ 0.004513 & 0.9037 & -0.01879 & -0.3834 \\ 0.009762 & 0.03388 & 0.9383 & 0.1302 \\ 0.0004922 & 0.001741 & 0.09677 & 1.007 \end{bmatrix}. \tag{6.12}$$

The observation matrix is the same that in (6.7) and the state vector $\mathbf{x} = [x_1 \; x_2 \; x_3 \; x_4]^T$ contains the horizontal and vertical velocities, pitch rate, and pitch angle, respectively.

Figure 6.6 depicts an example in which two structural outliers occur simultaneously from $t =$ 15 s to $t = 36$ s in the following manner: $F_{3,3} = 10\kappa_1 F_{3,3}$ and $H_{3,3} = 3\kappa_2 H_{3,3}$, where κ_1 and κ_2 are random numbers drawn from the uniform distribution. Clearly, the GM-KF is able to suppress concomitant structural outliers due to redundancy and the use of a robust estimation procedure. More importantly, this example represents a problem of time-varying bias caused by model uncertainties, which the GM-KF solves efficiently and online without a priori knowledge of the contamination distribution.

Figure 6.7 depicts another example in which all three types of outliers occur simultaneously, namely one structural, one observation, and one innovation outlier. The state transition matrix is corrupted as described in the preceding paragraph, but only from $t = 15$ s to $t = 18$ s. In addition, the third element of the observation vector $\mathbf{z}$ is corrupted at each time step by an observation outlier, i.e. $[z_1^3 5, z_1^3 6, z_1^3 7, z_1^3 8] = [18.4, 21.4, 58.6]$ replaces $[z_1^3 5, z_1^3 6, z_1^3 7, z_1^3 8] = [-0.04, 0.84, 3.49]$. Finally, the predicted vector $\hat{\mathbf{x}}_{k|k-1}$ contains an innovation outlier in the third element, which is induced by replacing

$$\begin{bmatrix} \hat{\mathbf{x}}^4_{15|14}, \hat{\mathbf{x}}^4_{16|15}, \hat{\mathbf{x}}^4_{17|16}, \hat{\mathbf{x}}^4_{18|17} \end{bmatrix} = \begin{bmatrix} 0.14 & 2.53 & 0.41 & 3.89 \end{bmatrix}, \tag{6.13}$$

by

$$\begin{bmatrix} \hat{\mathbf{x}}_{15|14}, \hat{\mathbf{x}}_{16|15}, \hat{\mathbf{x}}_{17|16}, \hat{\mathbf{x}}_{18|17} \end{bmatrix} = \begin{bmatrix} 54.4 & 20.8 & 45.5 & 85.2 \end{bmatrix}. \tag{6.14}$$

Figures 6.8, 6.9, and 6.10 depict the state estimates from a sample scenario that contains 3 outliers from $t = 16$ s to $t = 18$ s, namely one structural, one observation, and one innovation outlier. Particularly, element $(3, 3)$ of the observation matrix is affected by a structural error, i.e. $H_{3,3} = 3$ instead of $H_{3,3} = 0$ from $t = 16$ s to $t = 18$ s. The observation vector $\mathbf{z}$ contains an observation outlier, i.e. the first elements of the vector are set equal to $[z^1_{16}, z^1_{17}, z^1_{18}] = [18.4, 21.4, 58.6]$ instead of $[z^1_{16}, z^1_{17}, z^1_{18}] = [-0.04, 0.84, 3.49]$. Finally, the predicted vector $\hat{\mathbf{x}}_{k|k-1}$ is corrupted by outliers such that the third and fourth elements of the uncorrupted prediction vector, given by

$$\begin{bmatrix} \hat{\mathbf{x}}^3_{16|15}, \hat{\mathbf{x}}^3_{17|16}, \hat{\mathbf{x}}^3_{18|17} \\ \hat{\mathbf{x}}^4_{16|15}, \hat{\mathbf{x}}^4_{17|16}, \hat{\mathbf{x}}^4_{18|17} \end{bmatrix} = \begin{bmatrix} 2.53 & 0.41 & 3.89 \\ 1.80 & 2.58 & 4.58 \end{bmatrix}, \tag{6.15}$$

are replaced by

$$\begin{bmatrix} \hat{\mathbf{x}}^3_{16|15}, \hat{\mathbf{x}}^3_{17|16}, \hat{\mathbf{x}}^3_{18|17} \\ \hat{\mathbf{x}}^4_{16|15}, \hat{\mathbf{x}}^4_{17|16}, \hat{\mathbf{x}}^4_{18|17} \end{bmatrix} = \begin{bmatrix} 20.8 & 45.5 & 85.2 \\ 25.1 & 95.1 & 39.3 \end{bmatrix}. \tag{6.16}$$

Extensive Monte-Carlo simulations have confirmed that the GM-KF has a breakdown point of 25%. On average, the filter converged within 5 iterations of the IRLS algorithm. Consequently, the GM-KF may be utilized in the helicopter platform's control system to obtain reliable state estimates for the aircraft longitudinal dynamics when there are up to 25% of gross errors in the model parameter values and/or the measurements. By contrast, as seen in Figure 6.11, the H_∞-filter cannot withstand even a single observation outlier.

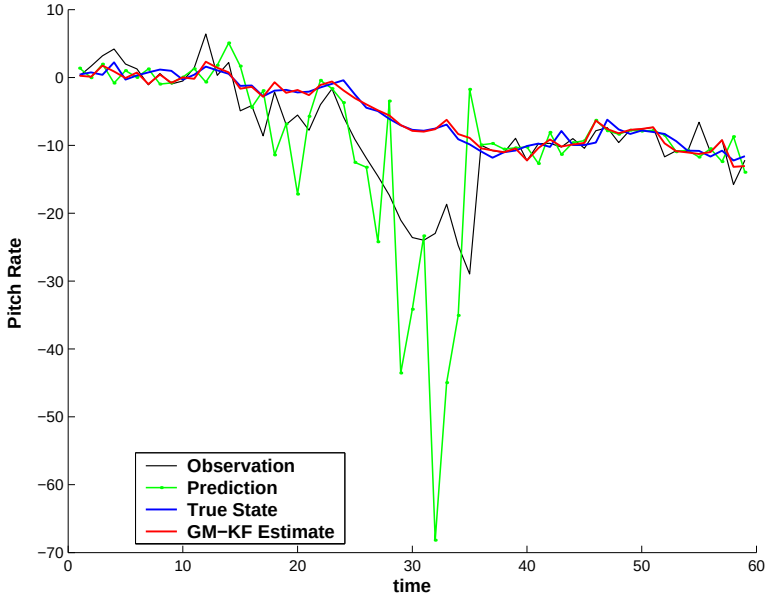


Figure 6.6: True (blue line), predicted (green line), and metered (black line) horizontal velocity vs. GM-KF estimated horizontal velocity (red line) of a helicopter in the presence of one observation, one innovation, and one structural outlier occurring simultaneously at from $t = 15$ s to $t = 36$ s.

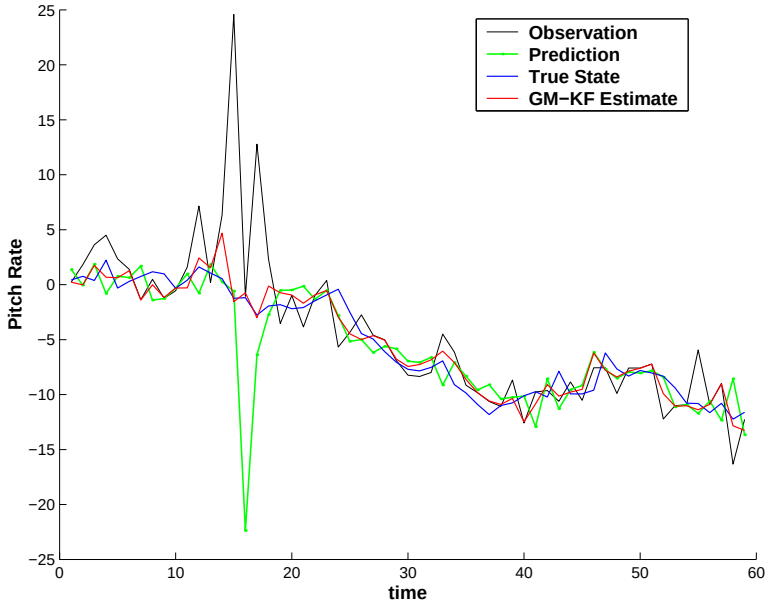


Figure 6.7: True (blue line), prediction (green line), and metered (black line) pitch rate vs. GM-KF estimated pitch rate (red line) of a helicopter in the presence of one observation, one innovation, and one structural outlier occurring simultaneously at from $t = 15$ s to $t = 18$ s.

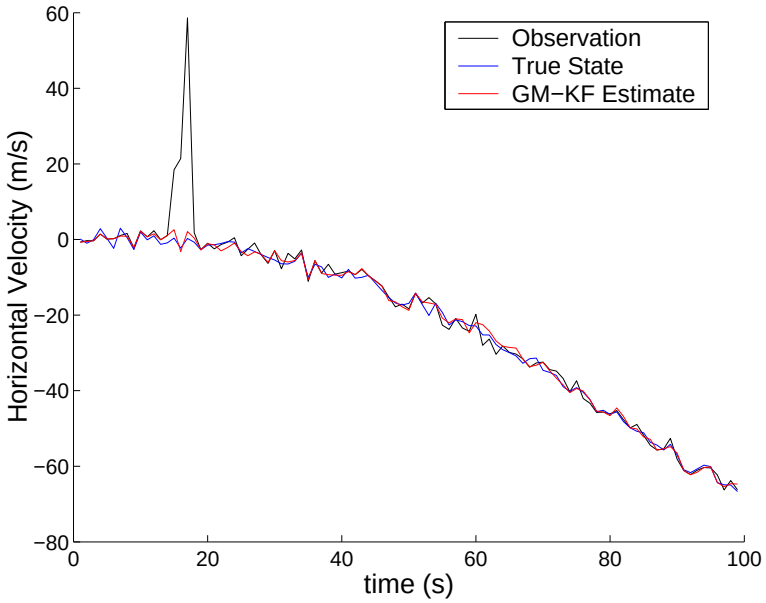


Figure 6.8: True (blue line) and metered (black line) horizontal velocity vs. GM-KF estimated horizontal velocity (red line) of a helicopter in the presence of one observation, one innovation, and one structural outlier occurring simultaneously at $t = 15s$, $t = 16s$, $t = 17s$, and $t = 18s$.

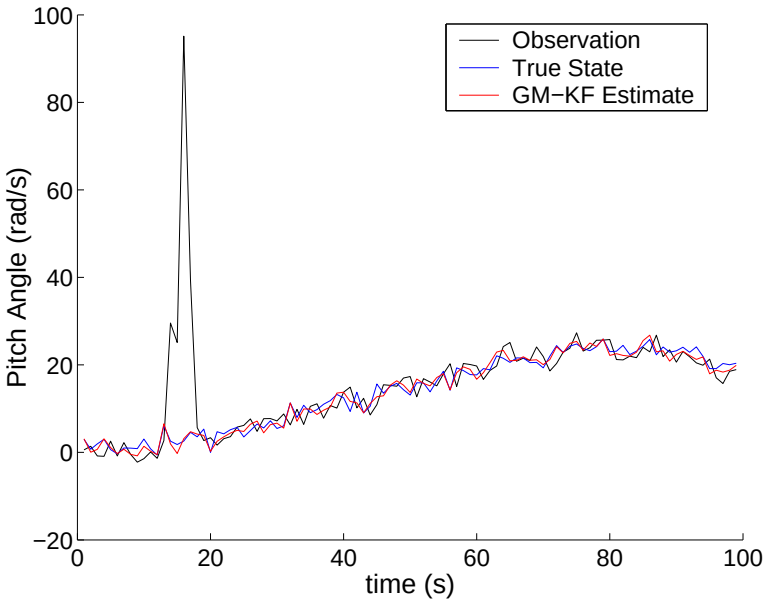


Figure 6.9: True (blue line) and metered (black line) pitch rate vs. GM-KF estimated pitch rate (red line) of a helicopter in the presence of one observation, one innovation, and one structural outlier occurring simultaneously at $t = 15s$, $t = 16s$, $t = 17s$, and $t = 18s$.

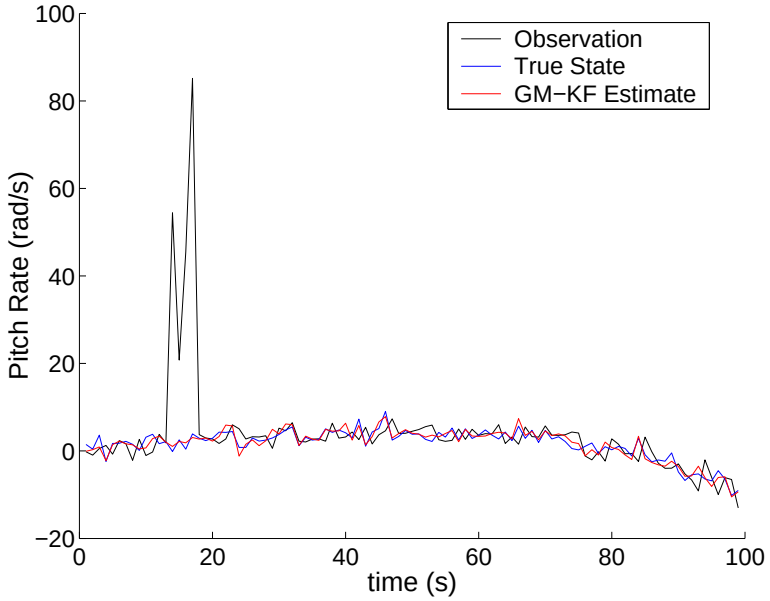


Figure 6.10: True (blue line) and metered (black line) pitch angle vs. GM-KF estimated pitch angle (red line) of a helicopter in the presence of one observation, one innovation, and one structural outlier occurring simultaneously at $t = 15s$, $t = 16s$, $t = 17s$, and $t = 18s$.

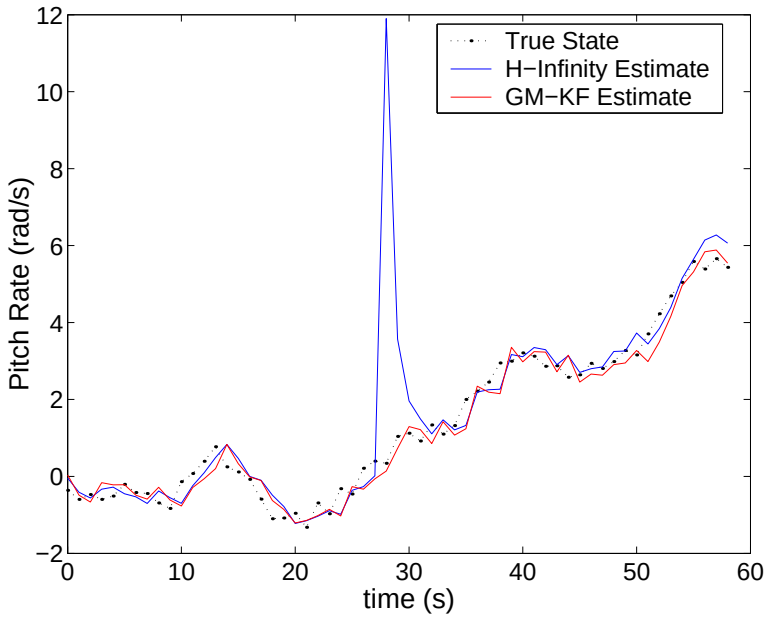


Figure 6.11: True pitch rate (dotted line) vs. GM-KF (red line) and H_∞-filter (blue line) estimated pitch rate of a helicopter in presence of a single observation outlier.

6.6 Arbitrary Discrete Dynamic Model

Next, we move to a more general dynamic model that represents just an arbitrary control system. The state and transition matrices of this model are given by

$$\mathbf{F} = \begin{bmatrix} 0.8 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.0 \\ 0.5 & 0.1 & 0.2 \end{bmatrix}, \tag{6.17}$$

and

$$\mathbf{H} = \begin{bmatrix} 0.4 & 2.0 & 3.0 & 2.0 & 0.3 & 0.5 \\ 0.1 & 0.1 & 0.3 & 0.1 & 4.0 & 3.0 \\ 1.0 & 1.0 & 0.5 & 0.0 & 0.1 & 2.0 \end{bmatrix}^T. \tag{6.18}$$

This example shows that the GM-Kalman filter derives one of its main advantages through its capability to perform a majority-minority comparison over the measurements and predictions together. If we revisit the equation for the maximum breakdown point from Chapter 3, given by

$$\epsilon^*_{max} = [(m-n)/2]/m, \tag{6.19}$$

we see that, for instance, with 6 measurements and 3 predictions in a 3-state system, the maximum number of outliers that can be handled is $[(m-n)/2] = [(9-3)/2] = 3$. In this case, the estimator is able to reject up to three outliers in theory; with more measurements, this breakdown point can be increased further. After carrying out extensive simulations, it was determined that the GM-Kalman filter's breakdown point in this setup can be no larger than 33%. In particular, with 6 measurements and 3 predictions to estimate 3 state variables, the system suppressed up to 3 simultaneous outliers. Figure 6.12 demonstrates the filter's power in the case of 3 such outliers occurring simultaneously and placed strategically to cause masking effects.

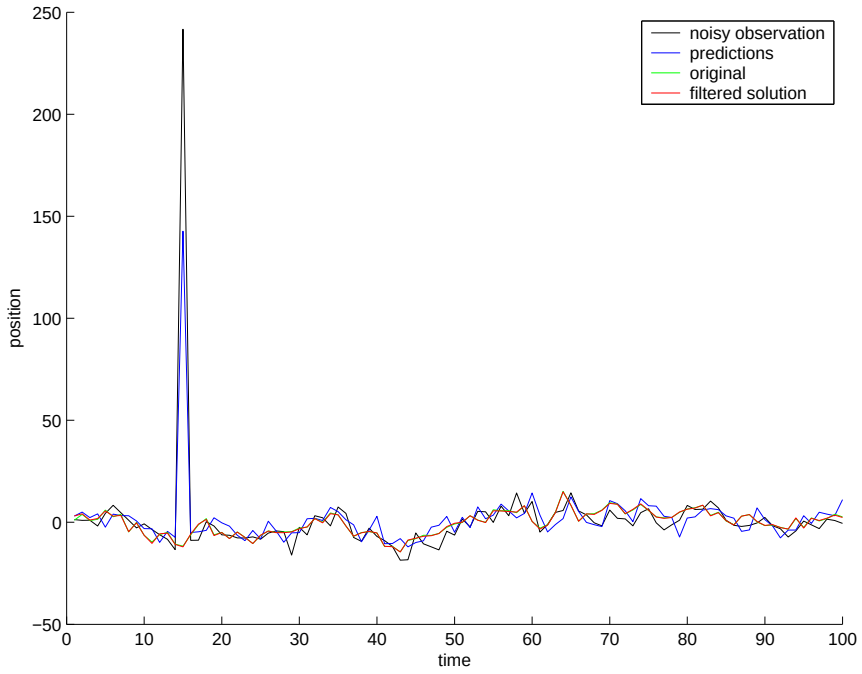


Figure 6.12: Estimation results from robust KF for one of the state variables and associated estimation errors from a 3 dimensional general dynamic model containing 1 additive outlier and 2 innovation outliers at $t = 16$.

6.7 Speech Enhancement via GM-KF

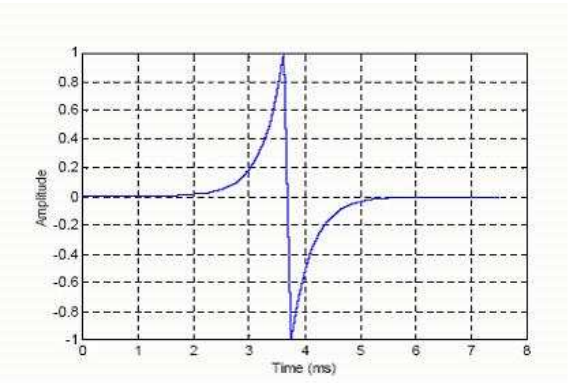
We now look at how the GM-KF can be applied to removing outliers in speech caused by impulsive noise. This type of noise may be modeled as exponential functions mirroring each other about a discontinuity point, as shown in Figure 6.13 in time and spectral domains. It may introduced in speech by co-channel and fading interferences and demodulation process in communications, and its negative effect is clear by observing the time domain original and corrupted speech waveforms in Figure 6.14. The clean and corrupted speech spectrograms, depicted in Figure 6.15, also show strong distortions in the form of vertical striations. The speech is sampled at 8 kHz with impulses lasting up to 240 samples.

Compensating for impulsive noise is a very difficult task as it completely destroys the segment of speech and classical statistical techniques based on least-squares estimation perform poorly. We will use the Projection Statistics (PS) method to first detect the corrupted speech segments. We then use a modified form of the popular Linear Predictive Coding (LPC) method to model and estimate the missing speech. This LPC scheme casts an autoregressive model (AR) of the speech into a linear regression framework, on which the Schweppe-type Huber GM-estimator is applied to robustly estimate the AR parameters. Effectively, we are using the GM-KF to estimate the speech parameters in the linear autoregressive model. We perform the processing over window lengths of 30 ms to obtain quasi-stationarity over segments of speech. A block diagram of the proposed scheme is presented in Figure 6.16.

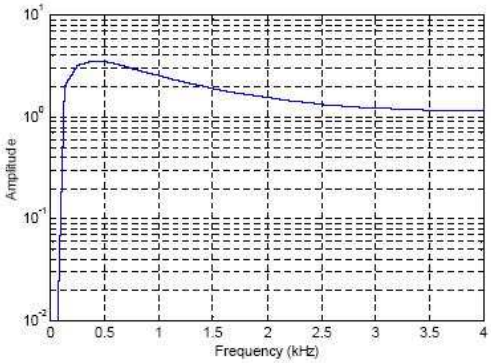
6.7.1 Outlier Detection in the Autoregressive Model

We model the speech signal via the LPC technique, which uses the autoregressive model given by

$$s_n = \sum_{k=1}^{n} a_k s_{n-k} + e_k. \tag{6.20}$$

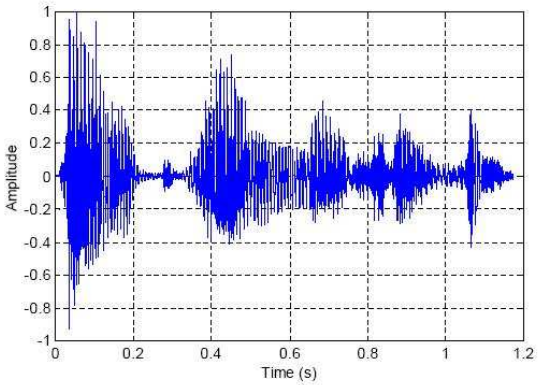


Time Domain

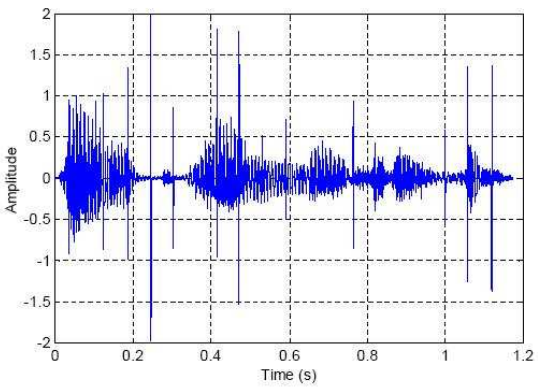


Spectral Domain

Figure 6.13: Impulse noise modeled as exponential functions mirroring each other about a discontinuity point is shown in time and spectral domain.

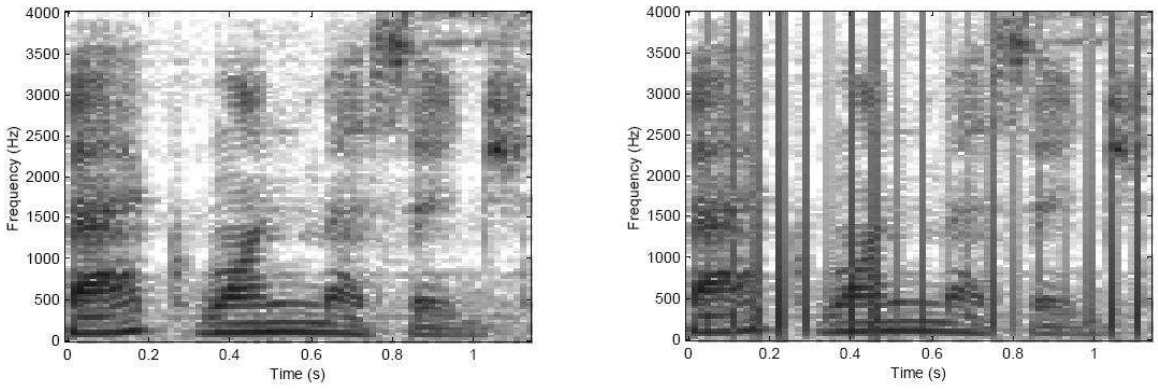


Clean Speech Signal



Corrupted Speech Signal

Figure 6.14: Speech waveform is shown before and after corruption by impulsive noise. Notice the visually (and audibly) disturbing spikes in the corrupted signal.



Clean Speech Signal

Corrupted Speech Signal

Figure 6.15: Speech waveform is shown before and after corruption by impulsive noise in the spectral domain. Notice the visually (and audibly) disturbing vertical striations in the corrupted signal.

This AR model can be cast into a linear regression form, $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$, as follows:

$$\begin{bmatrix} s_{n+5} \\ s_{n+6} \\ \vdots \\ s_{n+m+5} \end{bmatrix} = \begin{bmatrix} s_{n+4} & s_{n+3} & s_{n+2} & s_{n+1} \\ s_{n+5} & s_{n+4} & s_{n+3} & s_{n+2} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n+m+4} & s_{n+m+3} & s_{n+m+2} & s_{n+m+1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + \begin{bmatrix} e_{n+5} \\ e_{n+6} \\ \vdots \\ e_{n+m+5} \end{bmatrix}. \tag{6.21}$$

The time series signal used in the experiments is an actual professionally recorded speech from a cellular phone with real impulsive noise introduced and real background noise present in the recording. Hence, the time series signal does not contain synthetic signal or noise generated from a model. Each row vector of $\mathbf{H}$ defines the time series signal as an n-dimensional point and the impulses will appear as outlier points distant from the bulk of the data set in this n-dimensional space via the measure of PS. Recall that Projection Statistics is defined as the maximum of the standardized projections of the data points on all directions passing through the origin, since in this case, speech is centered at zero. A projection statistic indicates how far the associated point is from the point cloud in the worst one-dimensional projection. The impulsive noise is then detected as outliers based on a threshold of 97.5% confidence from the χ^2_n distribution since PS_i^2 roughly follows this distribution for a redundancy larger than 4, that is for $\zeta = m/n > 4$, when the points

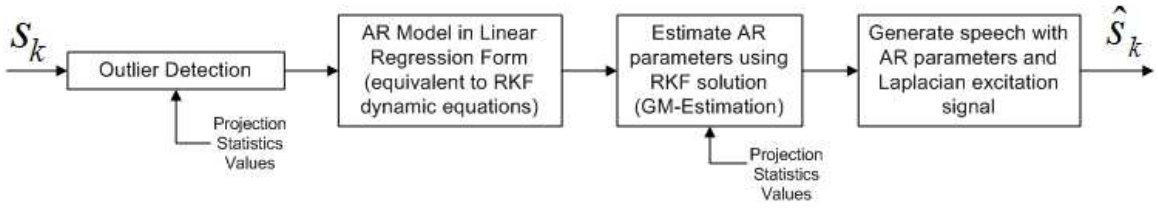


Figure 6.16: Proposed robust LPC processing is depicted in this figure.

are drawn from a multivariate standard Normal distribution.

Thus, to detect the outliers in the speech signal, a matrix $\mathbf{H}$ of an AR(4) containing consecutive values of the signal is built after the signal is segmented over 30 ms windows. The dimension of each vector $\mathbf{h}_i$ of $\mathbf{H}$ was chosen to be 4, which was empirically determined to yield good results. Each row of this matrix will have an associated Projection Statistic, which is compared to a threshold. For the given speech signal, we obtain Projection Statistics as shown in Figure 6.17. It is clear that the impulsive segments of speech are flagged with significantly higher PS values, representing outliers in the 4-dimensional space associated with the matrix $\mathbf{H}$.

6.7.2 Robust Linear Predictive Coding using GM-KF

Once the impulses have been removed, we need to replace the missing values in the signal. To this end, we apply an autoregressive model of order 20 to reconstruct speech samples along with a robust GM-estimator to estimate the AR model parameters. Outliers in $\mathbf{z}$, the response variable, are referred to as vertical outliers. Due to the AR model, an outlier in $\mathbf{z}$ will also be one in $\mathbf{h}$. These outlying points $(\mathbf{z}_i, \mathbf{h}_i)$ associated with the impulses are the bad leverage points that have a strong negative influence on an M-estimator solution in linear regression. Hence, we use a Schweppe-type Huber GM-estimator due to its bounded influence properties under infinitesimal contamination with a positive breakdown point; in fact, this estimator can handle bad leverage points, so long as the number of these outliers is within limits of the estimators breakdown point. All data points are given equal weight in least-squares estimation method; in contrast, the Schweppe-type Huber GM estimator involves Projection Statistics to obtain robust weights for the data points with respect to

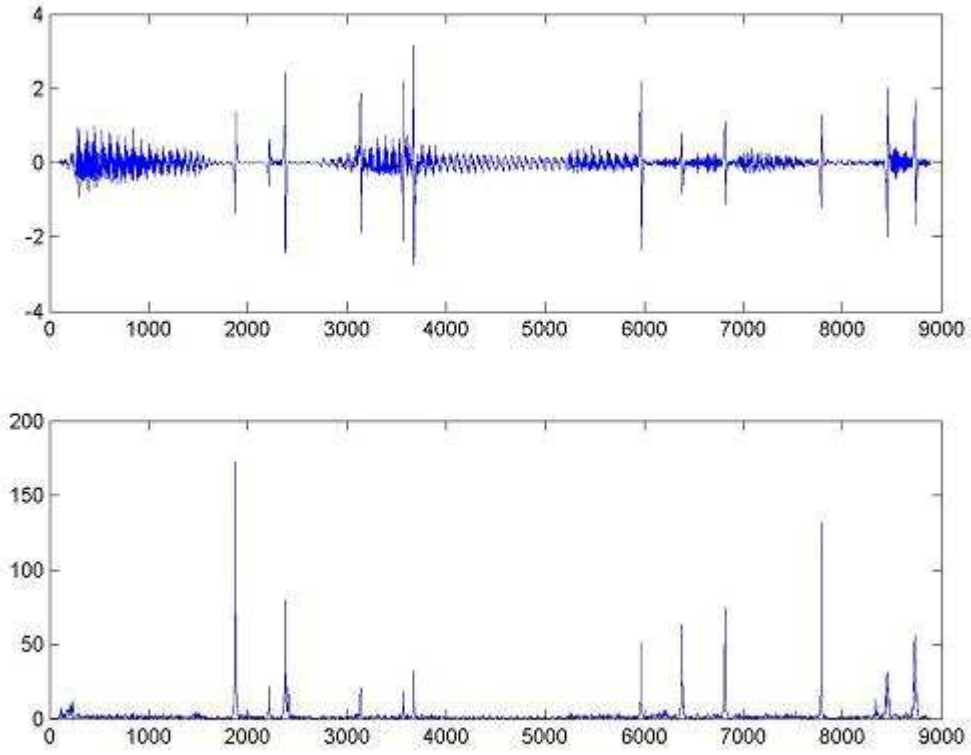


Figure 6.17: Resulting values for Projection Statistics are shown corresponding to the impulses induced in the speech signal. Note the good alignment between the two.

the bulk of the point cloud. The weight definition given by (4.87) provides robustness against bad leverage points by bounding the influence function. While the technique strongly down-weights bad leverage points arising from contributions of impulsive noise edges, it assigns weights of one to good leverage points arising from normal speech.

We now discuss the speech reconstruction process. Again, each of the points deemed as an impulsive noise sample is replaced with estimated values based on linear prediction with the parameter vector being obtained by (4.84). Accordingly, the following procedure is applied:

- Obtain a 120 sample signal segment immediately preceding the first outlier to be replaced;
- Formulate the $\mathbf{z}$ vector and $\mathbf{H}$ matrix using these 120 samples. The AR order is chosen to be 20, which is recognized as providing a good speech model [113].
- Estimate the parameter vector $\mathbf{x}$. Note that the weight matrix may be initialized with an identity matrix and iteratively derived. The value for MAD in the process is only estimated the first 4 times to avoid possible instability.
- Once $\mathbf{x}$ has been computed, excite the AR model with Laplacian noise (since speech follows closely the Laplacian distribution) to obtain the replacement samples.

Temporal and Statistical Results

We show four of the segments being reconstructed in Figure 6.18. It demonstrates the original spiky signal and the reconstruction. Also shown subsequently in Figure 6.19 is a plot of the associated projection statistics and final weights used in the parameter vector estimation for each of the signal samples over time corresponding to the third reconstructed segment in Figure 6.18. An appreciable suppression in impulsive noise can be noticed in Figure 6.20 when compared to the corrupted speech in Figure 6.14. Note the filtered speech signal overlapped onto the corrupted signal in Figure 6.21. Figure 6.22 allows us to better evaluate the improvement realized after filtering. Particularly, the overall and maximum errors in the clean/filtered signal pair have decreased

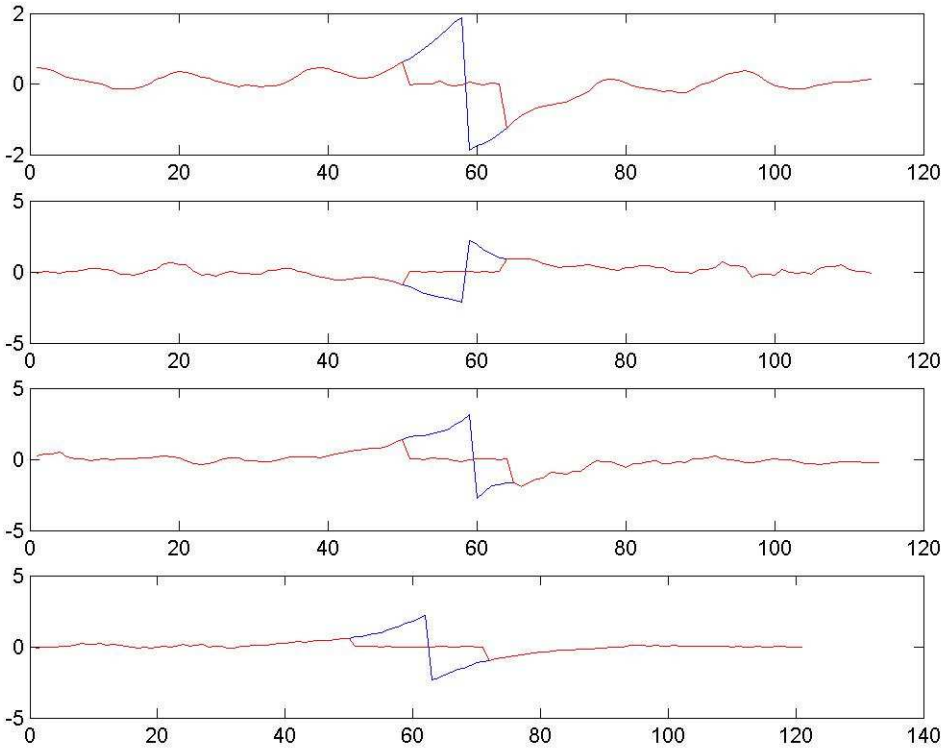


Figure 6.18: Four speech segments corrupted by impulsive noise are shown in black with speech reconstructed by the robust LPC overlaid in red.

appreciably in comparison to the clean/corrupted signal pair. Thus, the histogram of the robust distances (projection statistics) shows that there are the difference between the clean and filtered signal is smaller overall than between the clean and corrupted signal.

Spectral Results

We present an example from a specific reconstruction segment in the time domain and its corresponding spectrum in Figures 6.23 and 6.24. It is clear from the spectral image that the filtered signal (solid line with markers) follows the original speech spectrum (dotted line) much better when compared to the noisy signal spectrum (light solid line).

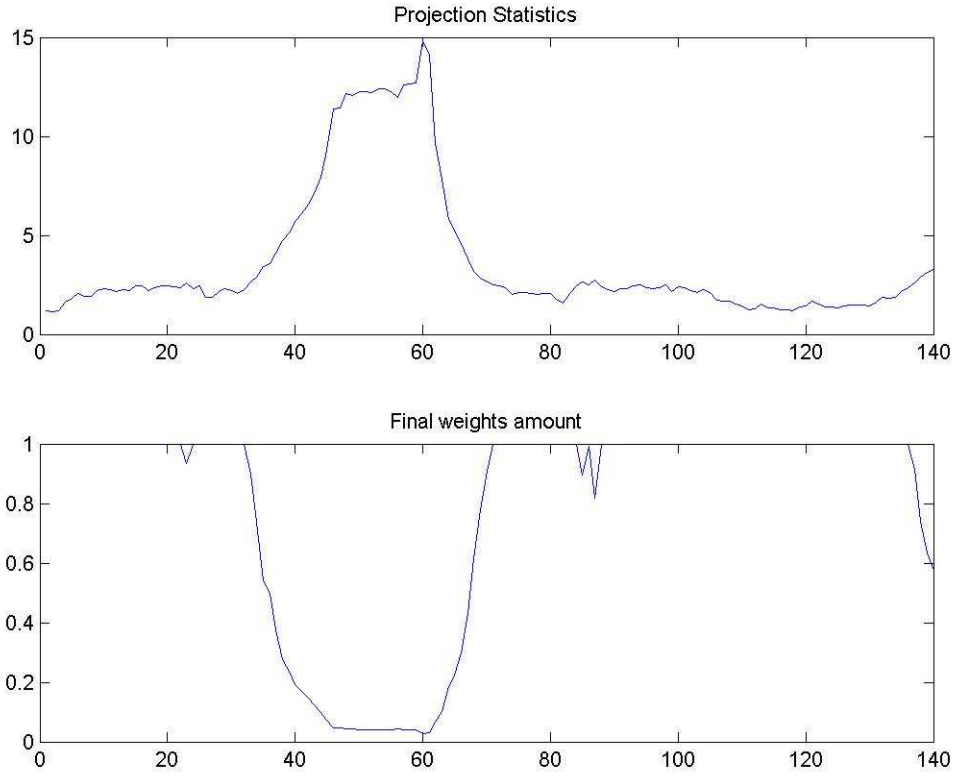


Figure 6.19: Projection statistics values and weights corresponding to the third speech segment depicted in Figure 6.18.

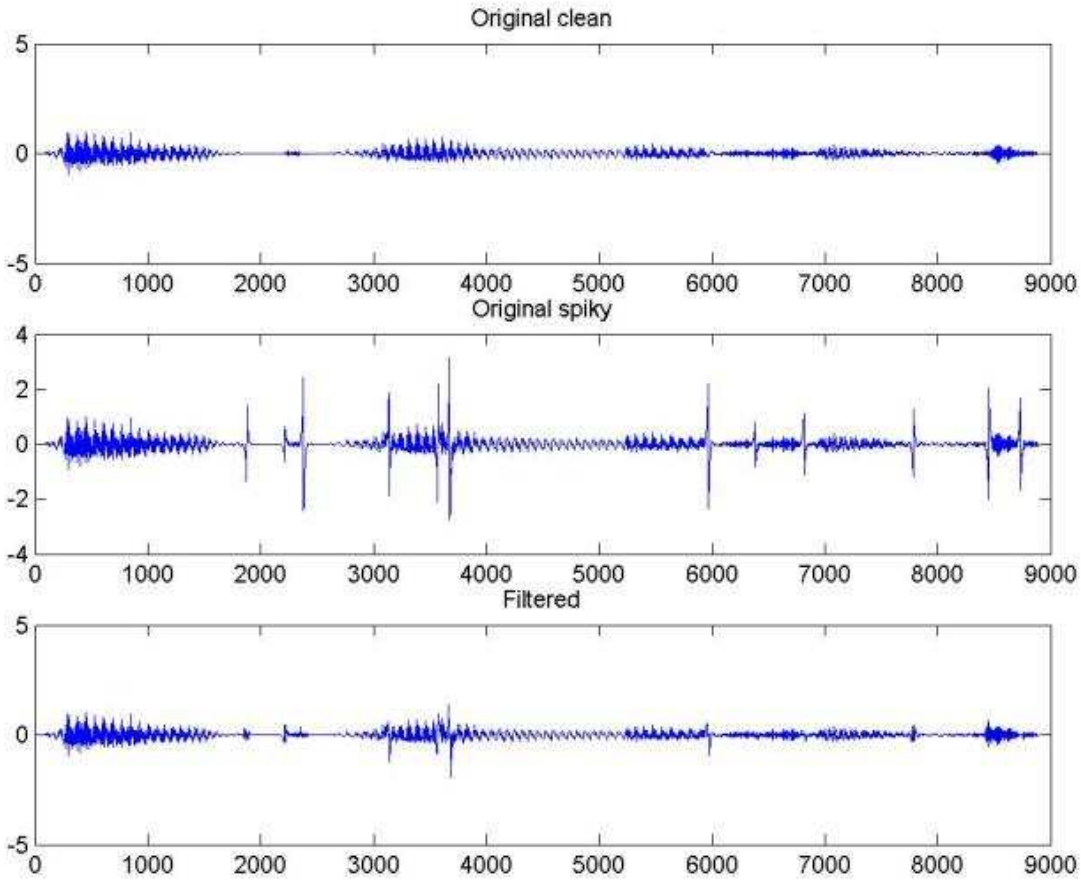


Figure 6.20: The speech waveform is its original form (top), corrupted by impulsive noise (center), and filtered and reconstructed using robust LPC and GM-KF method (bottom).

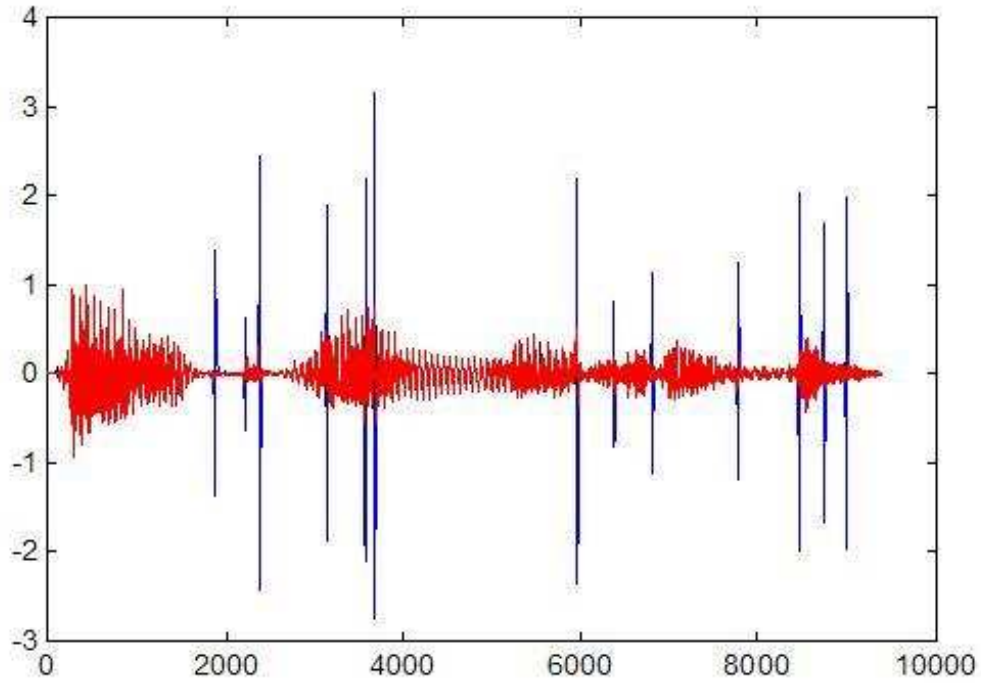
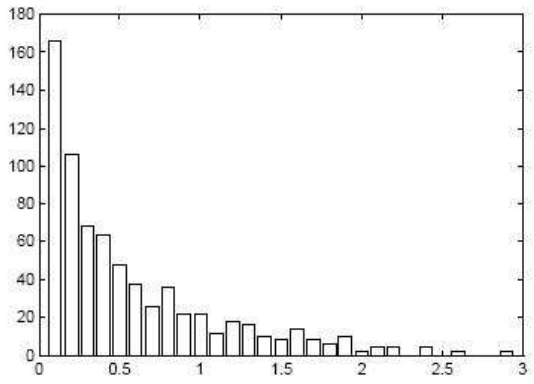
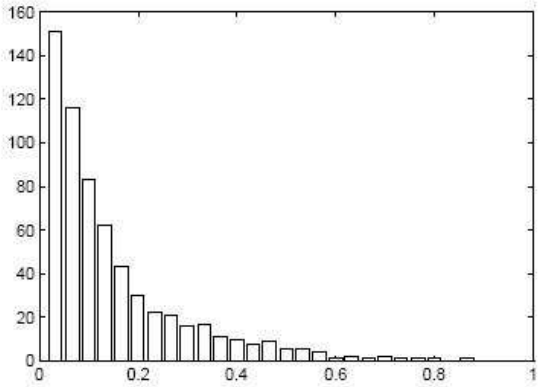


Figure 6.21: Original noisy speech waveform in black, overlaid with the reconstructed waveform in red.



Amount of difference between
original and corrupted waveforms on x-axis



Amount of difference between
original and filtered waveforms on x-axis

Figure 6.22: Histogram depicting a larger difference between the original and corrupted waveforms (left figure) when compared to the original and filtered waveform.

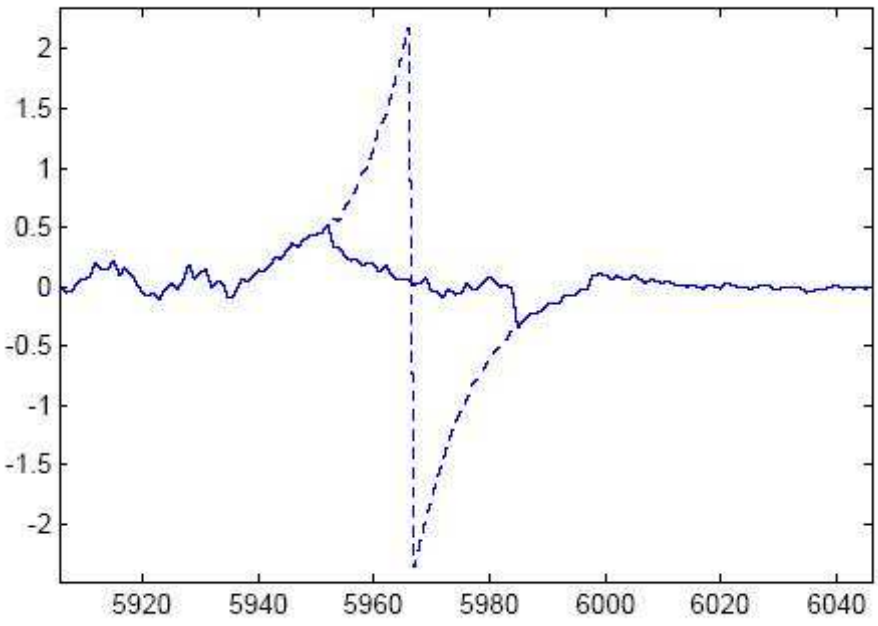


Figure 6.23: Sample time domain segment depicting the corruption by an outlier and reconstruction using the robust LPC and GM-KF method

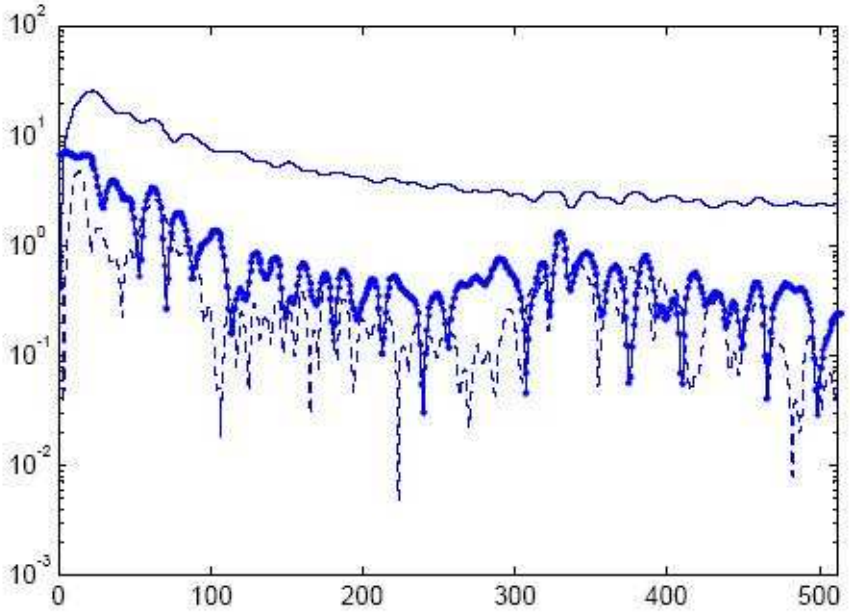


Figure 6.24: Spectral domain of the segment depicting the impulsive noise, original speech segment spectrum, and reconstructed spectrum.

Remarks

It should be noted that gain factor problems may arise if one is not careful with such an AR process. Note also that the edges of the impulse are often not detected based on the comparison between the threshold and the Projection Statistics. Therefore, we may see a rise in the filtered signal amplitude, representing portions that were not replaced immediately, followed by estimated lower amplitude values. To account for the edges, we followed a heuristic approach of replacing an additional 10 values preceding and following the first and last outliers. Due to many reasons, this time domain approach to the problem is advantageous over frequency domain processing, which has become default in many speech processing algorithms. First, Projection Statistics by itself is a fast algorithm as it does not require iterative calculations, making the impulse detection computationally efficient. The estimation takes $O(n)$ operations, but it leads to a reasonable solution at this computation level. We can obtain the FFT through $O(n\ log\ n)$ operations; however, the impulse detection and the more difficult task of re-estimating the corrupted segment still remain and would add significant computational complexity. Besides the computational advantage, the time domain approach is also more appealing and intuitive than the frequency domain counterpart from an algorithmic viewpoint of (a) noise detection and (b) speech estimation. Identifying the impulses along with the edges may not be straightforward task in frequency. In particular, the large PS values align very well with the impulses in the time domain; however, the strong vertical striations visible in the spectrogram are slightly smeared and eventually misaligned with the impulses after $1s$.

Chapter 7

Development and Application of the GM-Extended Kalman Filter

An optimal estimator for a linear system with additive Gaussian noise is the classical Kalman filter. In chapter 4, we proposed a new filter called GM-KF designed for robustness to outliers in the linear case. In the literature, the KF has been expanded to nonlinear systems with additive Gaussian noise in the form of the extended Kalman filter (EKF). In this chapter, we propose a new GM-EKF to achieve robust state estimation in the presence of sudden transient behavior caused by system process noise, or observation noise, or outliers in a class of nonlinear systems, including those that possess multiple equilibrium points and strong nonlinearity. Furthermore, the influence function of the GM-estimator under nonlinear regression is derived. Finally, we evaluate the performance of the EKF and GM-EKF through state estimation experiments on the Langevin model, which is commonly used to model climate transitions.

7.1 Nonlinear Systems with Multiple Equilibrium Points

The focus of this chapter is robust state estimation in nonlinear systems with multiple equilibrium points. It has been observed that such systems often have small fluctuations around an equilibrium

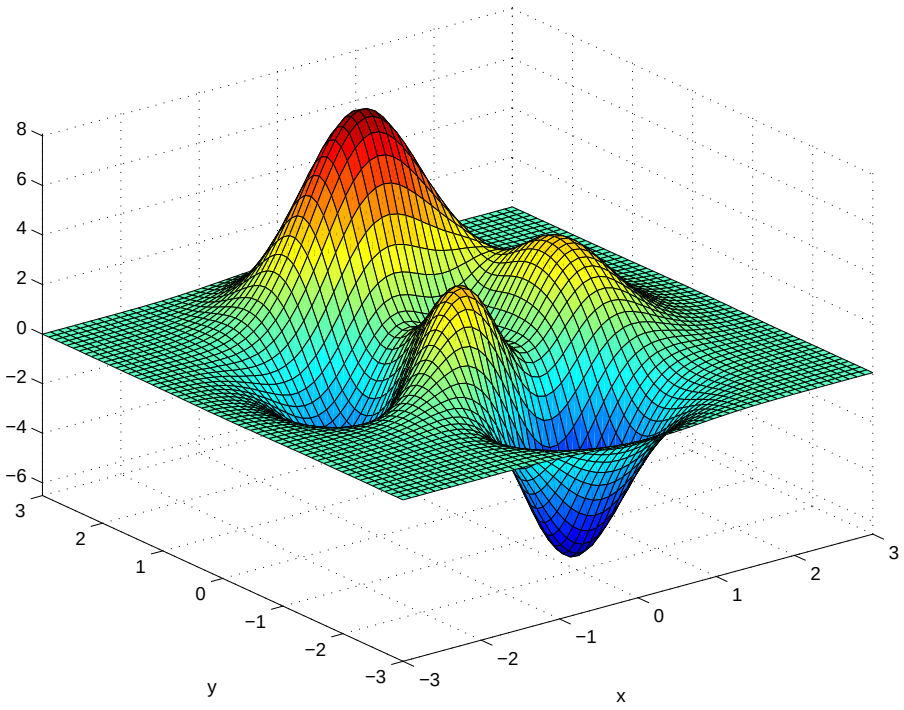


Figure 7.1: A nonlinear system with 3 stable equilibrium points.

point for some time, followed by a sudden shift in the system state to another equilibrium point. The buckling of an elastic beam is one example of such behavior from solid mechanics [93]. The system is in a state of equilibrium in terms of its structural properties under no outside pressure. The beam will shorten according to Hooke's law under outside compression [93], which can be seen as fluctuations around the equilibrium point. But, the beam buckles when the compression force is large enough and the linear theory fails [93]. In other words, the fluctuation around the equilibrium point is large enough causing the system to suddenly shift to a different equilibrium point.

As a general example, let us consider the system depicted in Figure 7.1 and the associated contour plot in Figure 7.2. This system contains three stable equilibrium points, with uneven saddle surfaces that act as boundaries between three regimes of the system. The system will reside in one of the basins of attraction for a certain amount of time, followed by a transition from one basin to another driven by noise with large enough energy or an external control input. Of course,

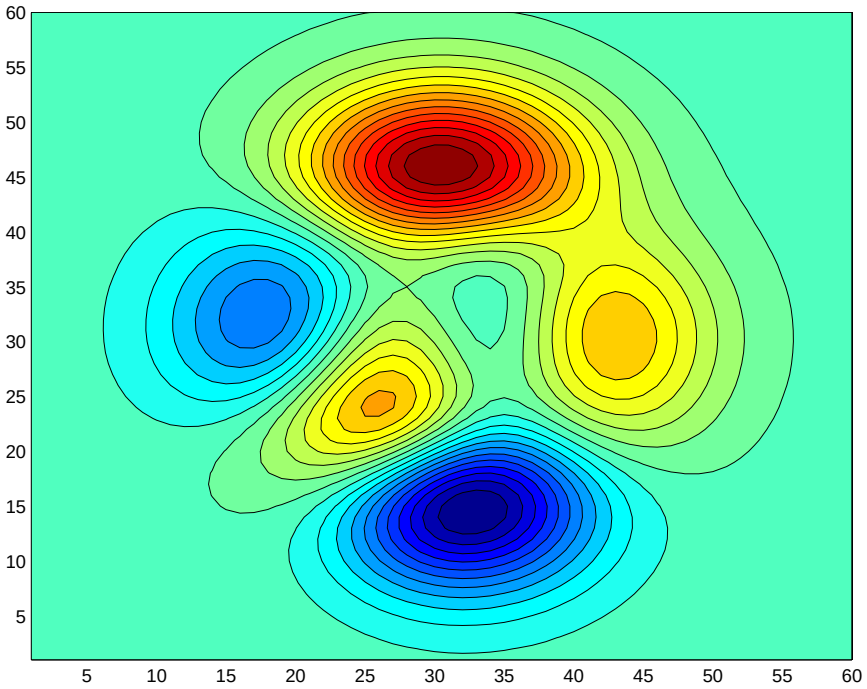


Figure 7.2: Contour plot of a nonlinear system with 3 stable equilibrium points.

the amount of energy required depends on how deep the basin of attraction is compared to the lowest saddle point along the boundary.

While the noise is often assumed to follow a Gaussian PDF, the assumed noise model is only an approximate one and the two types of noises may actually follow thick-tailed, non-Gaussian PDFs, inducing innovation and observation outliers. These outliers may once again easily lead to erroneous state estimates. Thus, a robust filter is desired that is able to output the correct qualitative state quickly in the presence of both outliers and strong, nonlinear, switch-like transitions [19, 35, 101]. Furthermore, the method should be able to provide reliable estimates over time, i.e. a "most probable history" of the states [35, 101]. In the case of the preceding example, this method's output should correctly indicate whether the beam is buckled or not.

This need is of growing importance in various fields, including meteorology, physical oceanography, and paleoclimatology [47, 143]. The vast majority of numerical techniques in these areas are based on variants of the least-squares estimator and the Kalman filter [93]. The KF is designed

for linear models, and therefore, is not effective in reliably tracking the nonlinear state transitions from one equilibrium point to another. It is not robust to outliers either.

Various other methods have been considered in these fields [35, 93, 101], including the least squares variational method (LSV) and the interactive Kalman filter. The LSV is actually an example from a class of methods developed in statistical physics, in which variational principles formulated as an optimization problem to determine the time-sequence of states have been investigated [35, 101]. A similar method has been developed in thermodynamics, in which the most probable state sequence is determined by minimizing the so-called Onsager-Machlup action function [35, 101, 102]. However, neither of these two methods are able to track the transient well because they are generally developed for problems with weak noise [35], which means that the probability of large deviations from an operating point of a nonlinear system is very low. In other words, sudden shifts from one equilibrium point to another are assumed to be unlikely, an assumption that is violated in many practical problems.

Burger and Cane [19] proposed the method known as the interactive Kalman filter, in which several error models are simultaneously used to solve for the state estimate. In this approach, several PDFs are explicitly assumed for the noise processes, and the system's state estimate is determined by the noise PDF that most closely represents the current system behavior. One disadvantage of this method is the need for a priori definition of specific noise models. In fact, the technique may easily lead to large computational costs if one uses too many noise models to represent the system, or on the other hand, inaccurate results if an insufficient number of noise regimes is applied to model the system.

Yet other methods from the literature that may be employed to track state transitions include the particle filter, particle Kalman filter, path integration, the ensemble Kalman filter and its second-order variant, and the singular extended interpolated filter [35, 60, 93, 107]. However, these techniques are often very complicated and computationally very intensive, requiring the use of Monte-Carlo methods in determining how the system's PDF evolves over time [92].

One popular method for state estimation in nonlinear systems is the extended Kalman filter

(EKF). It is based on a nonlinear model assumed in the prediction stage, over which linearization and discretization are performed around the previous filtered state estimate to obtain a set of discretized, linear perturbation equations. The EKF equations are then derived from the classical KF recursion. While the EKF gives good results under the assumptions, unfortunately it yields rather poor performance in the presence of transients and outliers.

This weak performance is attributed to several reasons. First of all, similar to the LSV, the EKF is developed for problems with weak noise; therefore, it does not track the system transitions well in the presence of Gaussian noise, and yields worse solutions or diverges completely in the presence of outliers [35, 133]. An example of the EKF's performance under weak and strong noise can be found in the work of Picard [108]. Secondly, the EKF strongly depends on accurate observations to track transitions caused by the system process noise. To understand this, we recognize that the state predictions based on the system model do not directly incorporate values of noise, and therefore, are unable to sense when a transition occurs due to system process noise. Thirdly, because the EKF is very dependent on hte observations, it may undergo a transition due to an observation outlier, ignoring a good prediction even when the system has not actually shifted between the basins of attraction. Finally, when the observation noise covariances are sufficiently larger than the system process noise covariances, the EKF may miss a system transition by ignoring good observations while incorrectly relying on predictions, which are not detecting the transitions in the first place. It may even initiate a transition when an innovation outlier is present in the predictions. Note that an outlier present in a control input vector also impacts the filter in the same manner as an innovation outlier does by directly affecting the prediction.

In contrast to all of these methods, we propose in this chapter a new filter to reliably track the states of a nonlinear dynamic system. Our filter is simple and robust against outliers. Its design framework is similar to the GM-KF and consists of three key steps, which include (a) creating a redundant observation vector, (b) performing robust prewhitening, and (c) robustly estimating the state vector. This framework allows us to employ one of many robust estimators within the extended Kalman filter methodology, so long as that estimator's covariance matrix can be calculated. Here,

we use the generalized maximum likelihood-type estimator again, yielding our proposed filter known as the GM-EKF.

In the following sections, we first study the Langevin model as a sample system that exhibits the transient behavior described in Section 7.1. Then, we revisit the standard EKF in Section 7.3 and recognize its limitations in tracking the state transitions from one equilibrium point to another in Section 7.5. Finally, we develop the robust GM-EKF scheme in Section 7.6 and use it to quickly and reliably track climate transitions under this model in Section 7.9.

7.2 The Langevin Model

A simple example of a nonlinear system exhibiting multiple equilibrium points is given by the dynamic equations expressed as [93]

$$\dot{x}_t = f(x_t) = -4x_t(x_t^2 - 1), \tag{7.1}$$

where the right-hand side can be viewed as the negative gradient of a potential function, $U(x_t)$, given by

$$U(x_t) = x_t^4 - 2x_t^2. \tag{7.2}$$

With a stochastic forcing term, the dynamic equation becomes

$$\dot{x}_t = 4x_t - 4x_t^3 + \kappa n_t, \tag{7.3}$$

which is termed the Langevin model when n_t is a white-noise process.

The double-well potential function represented by (7.2) is shown in Figure 7.3, with the negative of the gradient, $f(x)$, shown in Figure 7.4. It is clear that without any external forcing term or noise, this system has three equilibrium points at $x = 0$, $x = 1$, and $x = -1$, of which the former is unstable and the latter two are stable. In other words, starting at any point, the system will

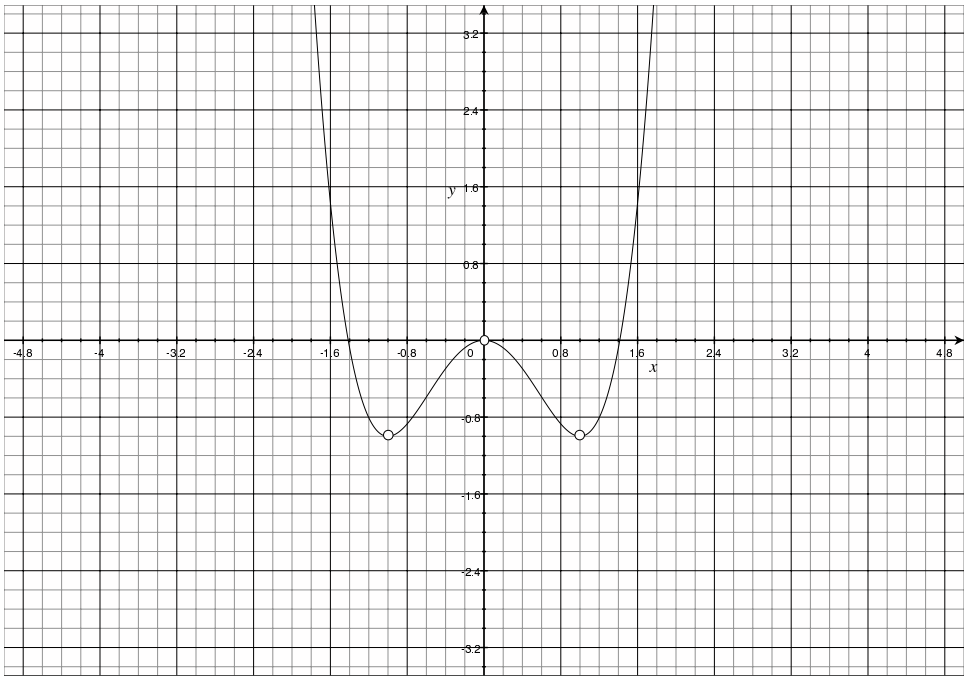


Figure 7.3: The double-well potential function, $U(x)$, shown with its 3 equilibrium points.

approach and settle at either $x = 1$ or $x = -1$. With small forcing or noise terms, the system will fluctuate around one of these two equilibrium points. With a large enough noise energy or external forcing term driving the system in the correct direction, the system shifts from one basin of attraction to another [19, 93], which is termed a transition in the system.

In practice, similar models are often used as a simple representation of climate system exhibiting transient behavior [13, 14, 35, 93, 99, 142]. The function $U(x_t)$ given by (7.2) can be interpreted as the climate potential, with x_t representing the average surface temperature of the earth at time t. The right minimum can be viewed as the present climate state while the other equilibrium represents ice ages, for example [35, 99]. With small values of κ, the state dynamics consists of small fluctuations around an equilibrium point for a long time with a large fluctuation leading to transitions occasionally. However, such behavior with small κ values is not very applicable to long-time climate dynamics that are analyzed on geological time scales, in which transitions occur frequently.

The evolution of the probability distribution of the system in (7.3) can also be derived explicitly.

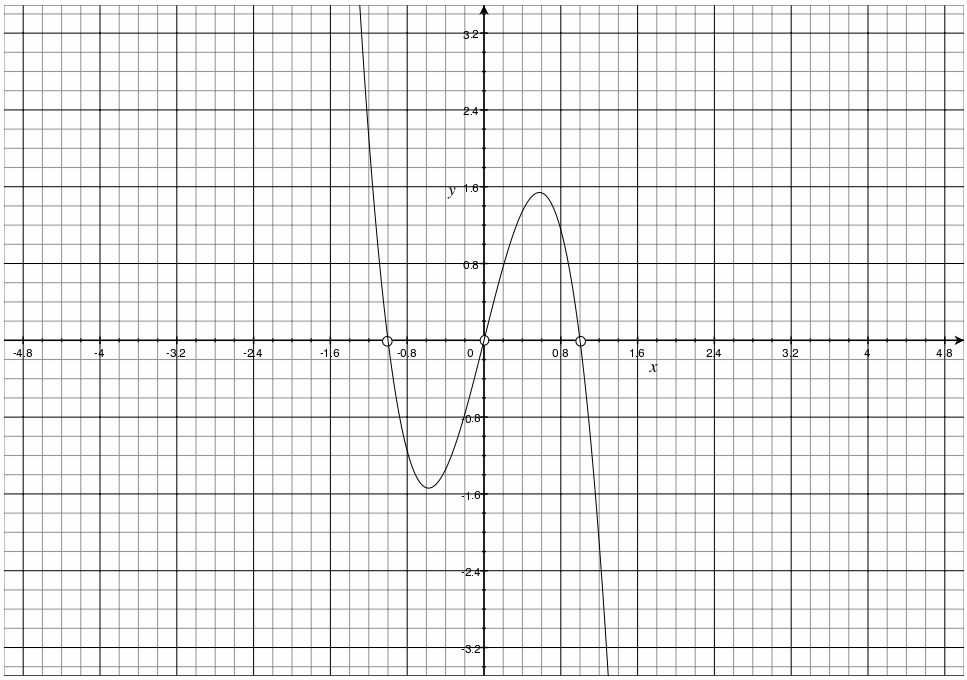


Figure 7.4: The double-well system dynamic equation, $f(x) = -U'(x)$, shown with its 3 equilibrium points.

In general, the PDF $F(\mathbf{x}_t, t)$ of the state $\mathbf{x}_t$ described by the Ito stochastic differential equation, given by

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + \mathbf{g}(\mathbf{x}_t, t)d\mathbf{b}_t, \tag{7.4}$$

can be obtained through the Fokker-Planck equation, where $\mathbf{g}(\mathbf{x}_t, t)$ is an arbitrary function, $\mathbf{b}_t$ is a Wiener process, and $\mathbf{f}(\mathbf{x}_t, t)$ specifies the system dynamics. Also known as the forward Kolmogorov equation, it is expressed as

$$\frac{\partial F(\mathbf{x}_t, t)}{\partial t} = -\nabla \cdot (\mathbf{f}(\mathbf{x}_t, t)F(\mathbf{x}_t, t)) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}(\mathbf{R}_g)_{ij} F(\mathbf{x}_t, t), \tag{7.5}$$

where $\mathbf{R}_g = \mathbf{g}(\mathbf{x}_t, t)\mathbf{g}^T(\mathbf{x}_t, t)$ [92, 93]. For the climate model described in (7.3), the Fokker-Planck equation is expressed as

$$\frac{\partial F(x_t, t)}{\partial t} = \frac{\partial}{\partial x_t}[4x_t(x_t^2 - 1)F(x_t, t)] + \frac{1}{2}\kappa^2 \frac{\partial^2 F(x_t, t)}{\partial x_t^2}, \tag{7.6}$$

which yields the steady state PDF solution [117] of

$$F(x_t, t) \propto e^{-2U(x_t,t)/\kappa^2}, \tag{7.7}$$

where $U(x_t, t)$ is given by (7.2).

As noted before, tracking the evolution of the PDF directly over time is a very computationally intensive task because Monte-Carlo type solutions such as the particle filter are needed. Instead, as Miller indicated [93], we consider a simpler data assimilation technique such as the EKF and assess its capabilities to follow state transitions on this simple model. We now revisit the formulation of the standard extended Kalman filter.

7.3 Review of the Extended Kalman Filter

Recall from Section 2.4.1 that the nonlinear system dynamics, assuming a vector state variable, is represented by the following equations [4, 88, 104]:

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t) + \mathbf{w}_t + \mathbf{u}_t, \tag{7.8}$$

$$\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t) + \mathbf{e}_t, \tag{7.9}$$

where $\mathbf{f}(\mathbf{x}_t)$ and $\mathbf{h}(\mathbf{x}_t)$ are assumed to be continuous and continuously differentiable with respect to all elements of the state vector, $\mathbf{x}_t$. In (7.9), the observations contained in $\mathbf{z}_t$ are treated as being obtained at discrete intervals. Let us denote by $\mathbf{W}_t$ and $\mathbf{R}_t$ the noise covariance matrices in the continuous time. To solve for the state estimates, the model must first be linearized around the previous corrected state estimate. Applying a first-order Taylor series expansion to the nonlinear equation and assuming the higher order terms are negligible, we get

$$\mathbf{f}(\mathbf{x}_t) = \mathbf{f}(\mathbf{x}_t^*) + \mathbf{F}_x(\mathbf{x}_t^*)\delta\mathbf{x}_t, \tag{7.10}$$

where $\mathbf{F}_x$ is the Jacobian matrix defined as

$$\mathbf{F}_x = \frac{\partial \mathbf{f}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\mathbf{x}_t^*} \tag{7.11}$$

and $\mathbf{x}_t^*$ is the nominal state vector. Let $\mathbf{x}_0$ denote the initial condition. For successive iterations, the corrected estimate $\hat{\mathbf{x}}_{k-1|k-1}$ is assigned as the nominal value at time k. The perturbation from the nominal value, $\delta\mathbf{x}_t$, is defined as

$$\delta\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_t^*, \tag{7.12}$$

where the nominal state vector satisfies the equation

$$\dot{\mathbf{x}}_t^* = \mathbf{f}(\mathbf{x}_t^*)\,. \tag{7.13}$$

Finally, substituting (7.10) into (7.8) yields the following linear perturbation state equation:

$$\delta\dot{\mathbf{x}}_t = \mathbf{F}_x(\mathbf{x}_t^*)\delta\mathbf{x}_t + \mathbf{w}_t + \mathbf{u}_t. \tag{7.14}$$

Linearizing (7.9) in a similar manner yields the following perturbation equation:

$$\delta\mathbf{z}_t = \mathbf{H}_x(\mathbf{x}_t^*)\delta\mathbf{x}_t + \mathbf{e}_t, \tag{7.15}$$

where $\mathbf{H}_x$ is a Jacobian matrix given by

$$\mathbf{H}_x = \frac{\partial \mathbf{h}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\hat{\mathbf{x}}_t^*}. \tag{7.16}$$

Next, using these perturbation equations and assuming zero-order hold on any inputs and continuous integration of the noise, the nonlinear model can be discretized and put into the following form

[17, 88]:

$$\mathbf{x}_{k+1} = \mathbf{F}_d \mathbf{x}_k + \mathbf{w}_k + \mathbf{B}_d \mathbf{u}_k, \tag{7.17}$$

$$\mathbf{z}_k = \mathbf{H}_d \mathbf{x}_k + \mathbf{e}_k, \tag{7.18}$$

where

$$\mathbf{F}_d = e^{\mathbf{F}_x T_s}, \tag{7.19}$$

$$\mathbf{H}_d = \mathbf{H}_x, \tag{7.20}$$

and T_s is the time step. To see this, the linearized model in (7.14) needs to be solved. A detailed derivation that shows the homogenous and particular solutions of this equation can be found in various texts [17, 36, 65, 88]. Here, we observe a simpler way to reach the solution when zero-order hold is assumed [17]. First, we multiply the linearized model

$$\dot{\mathbf{x}}_t = \mathbf{F}_x \mathbf{x}_t + \mathbf{w}_t + \mathbf{u}_t, \tag{7.21}$$

with an exponential function of time, $e^{-\mathbf{F}_x t}$, to get

$$e^{-\mathbf{F}_x t} \dot{\mathbf{x}}_t = e^{-\mathbf{F}_x t} \left[\mathbf{F}_x \mathbf{x}_t + \mathbf{w}_t + \mathbf{u}_t \right], \tag{7.22}$$

which can be rewritten as

$$\frac{d}{dt} \left(e^{-\mathbf{F}_x t} \mathbf{x}_t \right) = e^{-\mathbf{F}_x t} \left[\mathbf{w}_t + \mathbf{u}_t \right]. \tag{7.23}$$

Then, we obtain the analytical solution of (7.23) by integration as follows:

$$e^{-\mathbf{F}_x t} \mathbf{x}_t - e^0 \mathbf{x}_0 = \int_0^t e^{-\mathbf{F}_x \tau} \left[\mathbf{w}_\tau + \mathbf{u}_\tau \right] d\tau, \tag{7.24}$$

$$\mathbf{x}_t = e^{\mathbf{F}_x t} \mathbf{x}_0 + \int_0^t e^{\mathbf{F}_x (t-\tau)} \left[\mathbf{w}_\tau + \mathbf{u}_\tau \right] d\tau. \tag{7.25}$$

To discretize the last equation, let us define $\mathbf{x}_k \equiv \mathbf{x}(kT_s)$. Then,

$$\mathbf{x}_k = e^{\mathbf{F}_x kT_s}\mathbf{x}_0 + \int_0^{kT_s} e^{\mathbf{F}_x(kT_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau, \tag{7.26}$$

and

$$\begin{aligned}
\mathbf{x}_{k+1} &= e^{\mathbf{F}_x(k+1)T_s}\mathbf{x}_0 + \int_0^{(k+1)T_s} e^{\mathbf{F}_x((k+1)T_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau, && (7.27)\\
&= e^{\mathbf{F}_x(k+1)T_s}\mathbf{x}_0 + \int_0^{kT_s} e^{\mathbf{F}_x((k+1)T_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau \\
&\qquad + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{F}_x((k+1)T_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau, && (7.28)\\
&= e^{\mathbf{F}_x T_s}\left[e^{\mathbf{F}_x kT_s}\mathbf{x}_0 + \int_0^{kT_s} e^{\mathbf{F}_x(kT_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau\right] \\
&\qquad + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{F}_x((k+1)T_s-\tau)}\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]d\tau. && (7.29)
\end{aligned}$$

Note that the expression in the bracket of the last equation is $\mathbf{x}_k$ as given in (7.26), resulting in

$$\mathbf{x}_{k+1} = e^{\mathbf{F}_x T_s}\mathbf{x}_k + \int_{kT_s}^{(k+1)T_s} e^{\mathbf{F}_x((k+1)T_s-\tau)}d\tau\left[\mathbf{w}_\tau + \mathbf{u}_\tau\right]. \tag{7.30}$$

Equation 7.30 can be written using a simpler notation by substituting $\tau_s = kT_s + T_s - \tau$, yielding

$$\mathbf{x}_{k+1} = e^{\mathbf{F}_x T_s}\mathbf{x}_k + \int_0^{T_s} e^{\mathbf{F}_x\tau_s}d\tau_s\left[\mathbf{w}_{\tau_s} + \mathbf{u}_{\tau_s}\right]. \tag{7.31}$$

Assuming that the noise and control vectors are constant during the integration period, we can rewrite (7.31) as

$$\mathbf{x}_{k+1} = e^{\mathbf{F}_x T_s}\mathbf{x}_k + \mathbf{w}_k + \mathbf{B}_d\mathbf{u}_k, \tag{7.32}$$

where $\mathbf{w}_k$ is statistically equivalent through its first two moments to

$$\int_0^{T_s} e^{\mathbf{F}_x\tau_s}\mathbf{w}_{\tau_s}d\tau_s, \tag{7.33}$$

and

$$\mathbf{B}_d = \int_0^{T_s} e^{\mathbf{F}_x \tau_s} d\tau_s. \tag{7.34}$$

The mean and covariance matrix of $\mathbf{w}_k$ is given as

$$E[\mathbf{w}_k] = E\left[\int_0^{T_s} e^{\mathbf{F}_x \tau_s}\ \mathbf{w}_{\tau_s}\ d\tau_s\right] = \mathbf{0}, \tag{7.35}$$

and

$$\mathbf{W}_k = E[\mathbf{w}_k \mathbf{w}_k^T] = \int_0^{T_s} e^{\mathbf{F}_x \tau_s}\ \mathbf{W}_t\ e^{\mathbf{F}_x \tau_s} d\tau_s. \tag{7.36}$$

Assuming the observations are available periodically, we can write the complete linearized and discretized state space equations as

$$\mathbf{x}_{k+1} = \mathbf{F}_d \mathbf{x}_k + \mathbf{w}_k + \mathbf{B}_d \mathbf{u}_k, \tag{7.37}$$

$$\mathbf{z}_k = \mathbf{H}_d \mathbf{x}_k + \mathbf{e}_k, \tag{7.38}$$

as indicated in (7.17) and (7.18). Because this is a discrete system with linear matrices, the solution can be written similarly to the traditional linear Kalman filter recursions as follows:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_d \hat{\mathbf{x}}_{k-1|k-1} + \int_{t_{k-1}}^{t_k} \mathbf{f}(\hat{\mathbf{x}}_{k-1|k-1})\ dt + \mathbf{B}_d \mathbf{u}_k, \tag{7.39}$$

$$\mathbf{\Sigma}_{k|k-1} = \mathbf{F}_d \mathbf{\Sigma}_{k-1|k-1} \mathbf{F}_d^T + \mathbf{W}_k, \tag{7.40}$$

$$\mathbf{K}_k = \mathbf{\Sigma}_{k|k-1} \mathbf{H}_d^T\ [\mathbf{H}_d \mathbf{\Sigma}_{k|k-1} \mathbf{H}_d^T + \mathbf{R}_k]^{-1}, \tag{7.41}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k[\mathbf{z}_k - \mathbf{h}(\hat{\mathbf{x}}_{k|k-1})], \tag{7.42}$$

$$\mathbf{\Sigma}_{k|k} = \mathbf{\Sigma}_{k|k-1} - \mathbf{K}_k \mathbf{H}_d \mathbf{\Sigma}_{k|k-1}. \tag{7.43}$$

7.4 Numerical Integration Techniques for the EKF

Note that the integral in the prediction formula given by (7.39) may not have an analytical solution. Numerical integration techniques are necessary in such situations. In this section, the first-order

Euler and fourth-order Runge-Kutta methods are presented following Gerald [46]. One of the simplest techniques, the Euler method uses a first-order Taylor series approximation of an ordinary differential equation to find successive points in the solution. Given a differential equation and a set of initial conditions respectively given by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}_t), \tag{7.44}$$

$$\mathbf{x}_{t_0} = \mathbf{x}(t_0), \tag{7.45}$$

one step of the Euler method from t_k to $t_{k+1} = t_k + T_s$ is expressed as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + T_s\mathbf{f}(\mathbf{x}_k). \tag{7.46}$$

In essence, the Euler method assumes that the slope over one step is constant and equivalent to the value of the slope at the beginning of the step. The errors arising from this method can be understood by comparing (7.46) with a Taylor series expansion, given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + T_s\dot{\mathbf{x}}_k + O(T_s^2). \tag{7.47}$$

Particularly, the local error introduced in a single iteration is represented by the term $O(T_s^2)$ and accumulates over multiple iterations on the order of $O(T_s)$, i.e. it is a first-order technique. In contrast, the 4^{th}-order Runge-Kutta method is much more accurate as it approximates the differential equation with a higher order Taylor series expansion. Specifically, this method performs numerical integration using the recursive equation expressed as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{T_s}{6}(n_1 + 2n_2 + 2n_3 + n_4) + O(T_s^5), \tag{7.48}$$

where

$$n_1 = \mathbf{f}(t_k, \mathbf{x}_k), \tag{7.49}$$

$$n_2 = \mathbf{f}\left(t_k + \frac{T_s}{2}, \mathbf{x}_k + \frac{T_s}{2} n_1\right), \tag{7.50}$$

$$n_3 = \mathbf{f}\left(t_k + \frac{T_s}{2}, \mathbf{x}_k + \frac{T_s}{2} n_2\right), \tag{7.51}$$

$$n_4 = \mathbf{f}(t_k + T_s, \mathbf{x}_k + T_s n_3). \tag{7.52}$$

7.5 Application of the EKF to the Langevin Equation

We now apply the standard EKF to the Langevin model and study how well it tracks state transitions. It has been discussed in the literature [19, 35, 92, 93] that the EKF may perform satisfactorily only under certain conditions, and fails to track the transitions otherwise. We evaluate the performance of the filter under the following conditions:

1. A system resident around one of its stable equilibrium points, with no transitions;
2. Varying number of observations available for processing;
3. System perturbed by different system process noise variance, observation noise variance, and outliers;
4. State transitions in the system occurring very rapidly.

The specific test cases incorporating these conditions are elaborated further in Table 7.1. Complete results for these test cases have been presented in Appendix A. Figure 7.5 shows the EKF state estimation for case 1, which can be considered in a general sense as evaluating the statistical efficiency of the filter (i.e. how well the filter works at the Gaussian distribution). As expected, the EKF estimates the states with good accuracy as seen by the very low MSE values depicted in the bottom frame of the figure. Note that Gaussian ambient noise is assumed, with values of

Table 7.1: Test Cases for Evaluating the EKF and GM-EKF on a Double-Well System

	Transition Length (samples)	σ_x^2	σ_z^2	Sampling Frequency (Hz)	Sampling Period (s)	kappa
Case 1: Statistical efficiency test for the filter	-	0.01	0.01	4.00	0.25	0.75
Case 2: Effects of low number of observation	-	0.01	0.01	4.00	0.25	0.75
	-	0.01	0.01	1.00	1.00	0.75
	-	0.01	0.01	0.50	2.00	0.75
Case 3: Effects of high observation noise variances	-	0.01	0.02	4.00	0.25	0.75
	-	0.01	0.07	4.00	0.25	0.75
	-	0.01	0.08	4.00	0.25	0.75
Case 4: Rapid transitions with increasing observation noise variances	5	0.01	0.04	4.00	0.25	0.75
	5	0.01	0.05	4.00	0.25	0.75
Case 5: Observation outliers in the absence of an actual state transition	-	0.01	0.01	4.00	0.25	0.75
Case 6: Low number of observations with high observation noise variance		0.01	0.02	4.00	0.25	0.75
				2.00	0.50	
				1.00	1.00	
				0.67	1.50	
				0.50	2.00	

$\sigma_z^2 = 0.01$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$, where σ_z^2 and σ_x^2 are the observation and system process noise variances, respectively.

Interesting limitations of the EKF begin to surface when we simulate the system with transitions and observe how quickly, if at all, the filter tracks the system's shift in equilibrium points. In the subsequent results, the key values in the system are set as follows: $\sigma_z^2 = 0.01$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$. Figure 7.6 depicts the filtering result when a new observation is available every time a prediction is made, i.e. sampling period of 0.25 seconds and corresponding sampling frequency of 4 Hz. Under these conditions, the EKF tracks the transition with a time delay of 2 seconds. When the available number of observations is reduced as described in the literature [35, 92, 93], the EKF tracks the transition but with larger delays. In Figure 7.7, every third sample of the observation is available, i.e. a sampling period of 1 second and sampling frequency of 1 Hz. The time delay increases to 3 seconds in this condition. Reducing the frequency further to 0.50 Hz increases the time delay and causes the filter to oscillate until the transition is tracked properly, as seen in Figure 7.8. Additional results for sampling frequencies of 2 Hz and 0.67 Hz are given in Appendix A.

Next, we evaluate the performance of the EKF under varying noise variances. Particularly, it is expected that if the observation noise variance is large enough in comparison to the system process noise and filter estimation error variances, the filter will completely rely on the predictions, potentially missing one or more transitions completely. Figure 7.9 depicts a simulation trial where the key noise values are set equal to $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$, for which the filter is able to track the transition. When the observation noise variance is increased to $\sigma_z^2 = 0.07$, we see a significant increase in the time taken to track the transition, as shown in Figure 7.10. Indeed, the filter takes over 25 seconds to track the transition, which is 100 observation samples at the sampling frequency of 4 Hz in this experiment. Finally, as shown in Figure 7.11, when the observation noise variance is increased to $\sigma_z^2 = 0.08$, the EKF does not track the state transition even after 160 observation samples are processed over 40 seconds after the system transition occurred. More examples of this pattern can be found in Appendix A.

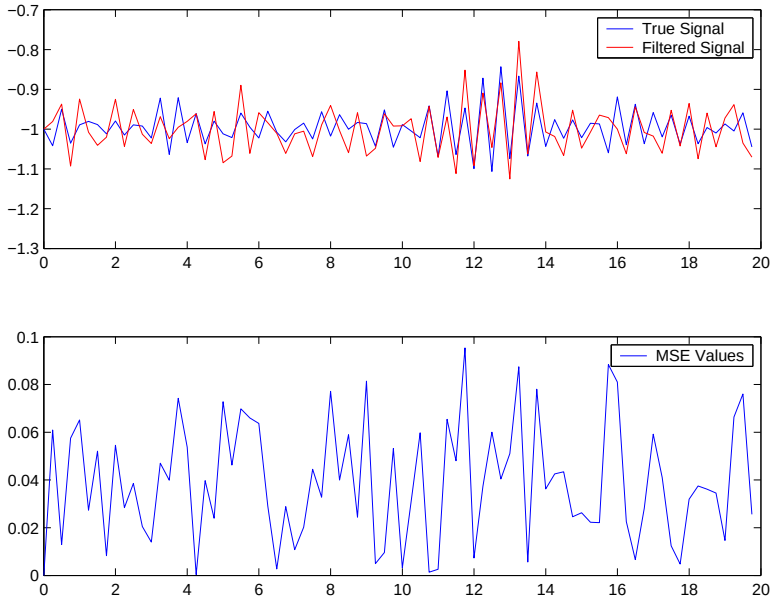


Figure 7.5: Case 1 - The EKF performs satisfactory state estimation in a double-well system that is residing around one of its stable equilibrium points. Lower frame depicts the associated MSE values for the estimation segment shown in the upper frame.

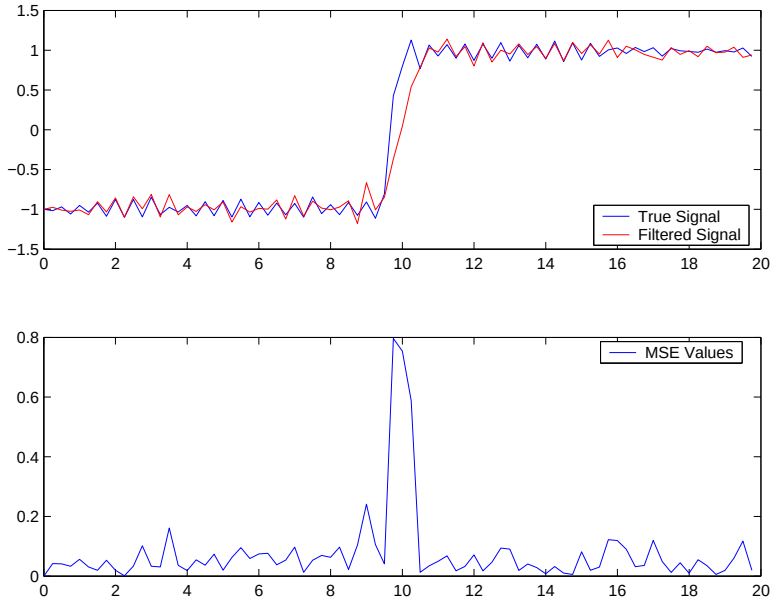


Figure 7.6: Case 2 - The EKF state estimate tracks the transition with a delay of 2 seconds when the observation frequency is 4 Hz, i.e. a new observation is available at every update step.

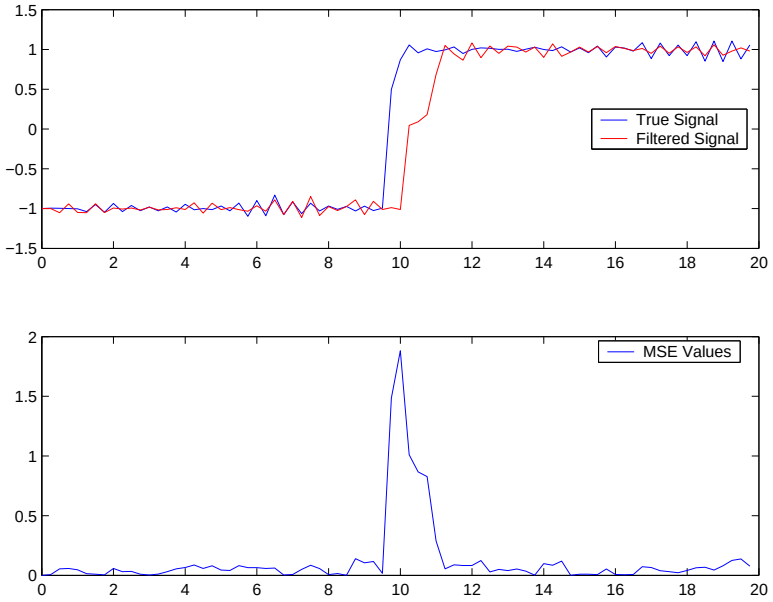


Figure 7.7: Case 2 - The EKF state estimate tracks the transition with a delay of 3 seconds when the observation frequency is 1 Hz.

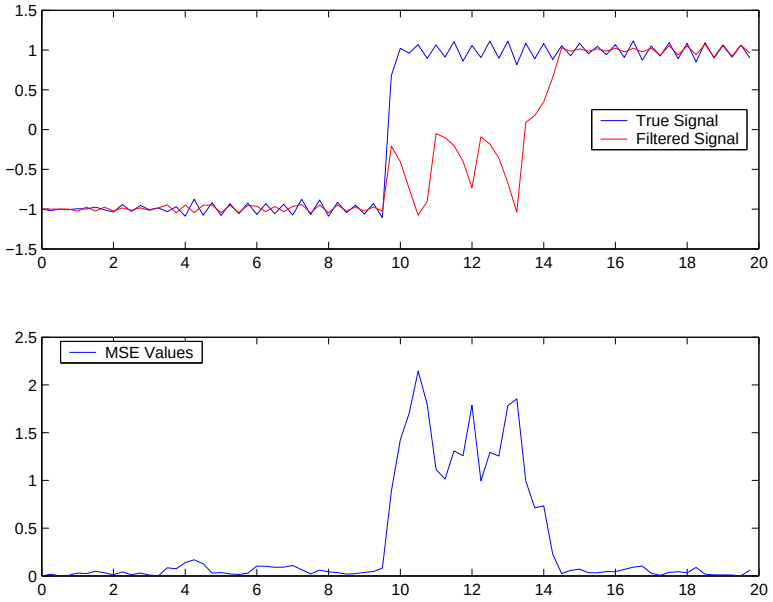


Figure 7.8: Case 2 - he EKF state estimate tracks the transition when the observation frequency is 0.50 Hz, but with a time delay of 5 seconds and marginally unstable behavior prior to the transition.

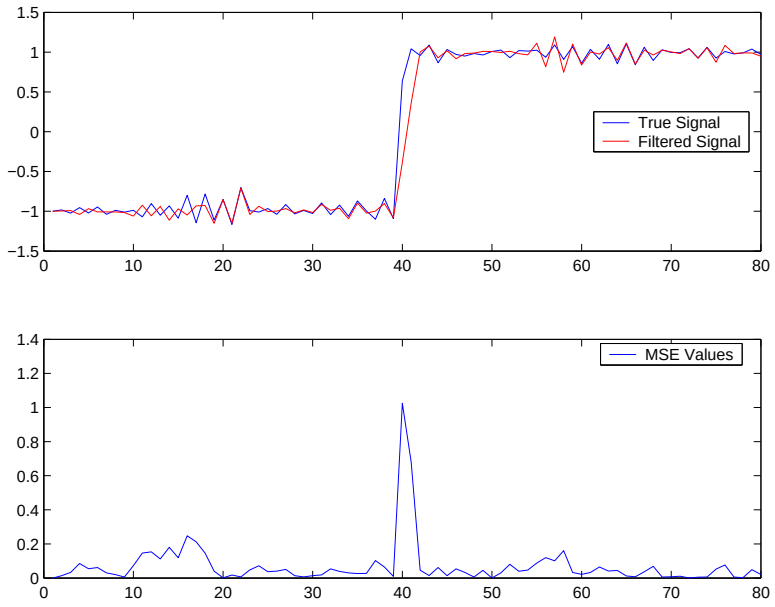


Figure 7.9: Case 3 - The EKF estimates the states and tracks the transition well when σ_z^2 is not much larger than σ_x^2, as shown with $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$.

Qualitatively, this type of behavior can be explained by an overconfidence in the predictions by the EKF in comparison to the observations [35, 92, 93]. In particular, it turns out that the filter error variance approaches a steady state value of 0.05, or 22% error about the mean. Thus, when the observation noise variance is equal to 0.08, or 28% error about the mean, the observations no longer affect the system and do not force the model into a different equilibrium. In the general case, the EKF will be subject to this type of result due to the way the Kalman filter gain matrix and estimates are computed in (7.41) and (7.42). In fact, if the observations are accurate enough to make this filter gain greater than 0.50, then the filter is expected to correctly track the system transition. As indicated by Miller [92], the EKF can be made to follow the state transitions by increasing the sampling frequency, such that the filter error variance does not reach its steady state value and processing subsequent observations nudges the state estimation to the accurate basin eventually (see Figure 7.8). However, if the observation noise is sufficiently larger than the system process noise, the filter gain may just not be large enough even over a long period, causing the filter to be too confident about the predictions and provide inaccurate state estimates indefinitely, as shown in Figure 7.11.

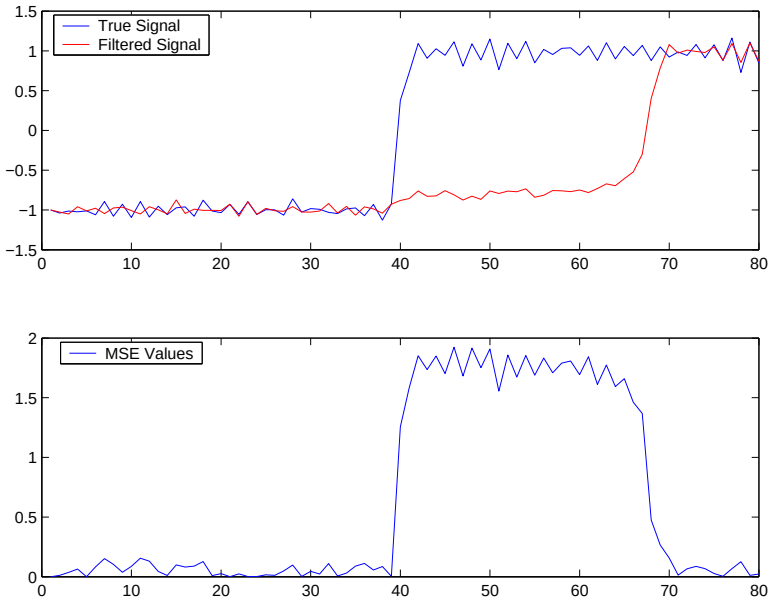


Figure 7.10: Case 3 - The EKF's state estimation and transition tracking deteriorates with a non-negligible delay of over 25 seconds (i.e. 100 samples at 4 Hz frequency) when σ_z^2 is sufficiently larger than σ_x^2, as shown with $\sigma_z^2 = 0.07$ and $\sigma_x^2 = 0.01$.

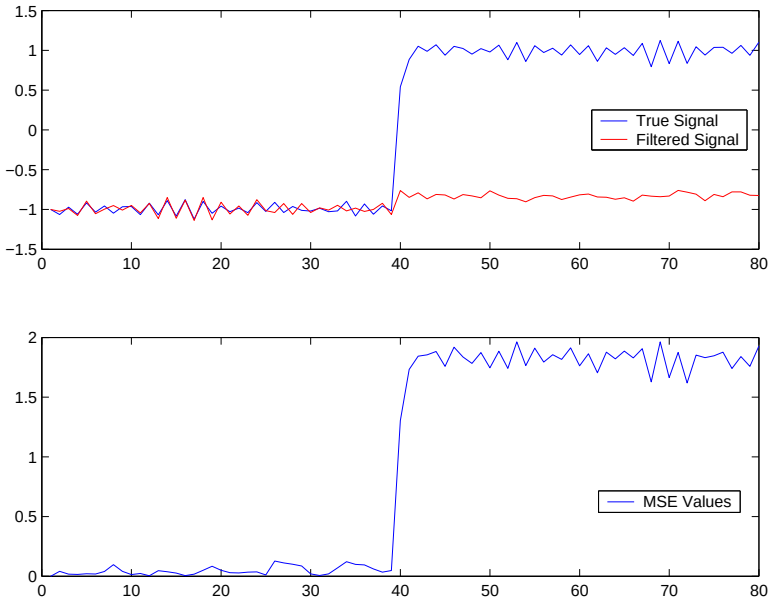


Figure 7.11: Case 3 - The EKF is unable to track the transition even after 40 seconds (i.e. 160 observation samples at 4 Hz frequency) when $\sigma_z^2 = 0.08$ and $\sigma_x^2 = 0.01$.

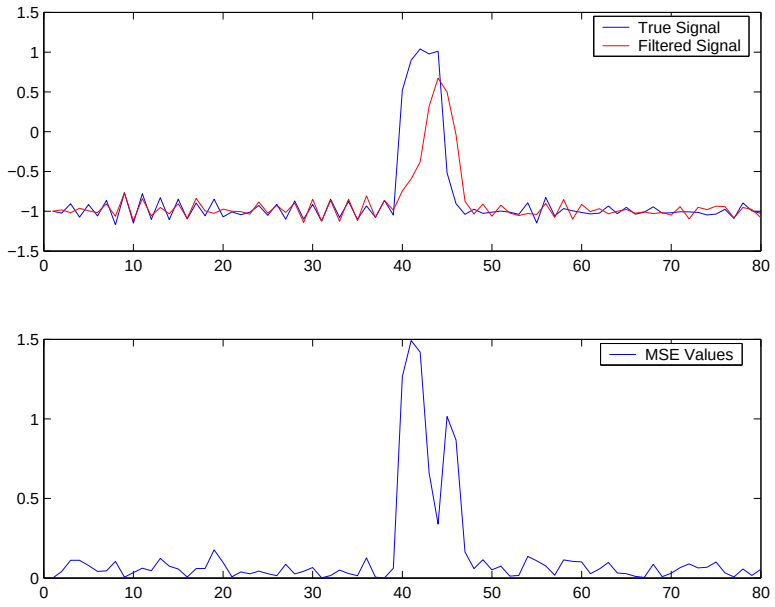


Figure 7.12: Case 4 - The EKF tracks a rapid transition if the observation noise variance is sufficiently low, i.e. $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, albeit with a noticeable delay

Next, we observe a set of scenarios in which the EKF tracks rapid transitions with an unduly time delay or is totally incapable of tracking rapid transitions when the observation noise variance is too large. In Figure 7.12, the system shifts from equilibrium around -1 to its equilibrium point around +1, and back again within 5 samples. With $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, the observations are not accurate enough to follow the transition immediately, but nevertheless, does track it with a finite delay. On the other hand, when $\sigma_z^2 = 0.05$, the EKF does not track the transition, as shown in Figure 7.13. In this case, the latter does not last long enough either to process a sequence of observations that may eventually force the filter to switch to another equilibrium point. Therefore, the EKF completely fails to follow the system state. Another example of such a scenario with a 10-sample long transition is given in Appendix A.

We now look at the opposite scenario, in which one or several observation samples are outliers. Figure 7.14 depicts an experiment in which sequential outliers are induced from $t = 45s$ to $t = 48s$. In this case, the EKF filter inaccurately outputs a state transition because it is unable to distinguish between good and bad observations.

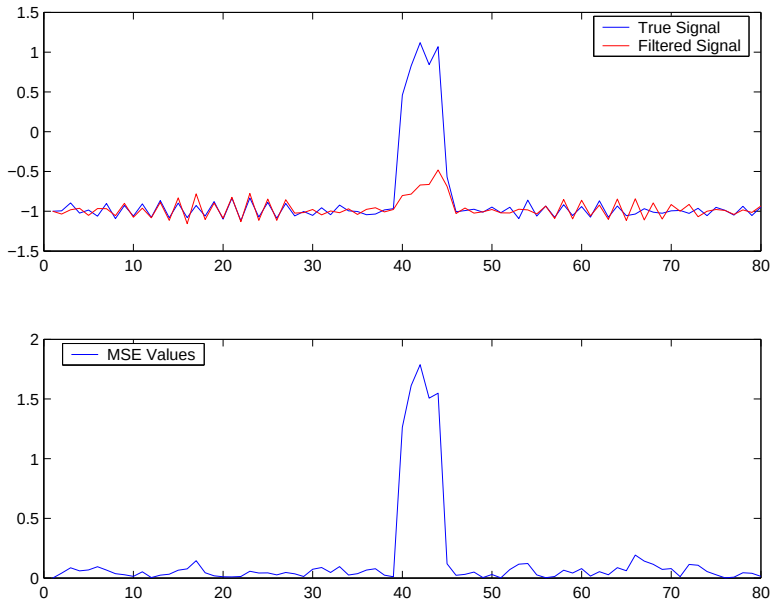


Figure 7.13: Case 4 - The EKF is unable to track a rapid transition if the observation noise variance exceeds a particular threshold, in this case $\sigma_z^2 = 0.05$ with $\sigma_x^2 = 0.01$.

Reliable estimation becomes even more difficult to get when two or more of the preceding situations occur concurrently. For example, the observation noise variance does not have to be much larger than that of the system process noise for the filter to fail, if the observation sampling period is increased simultaneously. This example is depicted by Figures 7.11 and 7.15. The system process noise variance is $\sigma_x^2 = 0.01$ in both cases. In Figure 7.11, it is shown that the EKF fails to follow the transition when $\sigma_z^2 = 0.08$ and the sampling period is 0.25 s. If the latter is increased to 1 s, the filter fails when σ_z^2 is just 0.02, as shown in Figure 7.15.

Clearly, the standard EKF provides poor estimates or completely fails in many different scenarios. What is needed to overcome such limitations of the standard EKF is a new robust filter with the following key advantages:

- Redundancy in the observations to overcome the sensitivity to the observation variance and overconfidence in the predictions
- Use of robust estimates of variance and weights along with a reliable iterative algorithm, making the filter resistant against outliers. In our application, we use the Projection Statistics

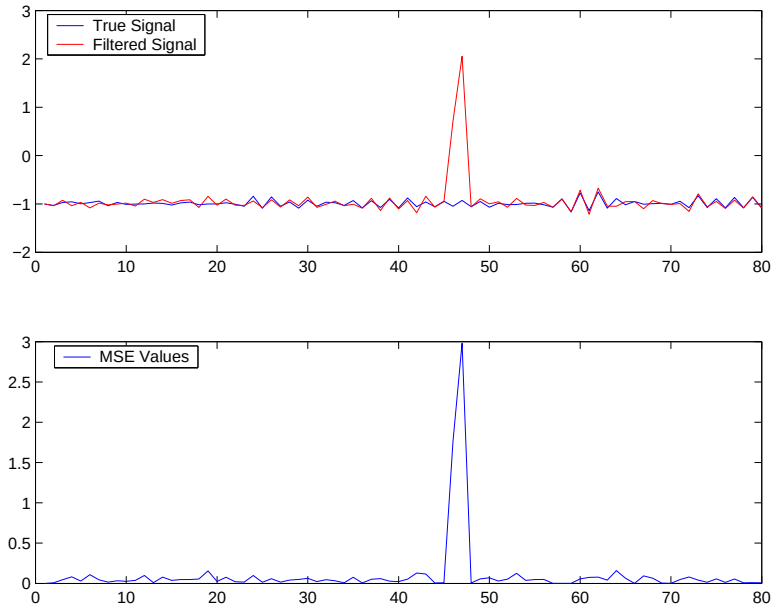


Figure 7.14: Case 5 - The EKF inaccurately outputs a state transition when outlying observations are processed.

(PS) and the Iteratively Re-weighted Least Squares (IRLS) algorithms to achieve our goals.

- Smooth ψ-function for a gradual down-weighting, helping maintain high statistical efficiency.

7.6 Development of the GM-EKF

The standard EKF filter leads to inaccurate estimates very easily in the presence of outliers, because just like the Kalman filter, the algorithm solves a least squares estimator with a breakdown point of zero. Furthermore, the standard formulation of the EKF degrades in many ways even when no outliers are present. For example, it was shown in the previous section that the EKF estimates are inconsistent with the true system behavior when a sudden state transition shift is not captured in the absence of accurate and precise observations.

To overcome these limitations, we now develop the robust GM-EKF method consisting of three key steps. In the first step, we convert the classical recursive approach into a batch-mode regression form so that the observations may be processed simultaneously. Recall that observation redundancy

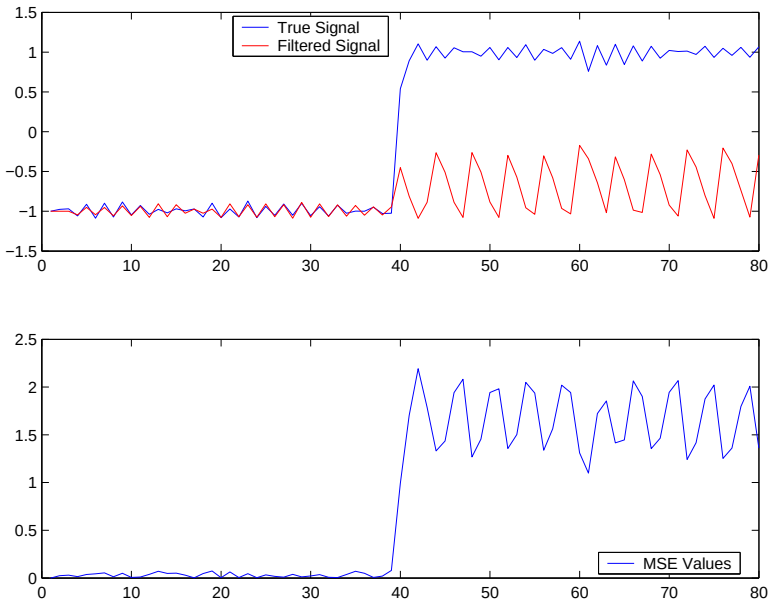


Figure 7.15: Case 6 - The EKF is unable to track the system transition even though $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$, due to the added complexity of a low sampling frequency of 1 Hz.

is required for an estimator to be capable of suppressing the outliers and tracking the transitions well, and can be achieved in practice by simply placing more sensors in the system. The second step consists of applying a prewhitening procedure that utilizes a robust estimator of location and covariance such as the Projection Statistics (PS) [42, 43, 27, 136] or minimum covariance determinant (MCD) [121]. The prewhitening procedure robustly uncorrelates the noise when outliers are present in the predictions and the observations. In the third step, the unconstrained nonlinear optimization in the GM-estimator is solved using the Iteratively Re-weighted Least Squares (IRLS) algorithm. Again, the influence function (IF) of the GM-estimator is employed to derive the asymptotic state estimation error covariance matrix of the GM-EKF [55, 37]. Recall that while the IF of an estimator is a measure of its sensitivity to infinitesimal contamination at a given distribution, its covariance matrix is equal to that of the estimator at that distribution ([55, 37]). In this chapter, we derive the IF of the GM-estimator for nonlinear regression models [98, 145].

We now develop the GM-EKF in detail. Let the system equations be given as

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t) + \mathbf{w}_t + \mathbf{u}_t, \tag{7.53}$$

$$\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t) + \mathbf{e}_t. \tag{7.54}$$

This model is discretized and linearized as described in Section 7.3, yielding

$$\mathbf{x}_k = \mathbf{F}_d\mathbf{x}_{k-1} + \mathbf{w}_k + \mathbf{B}_d\mathbf{u}_k \tag{7.55}$$

$$\mathbf{z}_k = \mathbf{H}_d\mathbf{x}_k + \mathbf{e}_k, \tag{7.56}$$

with

$$\mathbf{F}_d = e^{\mathbf{F}_x T_s} \tag{7.57}$$

$$\mathbf{H}_d = \mathbf{H}_x, \tag{7.58}$$

and

$$\mathbf{F}_x = \frac{\partial \mathbf{f}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\hat{\mathbf{x}}_{k-1|k-1}}, \tag{7.59}$$

$$\mathbf{H}_x = \frac{\partial \mathbf{h}(\mathbf{x}_t)}{\partial \mathbf{x}_t}\Big|_{\mathbf{x}_t=\hat{\mathbf{x}}_{k-1|k-1}}. \tag{7.60}$$

We first predict the next state using the following equation:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_d\hat{\mathbf{x}}_{k-1|k-1} + \int_{t_{k-1}}^{t_k} \mathbf{f}(\hat{\mathbf{x}}_{k-1|k-1})\, dt + \mathbf{B}_d\mathbf{u}_k. \tag{7.61}$$

Then, we combine these predictions with the observations to obtain the batch regression form as follows:

$$\begin{bmatrix} \mathbf{z}_k \\ \hat{\mathbf{x}}_{k|k-1} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_d \\ \mathbf{I} \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} \mathbf{e}_k \\ \boldsymbol{\delta}_{k|k-1} \end{bmatrix}, \tag{7.62}$$

where $\boldsymbol{\delta}_{k|k-1}$ is the error between the true state and its prediction, yielding

$$\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}\mathbf{x}_k + \tilde{\mathbf{e}}_k. \tag{7.63}$$

The second step of the filter is to uncorrelate the data. However, we may not perform data pre-whitening using classical methods since these techniques may negatively affect any outlying data. Instead, we use again a robust data pre-whitening procedure in the GM-EKF. In this procedure, we must first identify the outliers using a robust outlier detection method. We use the PS algorithm again for this purpose. Then, we compute the weights $\bar{\omega}_i$ for the elements of the vector $\tilde{\mathbf{z}}_k$ using the PS values, as follows:

$$\bar{\omega}_i = min\left(1, \frac{d^2}{PS_i^2}\right). \tag{7.64}$$

Next, the outliers are down-weighted by applying $\bar{\omega}_i$ to the elements of $\tilde{\mathbf{z}}_k$. After down-weighting the data vector, compute the covariance matrix $\tilde{\mathbf{R}}_k$ of the error $\tilde{\mathbf{e}}_k$, given by

$$\tilde{\mathbf{R}}_k = \begin{bmatrix} \mathbf{R}_k & 0 \\ 0 & \boldsymbol{\Sigma}_{k|k-1} \end{bmatrix}, \tag{7.65}$$

with $\mathbf{R}_k$ being the noise covariance matrix of $\mathbf{e}_k$ and $\boldsymbol{\Sigma}_{k|k-1}$ the propagated filter error covariance matrix after prediction. Recall that $\boldsymbol{\Sigma}_{k|k-1}$ is given by

$$\boldsymbol{\Sigma}_{k|k-1} = \mathbf{F}_d \boldsymbol{\Sigma}_{k-1|k-1} \mathbf{F}_d^T + \mathbf{W}_k. \tag{7.66}$$

We continue the robust pre-whitening procedure as follows. Using either upper diagonal factorization or Cholesky decomposition, we obtain the matrix $\mathbf{S}_k$ such that $\tilde{\mathbf{R}}_k = \mathbf{S}_k \mathbf{S}_k^T$. Equivalently, we may use the square-root method to obtain $\sqrt{\tilde{\mathbf{R}}_k}$ such that $\tilde{\mathbf{R}}_k = \sqrt{\tilde{\mathbf{R}}_k}\sqrt{\tilde{\mathbf{R}}_k}$. Finally, we multiply the linear regression model $\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}\mathbf{x}_k + \tilde{\mathbf{e}}_k$ on the left-hand side by $(\mathbf{S}_k)^{-1}$ or $\left(\sqrt{\tilde{\mathbf{R}}_k}\right)^{-1}$ to

perform pre-whitening; for example,

$$(\tilde{\mathbf{R}}_k)^{-\frac{1}{2}}\tilde{\mathbf{z}}_k = (\tilde{\mathbf{R}}_k)^{-\frac{1}{2}}\tilde{\mathbf{H}}\mathbf{x}_k + (\tilde{\mathbf{R}}_k)^{-\frac{1}{2}}\tilde{\mathbf{e}}_k, \tag{7.67}$$

yielding the final form of the regression as

$$\mathbf{y}_k = \mathbf{A}_k\mathbf{x}_k + \boldsymbol{\eta}_k. \tag{7.68}$$

The third and last step of the GM-EKF is to solve for the state estimates iteratively using the GM-estimator with the IRLS algorithm, with the solution given by

$$\hat{\mathbf{x}}_{k|k}^{(\nu+1)} = \left(\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{Q}^{(\nu)}\mathbf{y}_k. \tag{7.69}$$

The weight matrix $\mathbf{Q}$ in (7.69) is given by

$$\mathbf{Q} = diag\left\{q\left(\frac{r_i}{s\bar{\omega}_i}\right)\right\}, \tag{7.70}$$

with

$$q\left(\frac{r_i}{s\bar{\omega}_i}\right) = \frac{\psi\left(\frac{r_i}{s\bar{\omega}_i}\right)}{\left(\frac{r_i}{s\bar{\omega}_i}\right)} \tag{7.71}$$

and

$$\bar{\omega}_i = min\left(1, \frac{d^2}{PS_i^2}\right). \tag{7.72}$$

7.7 Influence Functions for GM-Estimators of Nonlinear Models

Linearization of the nonlinear model around the nominal values of the estimate enabled a simple formulation of the GM-EKF. The filter error covariance matrix $\boldsymbol{\Sigma}_{k-1|k-1}$ remains a key component of this filter, as it is required to compute $\tilde{\mathbf{R}}_k$ in the pre-whitening procedure and also influences the state estimation via (7.69). For the standard EKF, the covariance matrix may be computed

using the known covariances $\mathbf{R}_k$ and $\mathbf{W}_k$ following (7.43). For the robust GM-EKF though, it needs to be redeveloped using the relationship between the GM-estimator's influence function and its asymptotic covariance matrix. However, instead of using the influence function of the GM-estimator for linear models, one needs to derive the IF corresponding to the particular nonlinear model under consideration. Next, we derive the general form of this influence function. Let the nonlinear regression model be expressed as

$$\mathbf{y} = \boldsymbol{\varphi}(\mathbf{a}, \mathbf{x}) + \boldsymbol{\eta}, \tag{7.73}$$

where $\mathbf{y}$ are the observations, $\mathbf{a}$ are the explanatory variables, $\mathbf{x}$ is the parameter vector, $\boldsymbol{\eta}$ is the observation noise, and $\boldsymbol{\varphi}$ is a vector-valued nonlinear regression function with respect to $\mathbf{a}$ and/or $\mathbf{x}$. Recall that $\mathbf{y}$ and $\boldsymbol{\eta}$ in this model are i.i.d. vectors following Gaussian distributions. Assuming the function $\boldsymbol{\varphi}$ is deterministic and independent of the residuals, the cumulative probability distribution function of the residual error vector $\mathbf{r}$, expressed as

$$\mathbf{r} = \mathbf{y} - \boldsymbol{\varphi}(\mathbf{a}, \hat{\mathbf{x}}), \tag{7.74}$$

is denoted by $\Phi(\mathbf{r})$. For this model, the GM-estimator in regression provides an estimate for $\mathbf{x}$ by processing the redundant observation vector $\mathbf{y}$ and solving the implicit equation given by

$$\sum_{i=1}^{m} \boldsymbol{\lambda}_i \left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) = \mathbf{0}, \tag{7.75}$$

where

$$\boldsymbol{\lambda}_i \left(\mathbf{r}, \mathbf{a}_i, \mathbf{x}\right) = \bar{\omega}_i \frac{\partial \boldsymbol{\varphi}_i(\mathbf{a}, \mathbf{x})}{\partial \mathbf{x}} \; \psi\left(\frac{r_i}{s\bar{\omega}_i}\right). \tag{7.76}$$

Given the empirical cumulative probability distribution function F_m, the functional form of the estimator, where $\mathbf{x}$ is replaced by $\mathbf{T}$, is given by the vector-valued functional

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}\right) dF_m = \mathbf{0}. \tag{7.77}$$

Asymptotically, $F_m \rightarrow G$ by virtue of the Glivenko-Cantelli Theorem [115] and (7.77) becomes

$$\int \boldsymbol{\lambda}\left(\mathbf{r}, \mathbf{a}, \mathbf{T}(G)\right) dG = \mathbf{0}. \tag{7.78}$$

Following [98, 145], we begin the derivation of the asymptotic influence function, given by

$$\mathbf{IF}(\mathbf{r}, \mathbf{a}; \Phi) = \frac{\partial \mathbf{T}(G)}{\partial \epsilon}\Big|_{\epsilon=0} = \lim_{\epsilon \downarrow 0} \frac{\mathbf{T}((1-\epsilon)\Phi + \epsilon H) - \mathbf{T}(\Phi)}{\epsilon}, \tag{7.79}$$

by substituting $G = (1-\epsilon)\Phi + \epsilon H$ into (7.78), yielding

$$\begin{aligned}
\int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ dG &= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d[(1-\epsilon)\Phi + \epsilon H] && (7.80)\\
&= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d[\Phi + \epsilon(H - \Phi)] && (7.81)\\
&= \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \epsilon \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) = \mathbf{0}. && (7.82)
\end{aligned}$$

Differentiating with respect to ϵ and applying the chain rule yields

$$\frac{\partial}{\partial \epsilon} \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) + \epsilon \frac{\partial}{\partial \epsilon} \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(H - \Phi) = \mathbf{0}. \tag{7.83}$$

Next, recall the following interchangeability of differentiation and integration:

$$\frac{d}{d\epsilon} \int_S f(\epsilon, x)\mu(dx) = \int_S f'(\epsilon, x)\mu(dx), \tag{7.84}$$

assuming that f is continuous and measurable and f' is measurable on S [137, 141]. Using this result and given $H = \Delta_r$, where Δ_r is the probability mass at r, (7.7) reduces to

$$\int \frac{\partial}{\partial \epsilon} \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(\Delta_r - \Phi) + \epsilon \int \frac{\partial}{\partial \epsilon} \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d(\Delta_r - \Phi) = \mathbf{0}. \tag{7.85}$$

Evaluating this expression at $\epsilon = 0$ makes the last term equal to zero. It follows that

$$\int \frac{\partial}{\partial \epsilon} \left[\boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\right]_{\epsilon=0}\ d\Phi + \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Delta_r = \int \boldsymbol{\lambda}(\mathbf{r}, \mathbf{a}, \mathbf{T}(G))\ d\Phi. \tag{7.86}$$

Assuming Fisher consistency at Φ, the right-hand side of (7.86) becomes zero. Then, using the sifting property of Δ_r, we obtain

$$\int \frac{\partial \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} \frac{\partial \mathbf{T}}{\partial \epsilon}\Big|_{\epsilon=0} d\Phi + \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)) = \mathbf{0}. \tag{7.87}$$

This results in the following expression for the influence function:

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = \frac{\partial \mathbf{T}}{\partial \epsilon}\Big|_{\epsilon=0} = -\left[\int \frac{\partial \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]}{\partial \mathbf{x}}\Big|_{\mathbf{T}(\Phi)} d\Phi\right]^{-1} \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)). \tag{7.88}$$

Substituting (7.76) into (7.88) yields

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = -\left[\int \frac{\partial}{\partial \mathbf{x}} \left[\boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})\right]\Big|_{\mathbf{T}(\Phi)} d\Phi\right]^{-1} \boldsymbol{\lambda}(\mathbf{r},\mathbf{a},\mathbf{T}(\Phi)). \tag{7.89}$$

Deriving $\lambda(\cdot)$ with respect to $\mathbf{x}$, and assuming that $\bar{\omega}$ and s are independent of $\mathbf{x}$ [145], we obtain

$$\frac{\partial \boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})}{\partial \mathbf{x}} = \bar{\omega}\left[\frac{\partial \psi(r_{\bar{\omega}})}{\partial \mathbf{x}}\right]\left[\frac{\partial \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial \mathbf{x}}\right]^T + \bar{\omega}\psi(r_{\bar{\omega}})\left[\frac{\partial^2 \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial x_i \partial x_j}\right], \tag{7.90}$$

where $r_{\bar{\omega}} = (r/s\bar{\omega})$ and $[\partial^2 \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})/\partial x_i \partial x_j]$ is the Hessian matrix of $\boldsymbol{\varphi}(\mathbf{y},\ \mathbf{x})$. Applying the chain rule to the derivative of $\psi(\cdot)$ yields

$$\frac{\partial \boldsymbol{\lambda}(\mathbf{y},\mathbf{a},\mathbf{x})}{\partial \mathbf{x}} = -\psi'(r_{\bar{\omega}})\left[\frac{\partial \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial \mathbf{x}}\right]\left[\frac{\partial \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial \mathbf{x}}\right]^T + \bar{\omega}\psi(r_{\bar{\omega}})\left[\frac{\partial^2 \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial x_i \partial x_j}\right]. \tag{7.91}$$

Finally, substituting (7.76) and (7.91) into (7.89) yields the following as the influence function of the GM-estimator for nonlinear regression:

$$\mathbf{IF}(\mathbf{r},\mathbf{a};\Phi) = \left(\int \left\{\psi'(r_{\bar{\omega}})\frac{\partial \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial \mathbf{x}}\frac{\partial \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial \mathbf{x}}^T + \bar{\omega}\psi(r_{\bar{\omega}})\frac{\partial^2 \boldsymbol{\varphi}(\mathbf{y},\mathbf{x})}{\partial x_i \partial x_j}\right\}\Big|_{\mathbf{T}(\Phi)}\right)^{-1} \bar{\omega}\frac{\partial \boldsymbol{\varphi}(\mathbf{a},\mathbf{x})}{\partial \mathbf{x}}\psi(r_{\bar{\omega}}). \tag{7.92}$$

7.8 Breakdown Point in Nonlinear Regression

We discuss next the concept of breakdown point in nonlinear regression. Recall that the finite sample breakdown point in the linear regression case is defined as

$$\epsilon^* = max\left\{\epsilon = \frac{f}{m};\ b_{max} \text{ finite}\right\}, \tag{7.93}$$

where f is the number of contaminant points, m the total number points in the sample $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$, and b_{max} the maximum bias. As indicated by Stromberg and Ruppert [140], this breakdown point definition cannot be extended to nonlinear models due to two reasons: (a) if the parameter space is bounded, then $\epsilon^* = 1$, and (b) the breakdown point is not invariant to nonlinear reparameterization. For example, if $y_i = sin(hx_i)+e_i$ with the parameter x being an angle bounded to $(-\pi/2, \pi/2)$, it cannot take on arbitrarily large values by definition leading to a finite sample breakdown point of 1 for the least squares estimator [140].

Instead, Stromberg and Ruppert [140] have suggested that the breakdown point be defined in terms of the regression function instead of the estimated parameter. In particular, is is defined as the minimum amount of contamination that carries the regression function towards its supremum or infimum, respectively known as upper breakdown or lower breakdown. Formally, these are defined as

$$\begin{aligned}
\epsilon_+ &= min\left\{\frac{f}{m}; sup\ h(\hat{x}) = sup_\theta\ h(x)\right\} \text{ if } sup_\theta > h(x) &(7.94)\\
&= 1 \text{ otherwise} \\
\epsilon_- &= min\left\{\frac{f}{m}; inf\ h(\hat{x}) = inf_\theta\ h(x)\right\} \text{ if } inf_\theta < h(x) &(7.95)\\
&= 1 \text{ otherwise} \\
\epsilon &= min\{\epsilon_+, \epsilon_-\}. &(7.96)
\end{aligned}$$

It turns out that the breakdown point of the least squares estimator using this new definition is generally the same as in linear regression, i.e. $1/n$. In contrast, the breakdown point of very robust

estimators such as the Least Mean Squares (LMS) and Least Trimmed Squares (LTS) depends on the regression function itself along with the sample. For more details about the case of unbounded nonlinear regression functions, one can refer to [140]. It is interesting to note that the breakdown point of estimators for bounded regression functions depends on the sample data's behavior near the boundaries of the regression function. Finally, it should be noted that in nonlinear regression, the concept of leverage point must be redefined. An approach using nonlinear manifolds and techniques from differential geometry is one possible route to perform a more complete study for these estimators, as briefly discussed in [23].

7.9 Tracking Climate Transitions Using GM-EKF

In this section, we revisit the simulation scenarios stated in Table 7.1 and compare the GM-EKF's results with those obtained from the standard EKF. We begin with Figure 7.16, which shows the GM-EKF state estimation performance at the Gaussian distribution with no transitions and outliers. The high statistical efficiency expected from the GM-EKF is demonstrated by the relatively accurate estimates of the states in this experiment. The following values are assumed in this simulation: $\sigma_z^2 = 0.01$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$.

The ability and advantages of the GM-EKF method begin to emerge with case 2. In particular, we now simulate the system with transitions and investigate how the EKF is able to track the system's shift in equilibrium points. In the subsequent results, the following values are once again applied in the model: $\sigma_z^2 = 0.01$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$, unless otherwise noted. Figure 7.17 depicts the result of a simulation trial in which a new observation is available every time a prediction is made. The sampling period in this case is 0.25 seconds, and the corresponding sampling frequency is 4 Hz. Clearly, the GM-EKF tracks the system transition as soon as the observation is available. Now, we reduce the frequency of observations to 1 Hz and 0.5 Hz. The EKF results are seen to deteriorate in these cases, whereas the good performance of GM-EKF can be observed in Figures 7.18 and 7.19, respectively. Just one redundant observation available at each time step enables

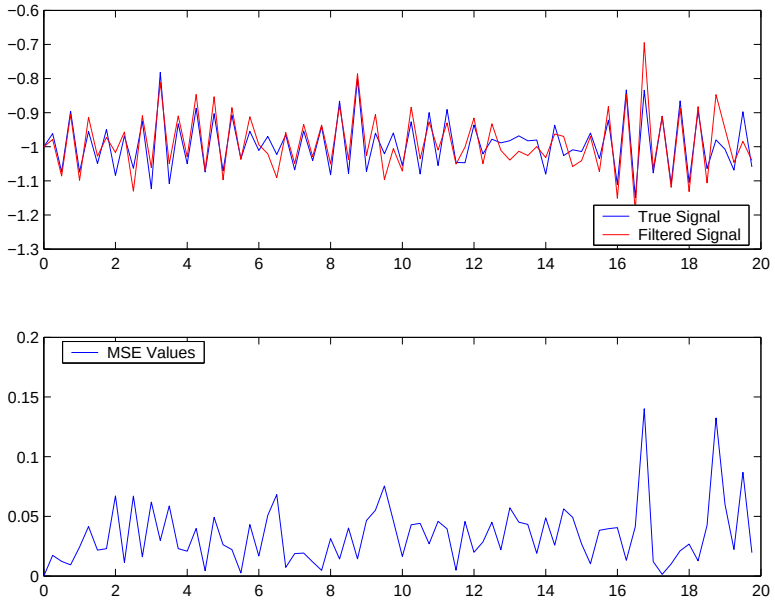


Figure 7.16: Case 1 - The GM-EKF performs satisfactory state estimation in a double-well system that is residing around one of its stable equilibrium points. Lower frame depicts the associated MSE values for the estimation segment shown in the upper frame.

the filter to sense the system transition as soon as the observations are recorded. Evidently, this characteristic plays an important role in this performance improvement.

Next, we evaluate the performance of the GM-EKF in situations where the observation noise variances are greater than the system process noise and filter error variances. First, let $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and $\kappa = 0.75$. As seen in Figure 7.20, the GM-EKF is able to track the transition in this case. Now, let $\sigma_z^2 = 0.07$ and $\sigma_z^2 = 0.08$. Figures 7.21 and 7.22 demonstrate the GM-EKF can easily detect and track the transition well for both cases. In contrast, the EKF is unable to follow the transition for over 160 samples when $\sigma_z^2 = 0.08$, as seen in Figure 7.11.

Qualitatively, the GM-EKF is able to reduce the confidence in the predictions in comparison with the observations by leveraging the observation redundancy and the GM-estimator in state estimation. The latter relies on robust estimators of scale and covariance to down-weight any non-conforming data, including the predictions, when compared to the majority observations. In other words, though no single observation is accurate enough to make the filter gain greater than 0.50, the GM-EKF computes a gain large enough to output the correct state estimate and correctly

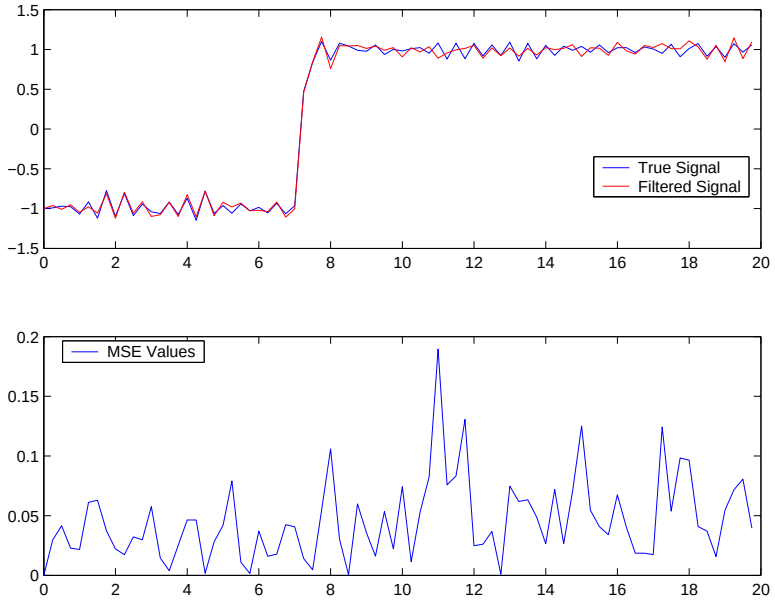


Figure 7.17: Case 2 - The GM-EKF state estimate tracks the transition with no delays when the observation frequency is 4 Hz, i.e. a new observation is available at every update step.

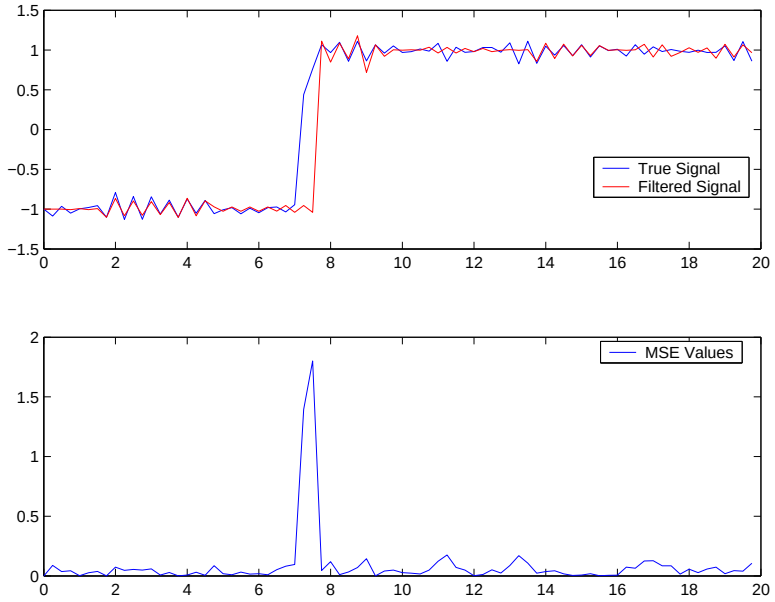


Figure 7.18: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 1 Hz.

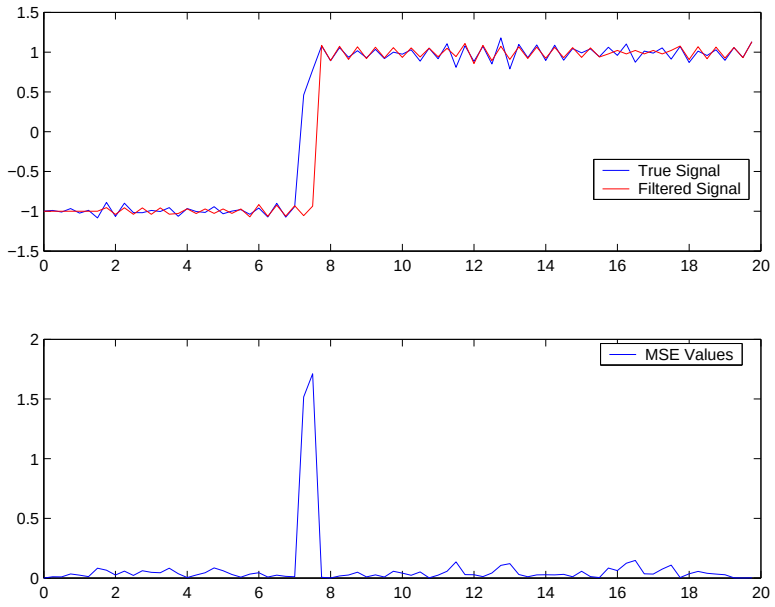


Figure 7.19: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 0.50 Hz.

senses the transition.

We now consider scenarios in which the transitions between the basins occur very rapidly. In Figure 7.23, the system shifts its dynamics from the basin of attraction at the equilibrium point -1 to that at equilibrium point +1, and back again within 5 samples. With $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, the GM-EKF once again is able to track the system's transition very rapidly and accurately. Even if the noise variance is increased to $\sigma_z^2 = 0.05$, the GM-EKF performs well as seen in Figure 7.24. This is in contrast to the EKF's poor performance as depicted in Figure 7.13. Another example of such a scenario with a 10-sample long transition is given in Appendix B.

We now look at the opposite scenario, in which one or more observation samples are outliers. Figure 7.25 depicts an experiment in which sequential outliers are induced from $t = 45s$ to $t = 48s$. In this case, the GM-EKF filter suppresses the outlying observation at each time step. Thus, a single redundant observation enables the filter to suppress up to 1 outlier in the system, whether it is in the predictions or observations.

We now evaluate the GM-EKF's breakdown point via extensive simulations with the system

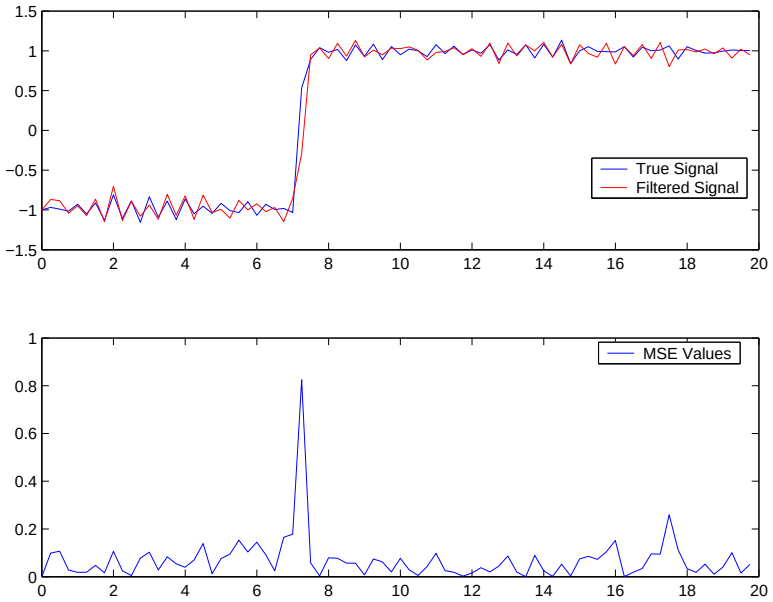


Figure 7.20: Case 3 - The GM-EKF estimates the states and tracks the transition well when σ_z^2 is not much larger than σ_x^2, as shown with $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$.

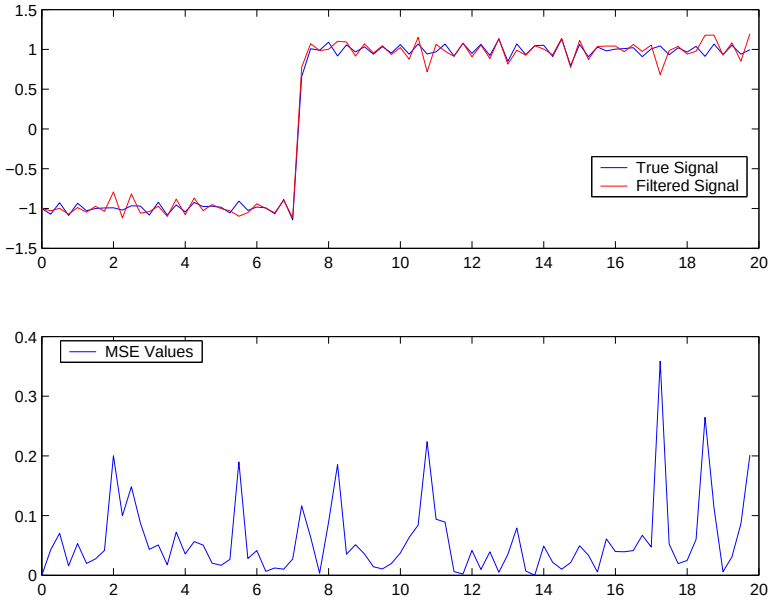


Figure 7.21: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown with $\sigma_z^2 = 0.07$ and $\sigma_x^2 = 0.01$.

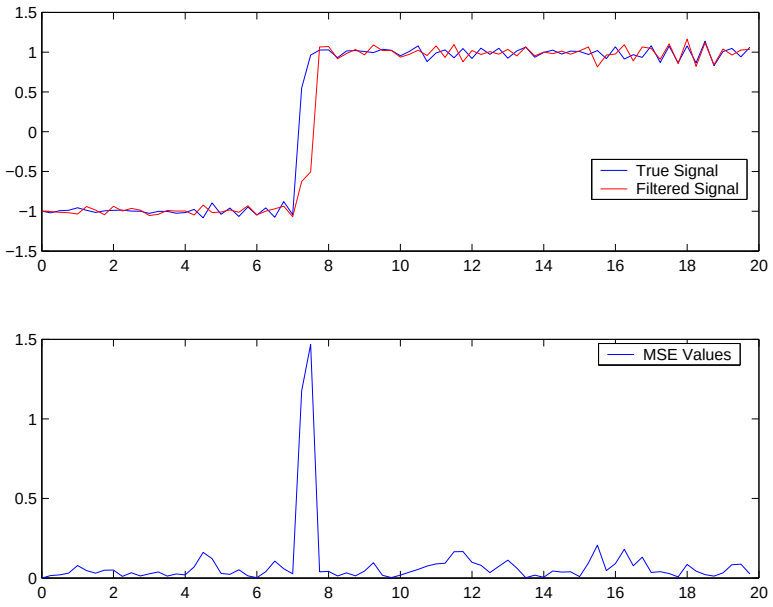


Figure 7.22: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown with $\sigma_z^2 = 0.08$ and $\sigma_x^2 = 0.01$ (a case in which the standard EKF could not track the transition even after 40 seconds of observation data was processed).

containing several redundant observations and a varying number of outliers. Because the GM-EKF uses a linearized dynamic model to perform state estimation, we compute its breakdown point using the definition given by Rousseeuw and Leroy [121] and Mili and Coakley [91]. Recall from Chapter 3 that the maximum possible finite-sample breakdown point is given by $\epsilon^*_{max} = [(m-n)/2]/m$, where m is the total number of observations with n variables. Table 7.2 contains some results from these simulations showing how many outliers the filter is able to suppress versus the maximum possible breakdown point given by ϵ^*_{max}. We observe that the finite-sample breakdown point of the GM-EKF is no larger than 25%.

Clearly, the GM-EKF results demonstrate significantly better performance in many different scenarios. The new filter is also able to handle more complicated scenarios, such as a larger than desired observation noise variance along with low availability of the observations (low frequency). Figure 7.26 depicts this example, which has a sampling period of 1 second; the filter tracks the system's shift in equilibrium well when $\sigma_z^2 = 0.02$, a situation in which the standard EKF fails.

Table 7.2: GM-EKF Breakdown Point in the Presence of Outliers

Total # of Observations $(m+n)$	# of Redundant Observations	Maximum # of Outliers that can be Suppressed	Maximum Possible Breakdown Point	Actual # of Outliers Suppressed	Breakdown Point in Experiment
3	2	1	33%	1	33%
4	3	1	25%	1	25%
5	4	2	40%	2	40%
6	5	2	33%	2	33%
7	6	3	43%	2	28.5%
8	7	3	37.5%	3	37.5%

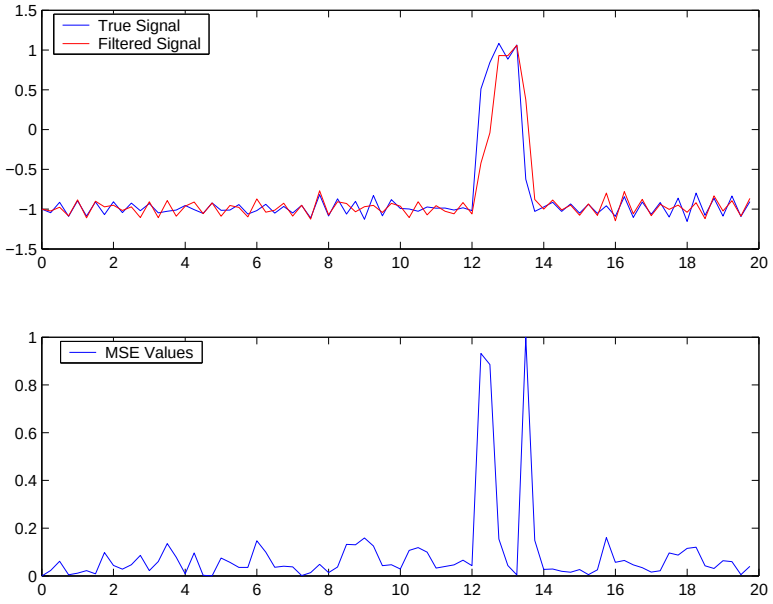


Figure 7.23: Case 4 - The GM-EKF tracks a rapid transition if the observation noise variance is sufficiently low, i.e. $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, with almost no delay.

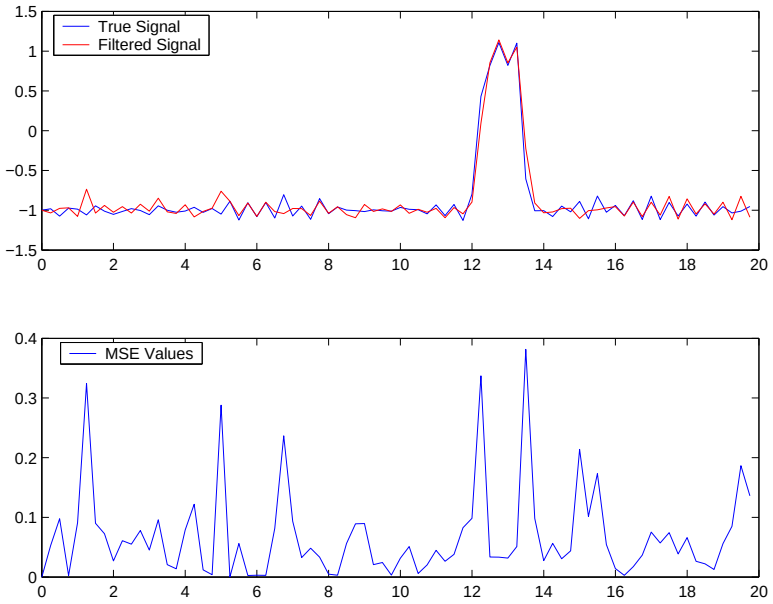


Figure 7.24: Case 4 - The GM-EKF tracks a rapid transition even if the observation noise variance is $\sigma_z^2 = 0.05$ and $\sigma_x^2 = 0.01$, a case in which the standard EKF was unable to track the transition.

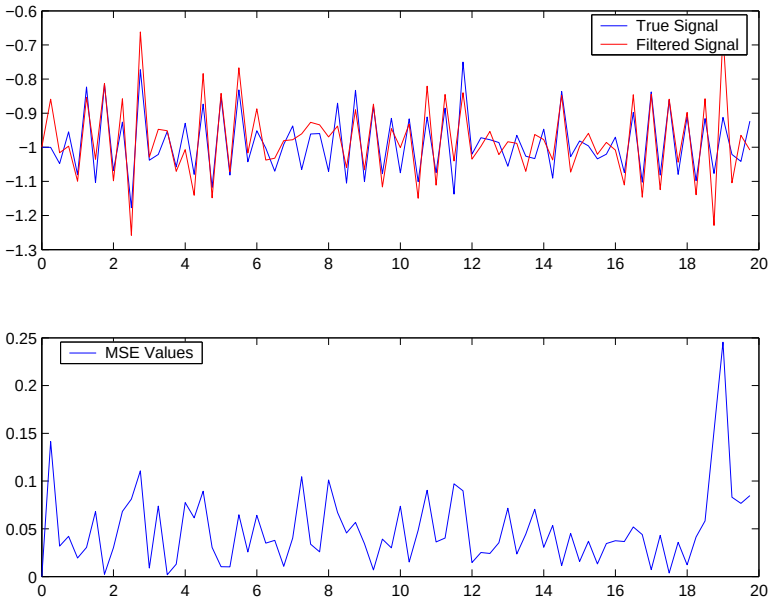


Figure 7.25: Case 5 - The GM-EKF accurately recognizes that one of the two observations being processed is an outlier, and does not output a state transition.

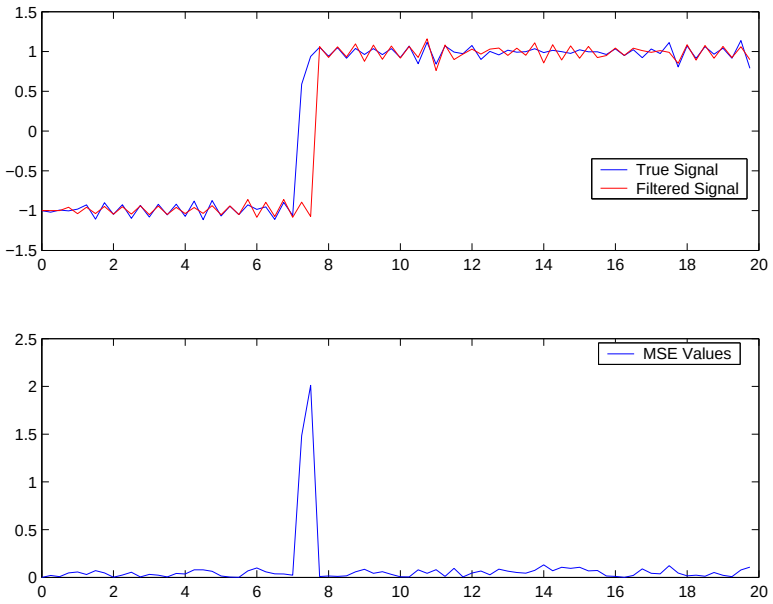


Figure 7.26: Case 6 - The GM-EKF tracks the system transition with multiple effects $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and sampling frequency is 1 Hz.

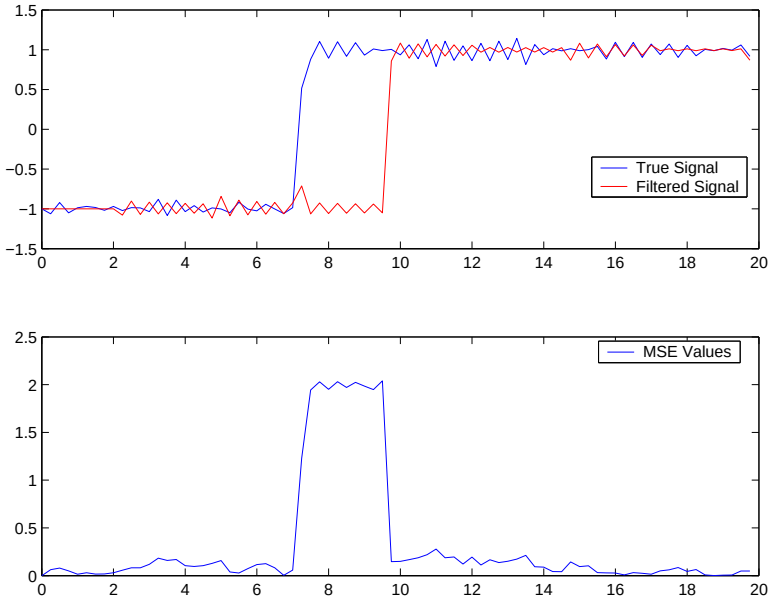


Figure 7.27: Case 6 - The GM-EKF shows signs of deterioration only when $\sigma_z^2 = 0.10$, $\sigma_x^2 = 0.01$, and sampling frequency is 0.4 Hz, which indicates that much more extreme cases of observation sampling frequency and noise variance can be handled by the new filter in comparison to the standard EKF.

Indeed, as seen in Figure 7.27, the observation noise variance has to be an order of magnitude larger than the system process noise variance, i.e. $\sigma_z^2 = 0.10$ and $\sigma_x^2 = 0.01$ with a sampling frequency of 0.4 Hz, for the GM-EKF to be slightly affected. Regardless, it still senses the transition and tracks it accordingly.

Chapter 8

Summary and Discussions

In this research, we have developed a new class of filters that are able to handle three types of outliers - observation, innovation, and structural - which may affect the following components of a system's model: observation noise, system process noise, control input, state transition matrix, and observation matrix. In particular, the generalized maximum likelihood-type Kalman filter is developed as one good method in this general class of filters. Robustness is achieved in this filter framework via several aspects, including redundancy in the observation vector, a new data prewhitening procedure, and use of a robust estimator to solve for the state estimates.

The need for redundancy in this framework is motivated by the concept of breakdown point borrowed from robust statistical theory. To achieve redundancy, the filter is cast into a general linear regression framework that combines the predictions and the observations for the state estimation. The new data prewhitening procedure incorporates robust estimators of location and variance to accurately detect and handle outliers while decorrelating the data. Finally, the linear regression framework allows us to make use of any robust estimator whose covariance matrix can be derived. In this work, the class of nonlinear GM-estimators and the IRLS algorithm are used to solve for the state estimates. Characterizing the three types of outliers as vertical outliers and bad leverage points enables us to develop an appropriate procedure that is able to suppress them. For example, structural outliers behave as bad leverage points in the linear regression space; hence, the GM-

estimator is employed given its robustness to the latter. Finally, a new error covariance matrix for the GM-KF is developed using the influence function of the GM-estimator. In addition, systems undergoing nonlinear dynamics with one or several stable equilibrium points are considered in this work. The GM-EKF is developed as a robust version of the traditional extended Kalman filter methodology, and is shown to significantly outperform the latter in climate transition tracking applications.

The GM-KF's efficiency and robustness to concomitant outliers are investigated through various simulations. From an efficiency viewpoint, it is shown that multiple observations are advantageous in the estimation as the one with the lowest noise variance drives the MSE value. The efficiency for the GM-KF relative to the classical KF turns out to be 85% in our simulations. From a robustness perspective, the GM-KF is shown to suppress all three types of outliers in various scenarios with a breakdown point of up to 25%. The GM-KF also proves to be more robust to outlier contamination than the H_∞-filter.

Inherently, this research opens the door to a methodology that can be used to develop a variety of other nonlinear filters. The ρ-function employed in the GM-KF and GM-EKF can be modified to obtain different convergence rate, robustness, and efficiency properties in the filter. For example, replacing the Huber ρ-function used in this work with the strictly convex logistic function allows one to iterate by means of the Newton method instead of the IRLS algorithm. The filter convergence rate would then be quadratic instead of linear. Besides variations to the GM-KF and GM-EKF, completely new filters can also be developed by replacing the GM-estimator with another regression estimator of choice, such as the MM-estimator [78], so long as that estimator's covariance matrix can be computed.

As seen in this dissertation, the proposed GM-KF and the H_∞-filter are very complementary to each other. The former withstands outliers while the latter handles noise uncertainties or system modeling errors. Several researchers have proposed to combine the Kalman filter with the H_∞-filter to find the best state estimator in the Kalman filter sense (high statistical efficiency at the Gaussian distribution) but subject to the constraint that the maximum estimation error is bounded.

A very interesting topic would be to combine the robust GM-Kalman filter with the H_∞-filter in this framework. The resulting filter would have the ability to deliver all three properties: high efficiency, robustness, and minimax estimation.

Several other research topics on the GM-KF and GM-EKF can be formulated. We may investigate how the property of stabilizability developed for the classical KF and EKF applies to the new filters. Particularly, it would be interesting to perform an in-depth study on the effects of outliers occurring in other parts of the system, for example in the noise covariance matrices $\mathbf{W}$ and $\mathbf{R}$. Furthermore, it would be interesting to see the effects of using another outlier detection method instead of the Projection Statistics, such as the one proposed by Maronna and Zamar [80]. Yet another topic of high interest would be a study of the breakdown point in nonlinear regression, as the traditional definition of breakdown point for linear systems does not apply in that case.

Appendix A

EKF Applied to Climate Model

Case 1: Statistical Efficiency Test on EKF

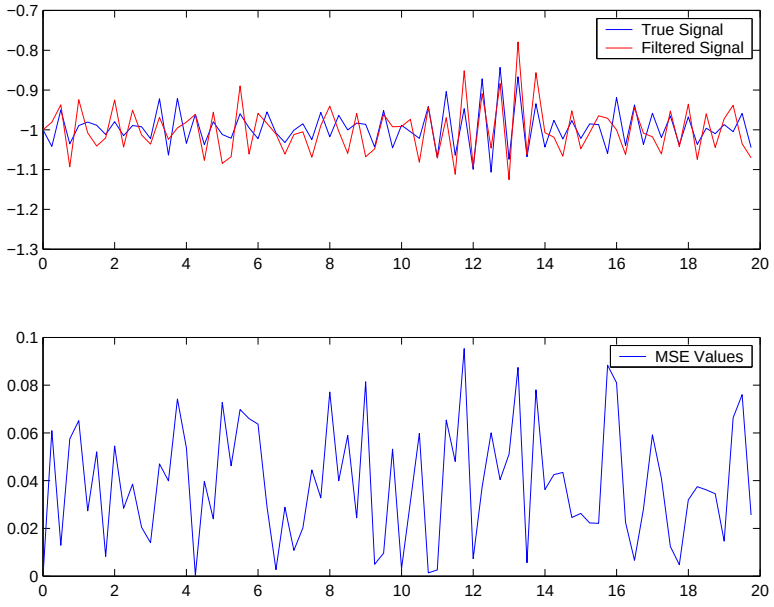


Figure A.1: Case 1 - The EKF performs satisfactory state estimation in a double-well system that is residing around one of its stable equilibrium points. Lower frame depicts the associated MSE values for the estimation segment shown in the upper frame.

Case 2: Decreasing the Observation Frequency

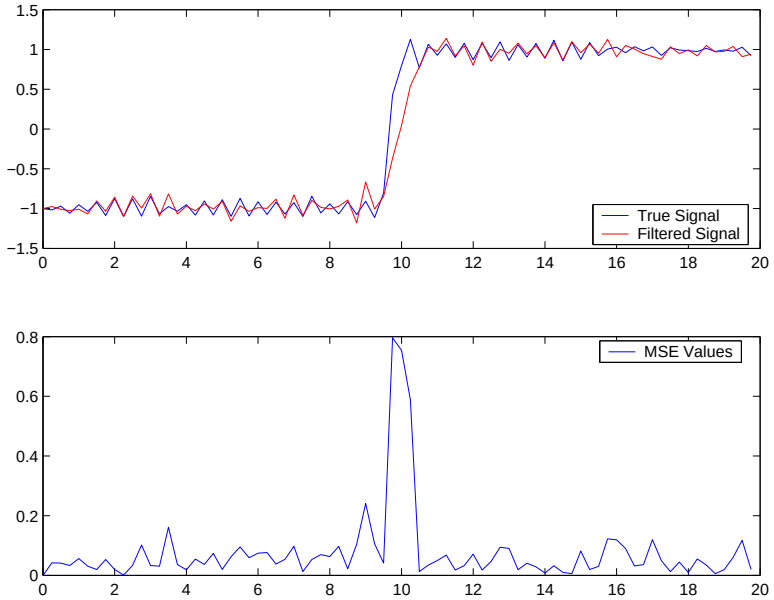


Figure A.2: Case 2 - The EKF state estimate tracks the transition with a delay of 2 seconds when the observation frequency is 4 Hz, i.e. a new observation is available at every update step.

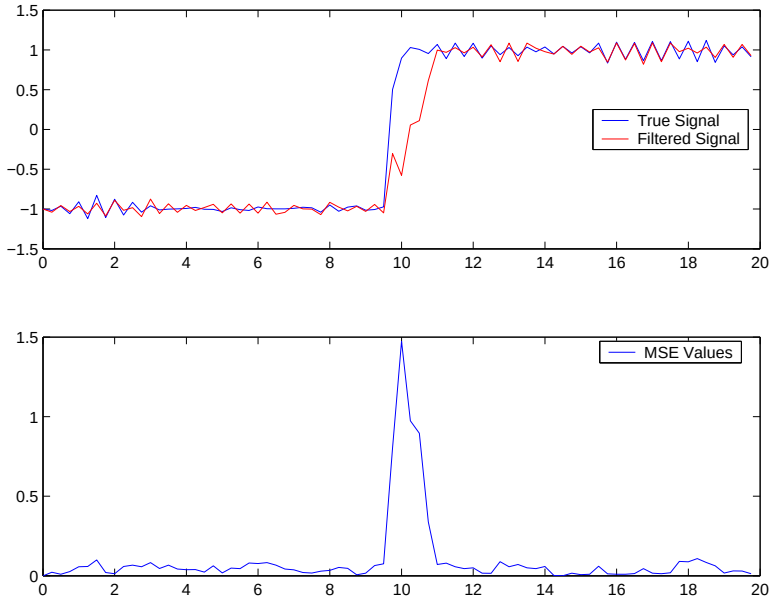


Figure A.3: Case 2 - The EKF state estimate tracks the transition with a delay of over 2 seconds when the observation frequency is 2 Hz, i.e. a new observation is available at every update step.

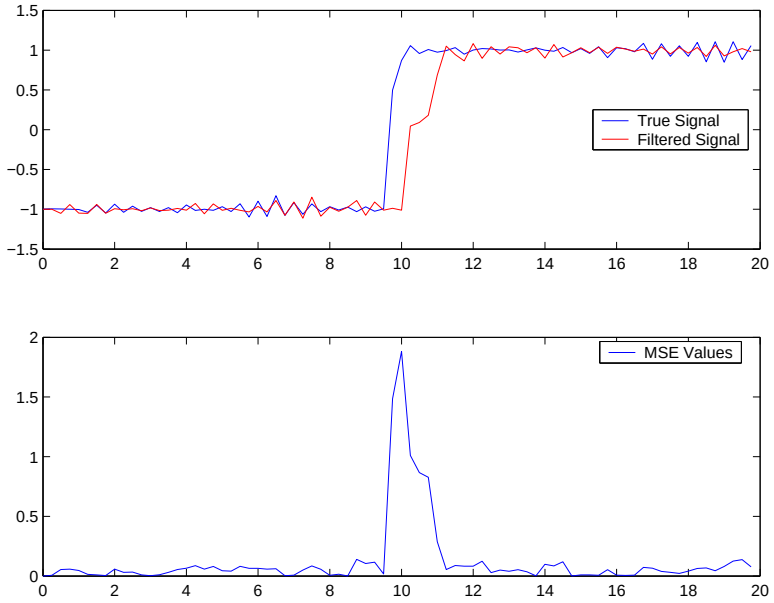


Figure A.4: Case 2 - The EKF state estimate tracks the transition with a delay of 3 seconds when the observation frequency is 1 Hz.

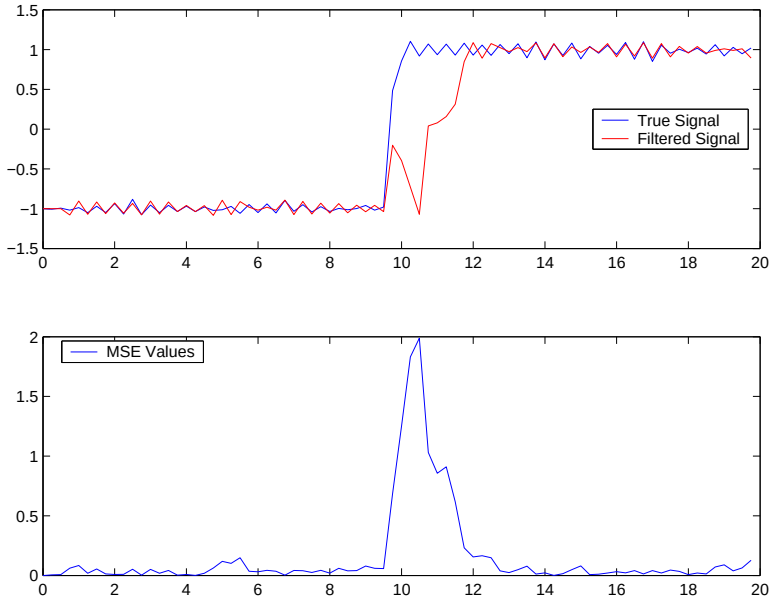


Figure A.5: Case 2 - The EKF state estimate tracks the transition with a delay of over 3 seconds when the observation frequency is 0.67 Hz, i.e. a new observation is available at every update step.

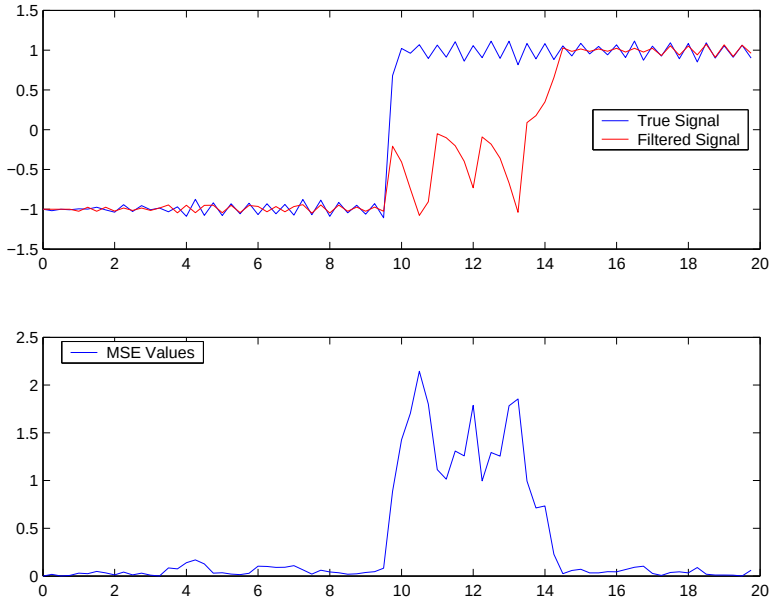


Figure A.6: Case 2 - The EKF state estimate tracks the transition when the observation frequency is 0.50 Hz, but with a time delay of 5 seconds and marginally unstable behavior prior to the transition.

Case 3: Varying the Observation Noise Variance

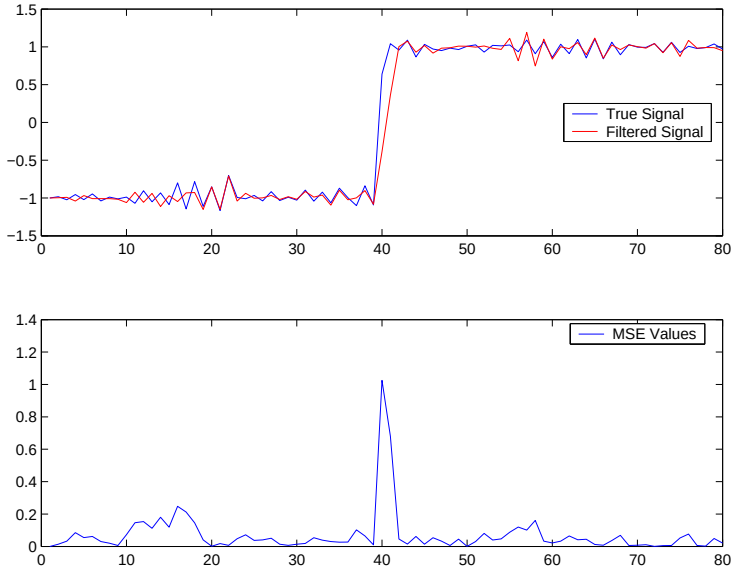


Figure A.7: Case 3 - The EKF estimates the states and tracks the transition well when σ_z^2 is not much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$.

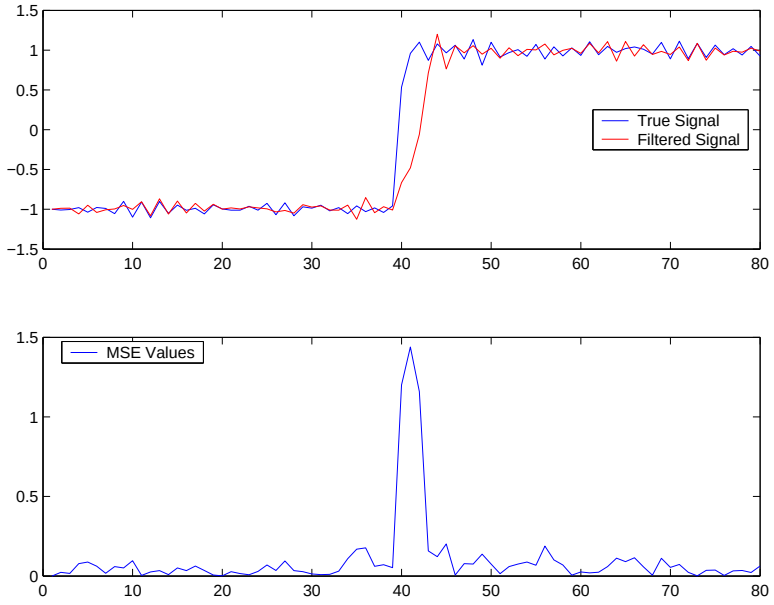


Figure A.8: Case 3 - The EKF's state estimation and transition tracking begins to deteriorate with a non-negligible delay of about 5 seconds (i.e. 20 samples at 4 Hz frequency) when σ_z^2 is sufficiently larger than σ_x^2, as shown above with $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$.

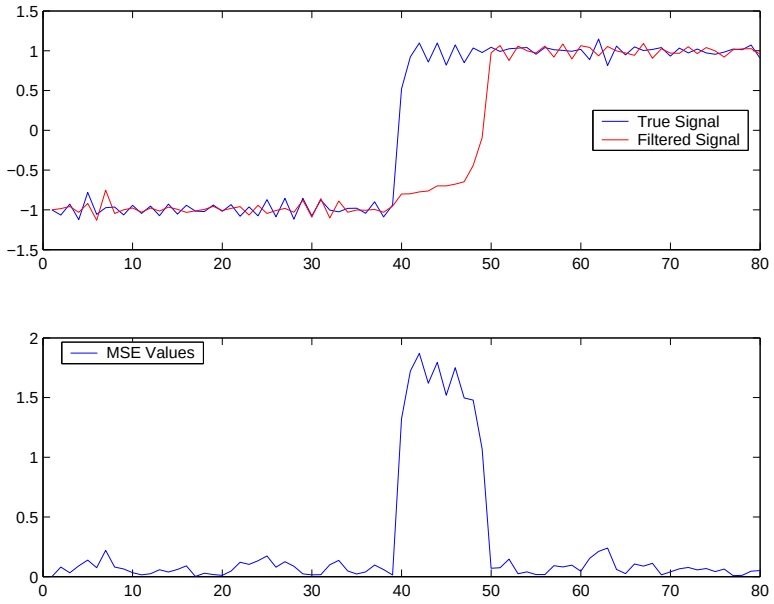


Figure A.9: Case 3 - The EKF's state estimation and transition tracking deterioration becomes more noticeable with a delay of about 10 seconds (i.e. 40 samples at 4 Hz frequency) when σ_z^2 is sufficiently larger than σ_x^2, as shown above with $\sigma_z^2 = 0.06$ and $\sigma_x^2 = 0.01$.

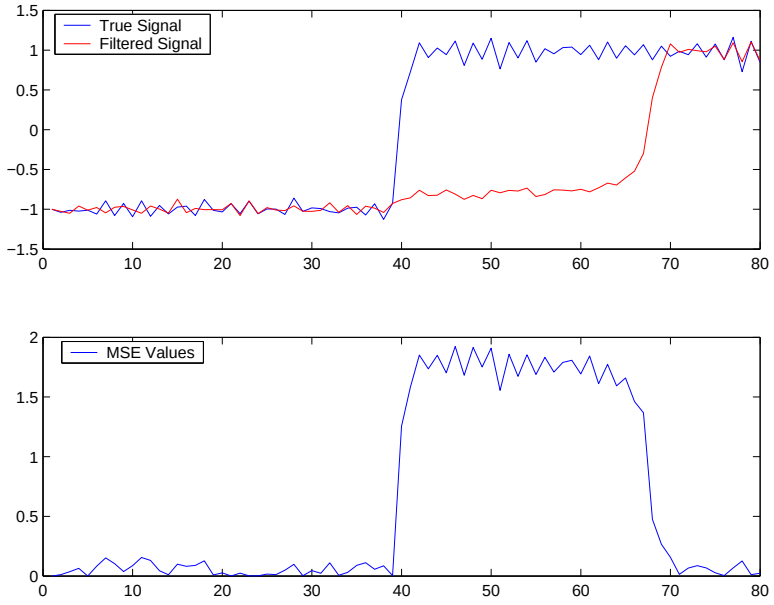


Figure A.10: Case 3 - The EKF's state estimation and transition tracking deteriorates with a non-negligible delay of over 25 seconds (i.e. 100 samples at 4 Hz frequency) when σ_z^2 is sufficiently larger than σ_x^2, as shown above with $\sigma_z^2 = 0.07$ and $\sigma_x^2 = 0.01$.

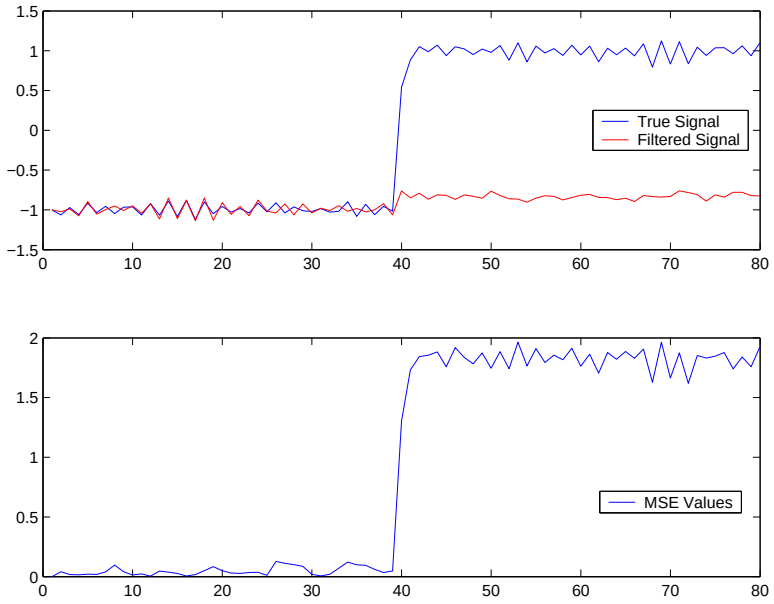


Figure A.11: Case 3 - The EKF is unable to track the transition even after 40 seconds of observations (i.e. 160 samples at 4 Hz frequency) have been processed when $\sigma_z^2 = 0.08$ and $\sigma_x^2 = 0.01$.

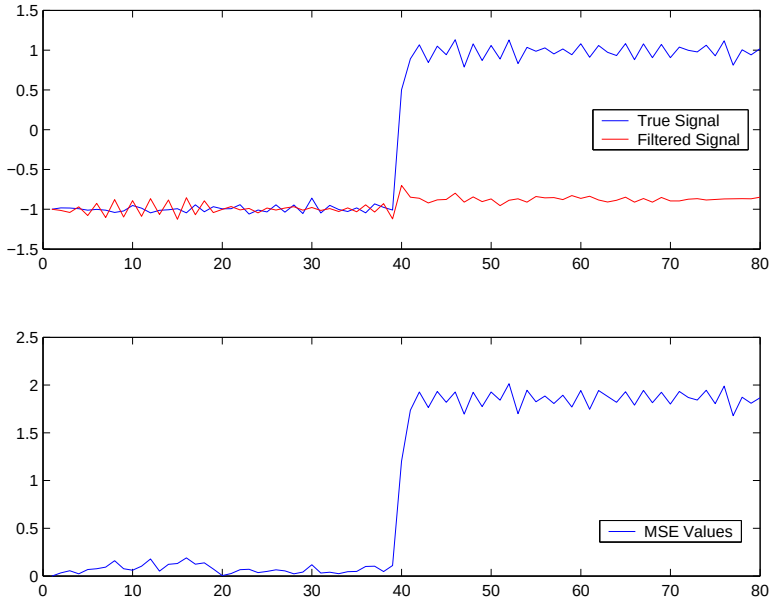


Figure A.12: Case 3 - The EKF is unable to track the transition even after 40 seconds of observations (i.e. 160 samples at 4 Hz frequency) have been processed when $\sigma_z^2 = 0.10$ and $\sigma_x^2 = 0.01$.

Case 4: Rapid Transitions with Increasing Observation Noise Variance

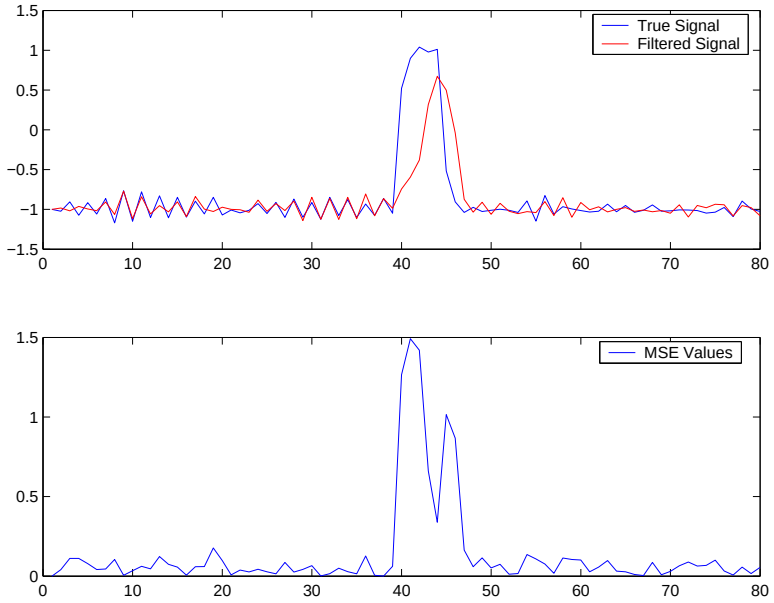


Figure A.13: Case 4 - The EKF tracks a transition that is 5 samples long if the observation noise variance is sufficiently low, i.e. $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, albeit with a noticeable delay.

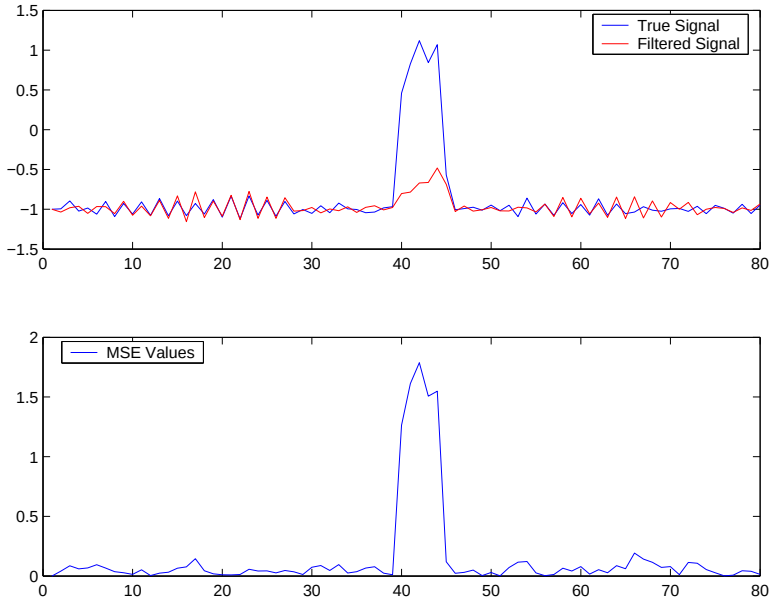


Figure A.14: Case 4 - The EKF is unable to track a transition that is 5 samples long if the observation noise variance exceeds a particular threshold, in this case $\sigma_z^2 = 0.05$ with $\sigma_x^2 = 0.01$.

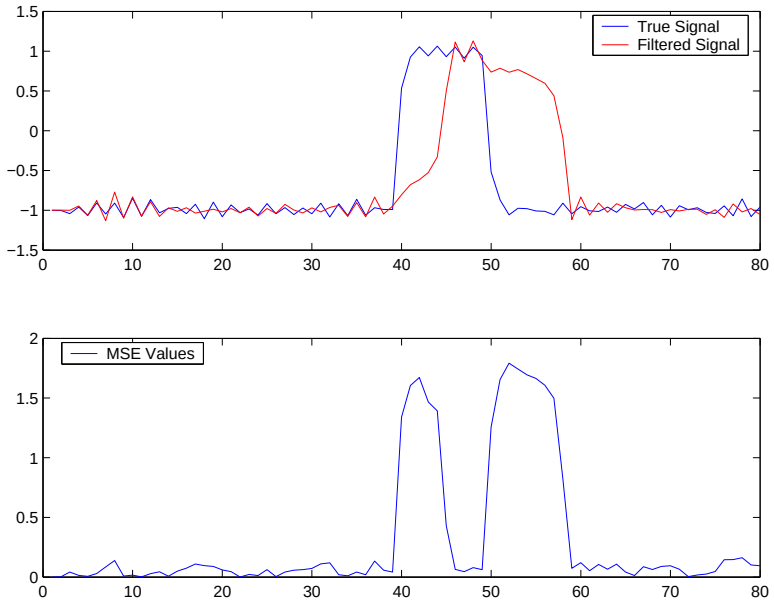


Figure A.15: Case 4 - The EKF tracks a transition that is 10 samples long if the observation noise variance is sufficiently low, i.e. $\sigma_z^2 = 0.06$ and $\sigma_x^2 = 0.01$, albeit with a noticeable delay.

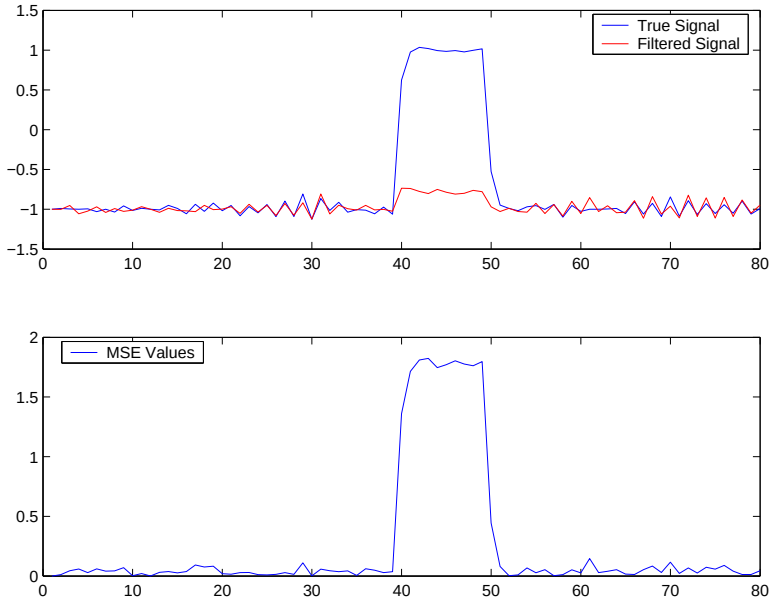


Figure A.16: Case 4 - The EKF is unable to track a transition that is 10 samples long if the observation noise variance exceeds a particular threshold, in this case $\sigma_z^2 = 0.07$ with $\sigma_x^2 = 0.01$.

Case 5: Effects of Observation Outlier

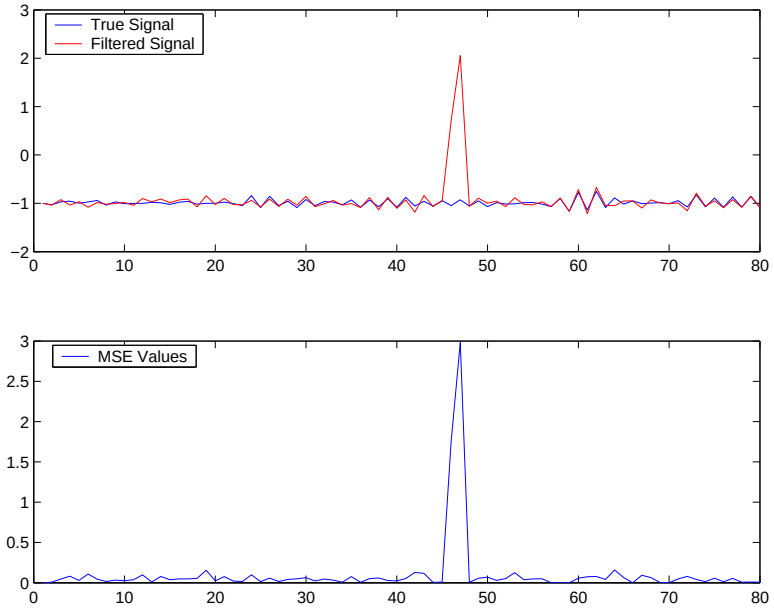


Figure A.17: Case 5 - The EKF inaccurately outputs a state transition when outlying observations are processed.

Case 6: Varying Sampling Frequency and Observation Noise Variance

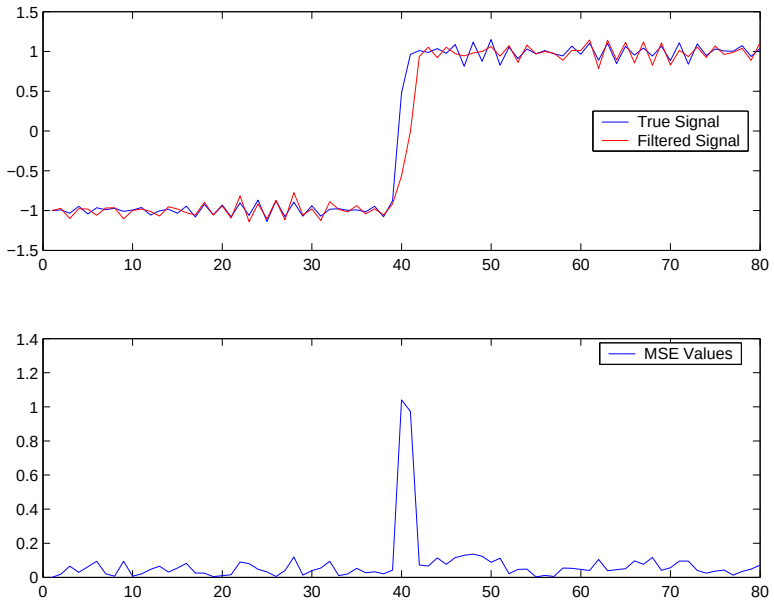


Figure A.18: Case 6 - The EKF tracks the system transition with $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and 4 Hz sampling frequency.

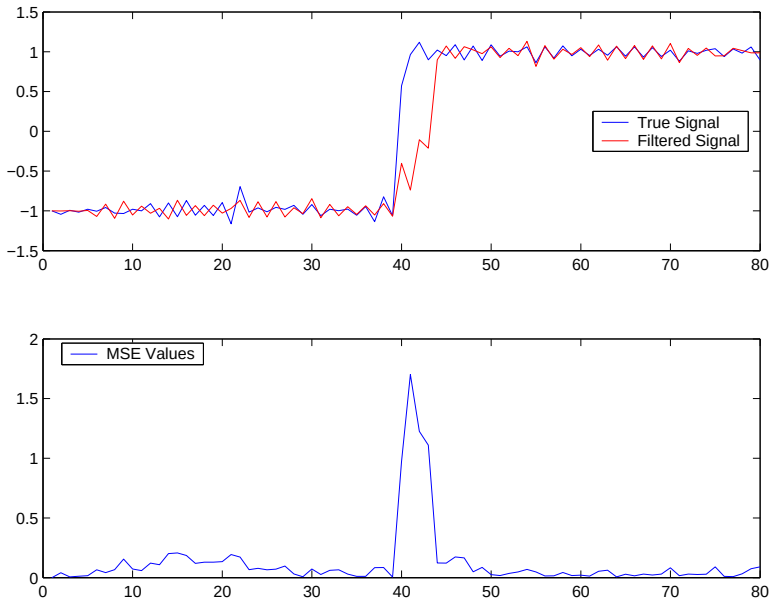


Figure A.19: Case 6 - The EKF begins to show some deterioration in tracking the system transition with $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and 2 Hz sampling frequency.

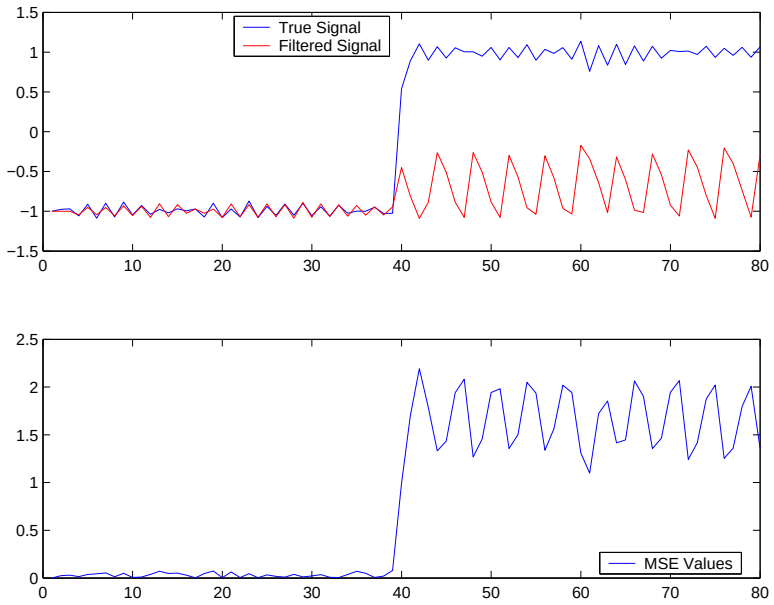


Figure A.20: Case 6 - The EKF is unable to track the system transition even though $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$, due to the added complexity of a low sampling frequency of 1 Hz.

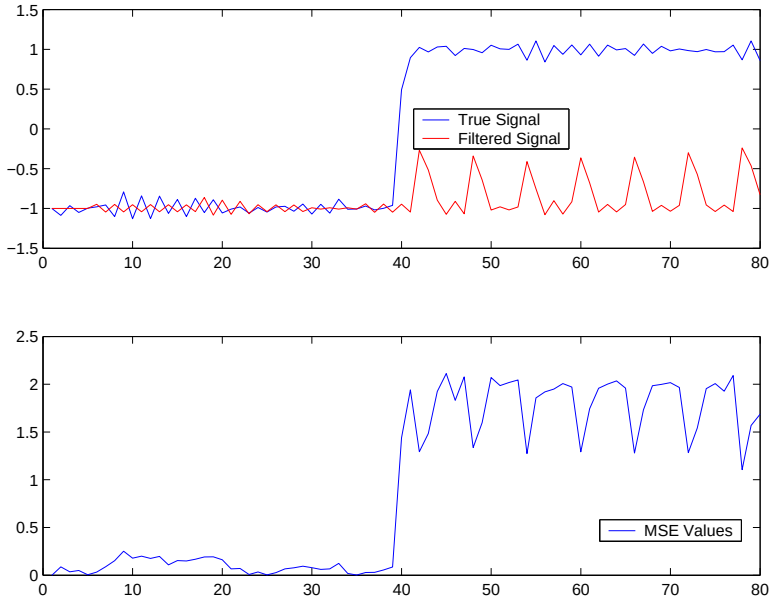


Figure A.21: Case 6 - The EKF is unable to track the system transition even though $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$, due to the added complexity of a low sampling frequency of 0.67 Hz.

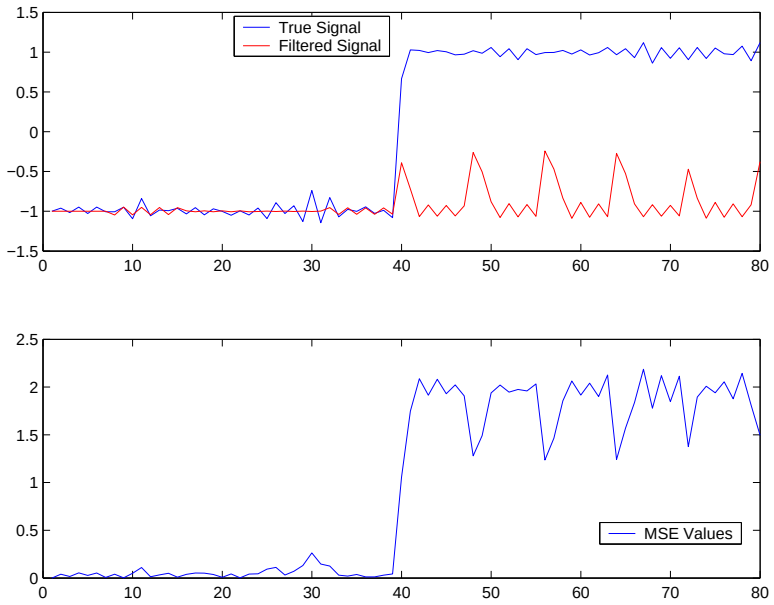


Figure A.22: Case 6 - The EKF is unable to track the system transition even though $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$, due to the added complexity of a low sampling frequency of 0.50 Hz.

Appendix B

GM-EKF Applied to Climate Model

Case 1: Statistical Efficiency Test on EKF

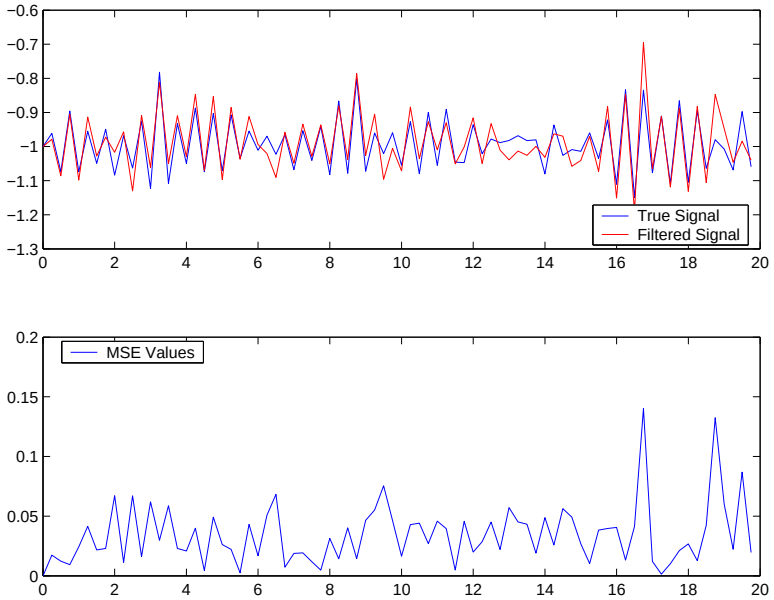


Figure B.1: Case 1 - The EKF performs satisfactory state estimation in a double-well system that is residing around one of its stable equilibrium points. Lower frame depicts the associated MSE values for the estimation segment shown in the upper frame.

Case 2: Decreasing the Observation Frequency

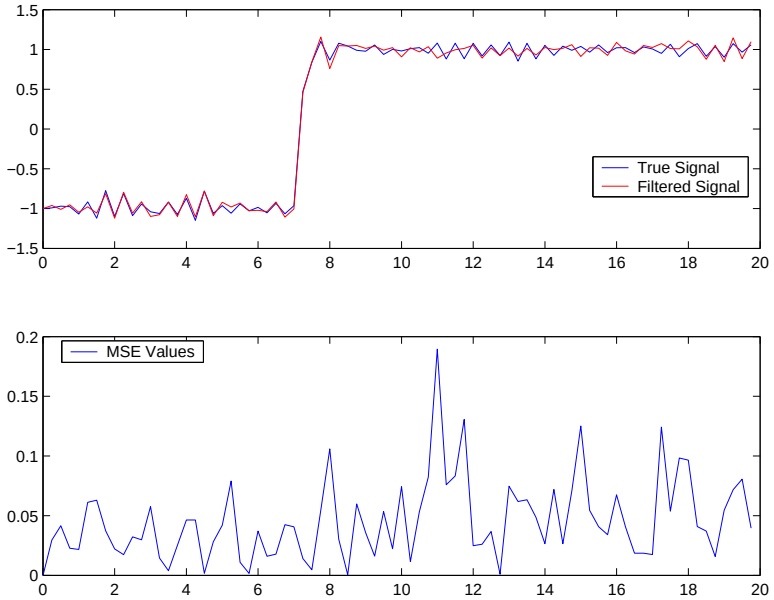


Figure B.2: Case 2 - The GM-EKF state estimate tracks the transition with no delays when the observation frequency is 4 Hz, i.e. a new observation is available at every update step.

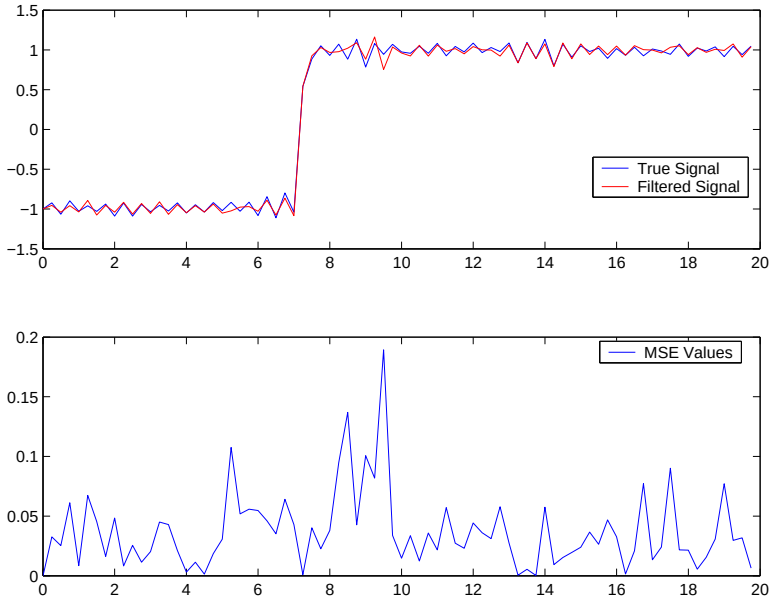


Figure B.3: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 2 Hz.

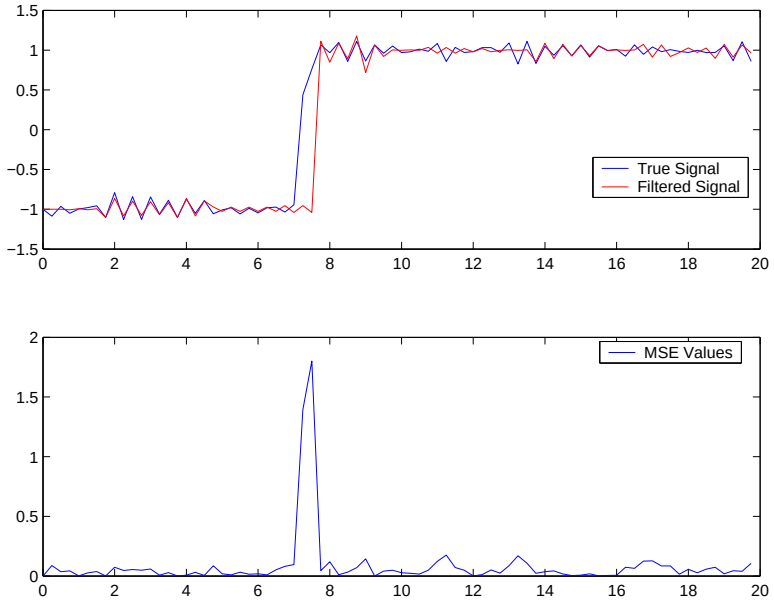


Figure B.4: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 1 Hz.

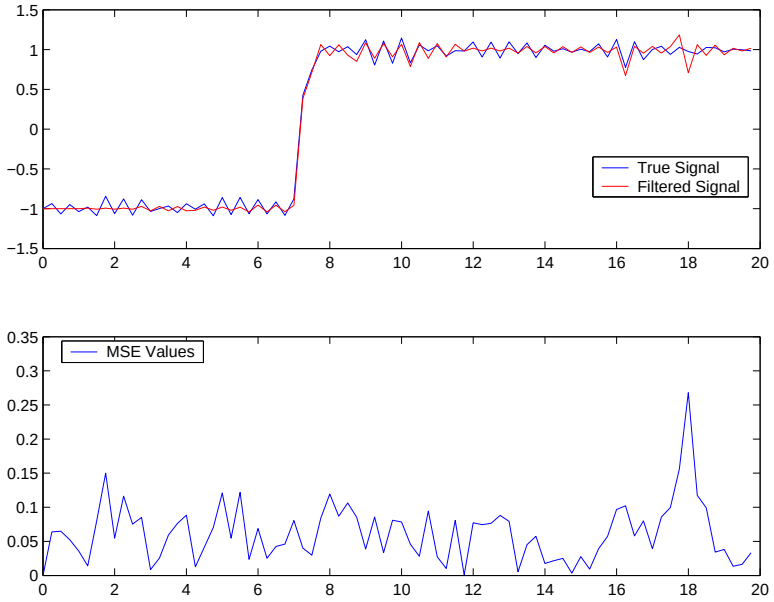


Figure B.5: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 0.67 Hz.

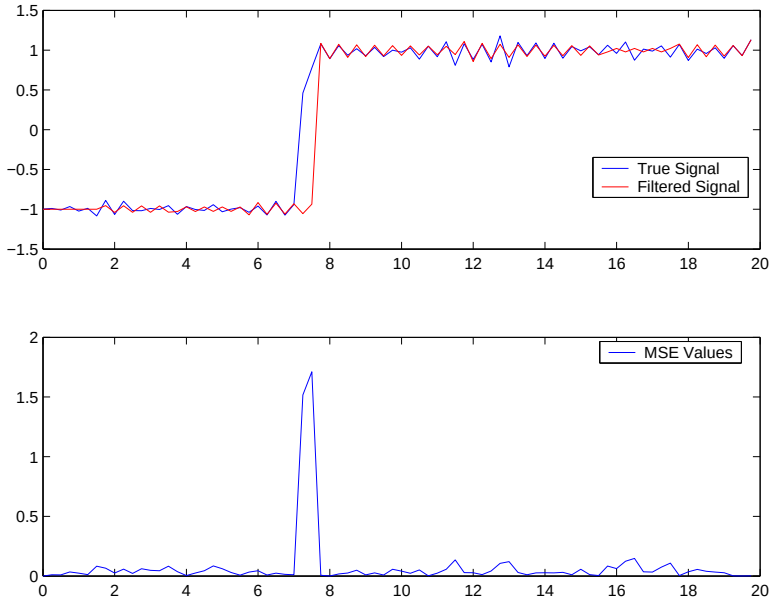


Figure B.6: Case 2 - The GM-EKF state estimate continues to track the transition very well when the observation frequency is 0.50 Hz.

Case 3: Varying the Observation Noise Variance

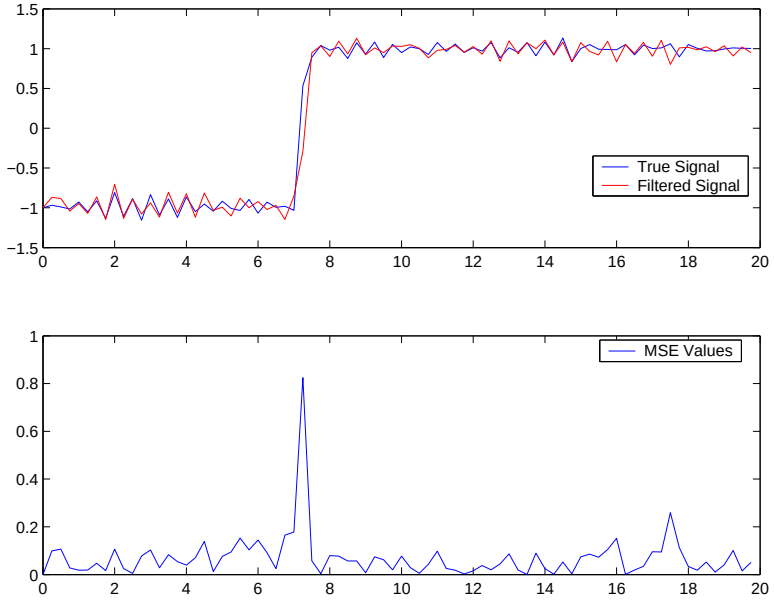


Figure B.7: Case 3 - The GM-EKF estimates the states and tracks the transition well when σ_z^2 is not much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.02$ and $\sigma_x^2 = 0.01$.

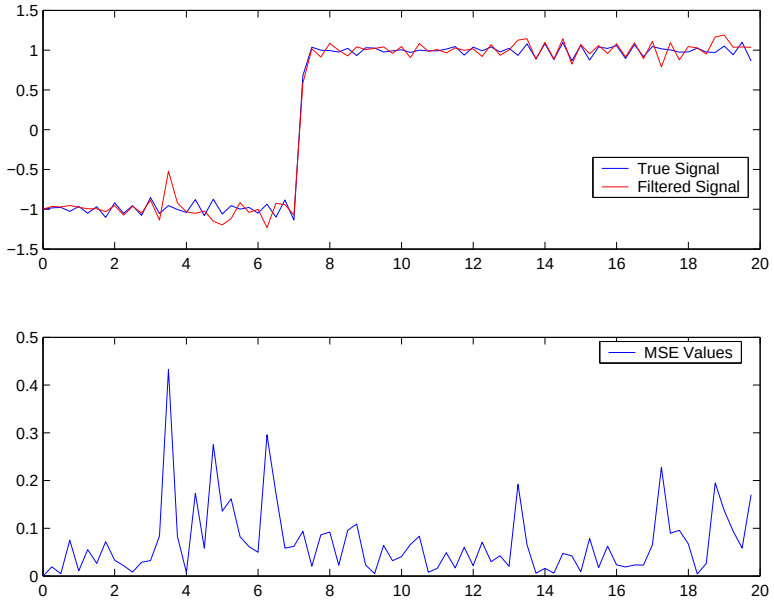


Figure B.8: Case 3 - The GM-EKF estimates the states and tracks the transition well when σ_z^2 is not much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$.

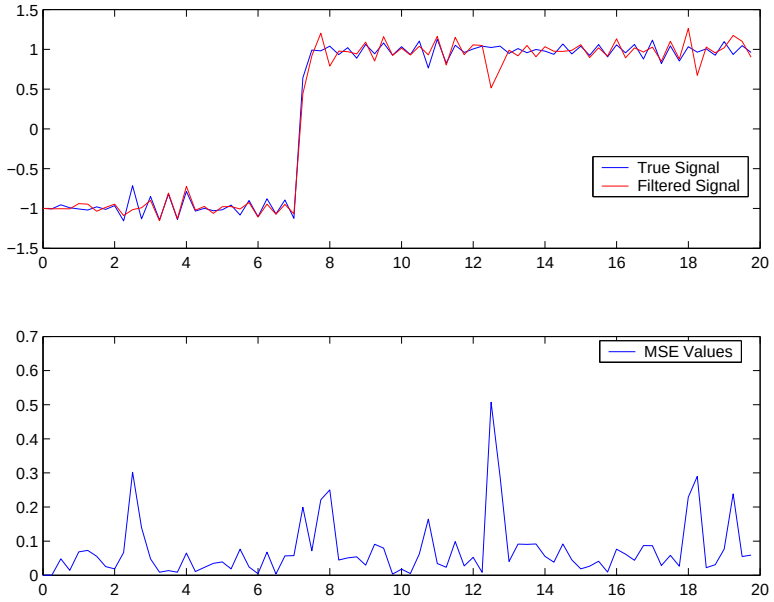


Figure B.9: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.06$ and $\sigma_x^2 = 0.01$.

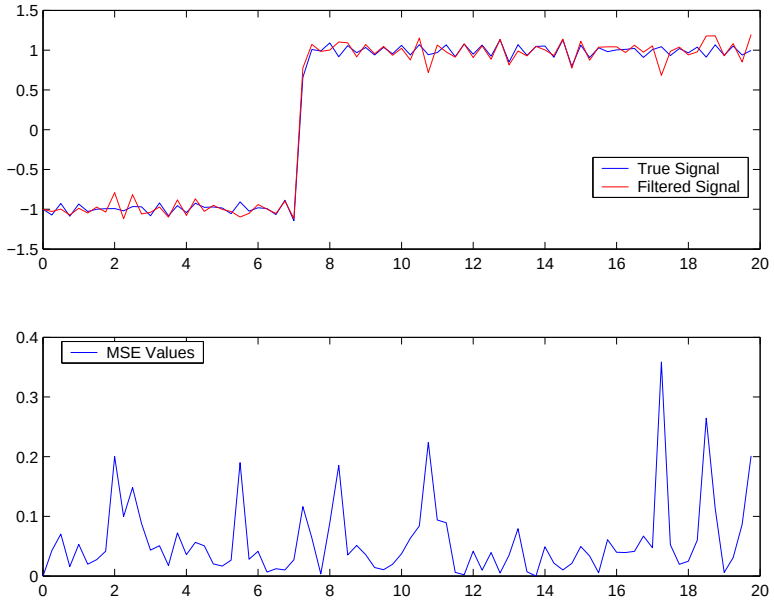


Figure B.10: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.07$ and $\sigma_x^2 = 0.01$.

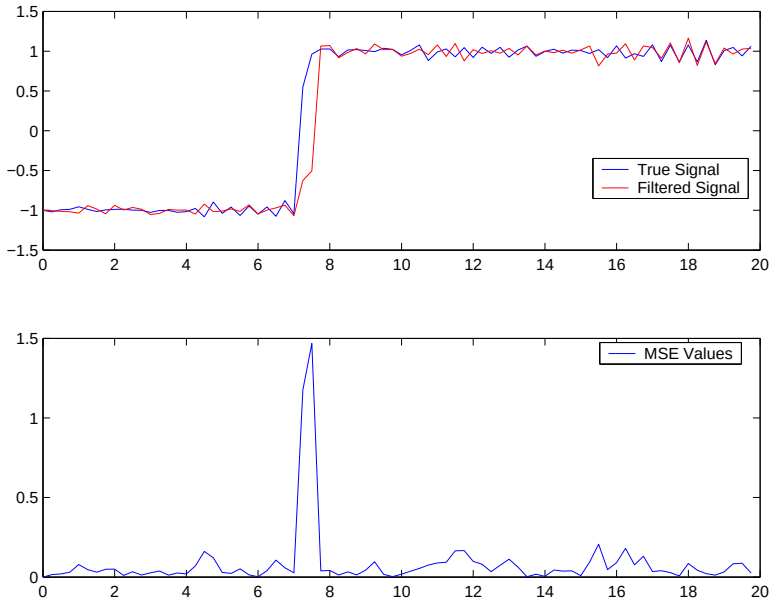


Figure B.11: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.08$ and $\sigma_x^2 = 0.01$ (a case in which the standard EKF could not track the transition even after 40 seconds of observation data was processed).

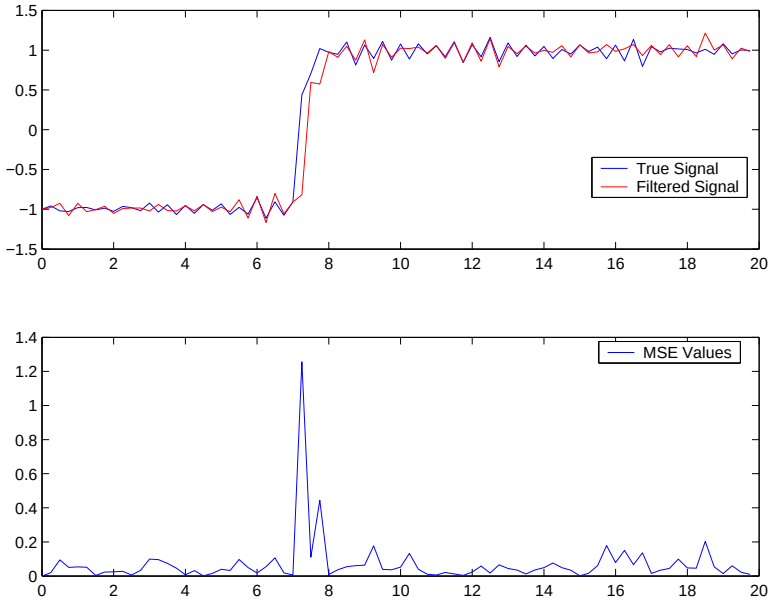


Figure B.12: Case 3 - The GM-EKF continues to maintain good state estimation and transition tracking even when σ_z^2 is much larger than σ_x^2, as shown above with $\sigma_z^2 = 0.10$ and $\sigma_x^2 = 0.01$ (a case beyond which the standard EKF could track transitions).

Case 4: Rapid Transitions with Increasing Observation Noise Variance

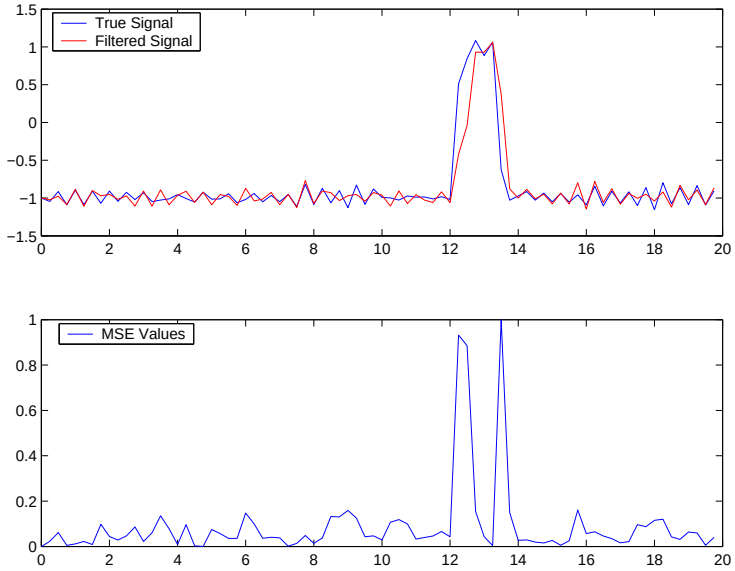


Figure B.13: Case 4 - The GM-EKF tracks a rapid transition if the observation noise variance is sufficiently low, i.e. $\sigma_z^2 = 0.04$ and $\sigma_x^2 = 0.01$, with almost no delay.

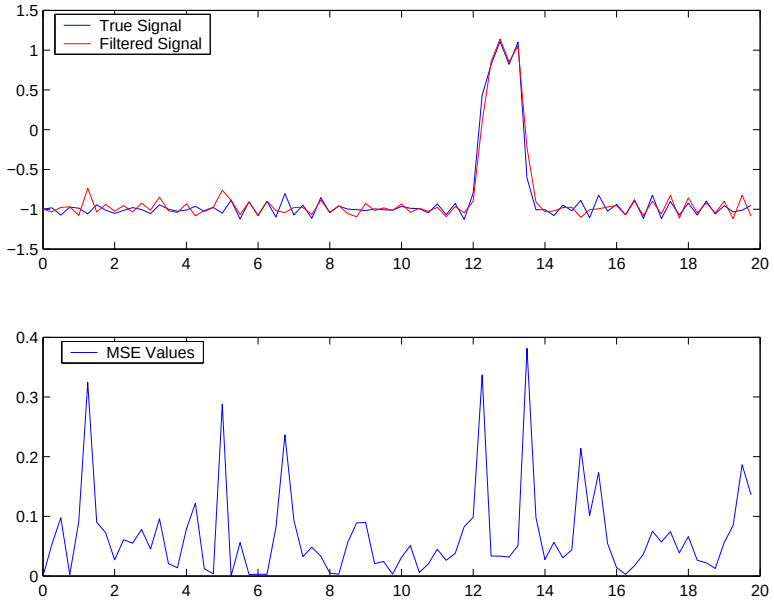


Figure B.14: Case 4 - The GM-EKF tracks a rapid transition even if the observation noise variance is $\sigma_z^2 = 0.05$ and $\sigma_x^2 = 0.01$, a case in which the standard EKF was unable to track the transition.

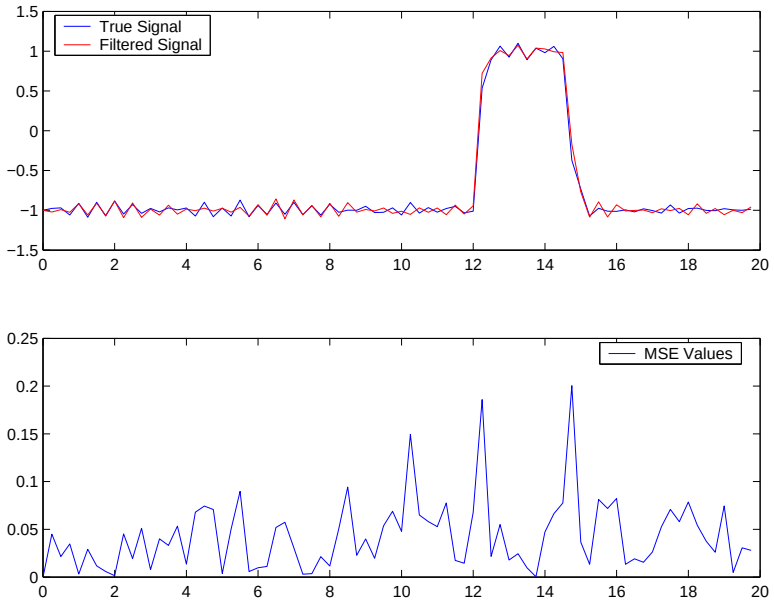


Figure B.15: Case 4 - The GM-EKF tracks a rapid transition even if the observation noise variance is $\sigma_z^2 = 0.06$ and $\sigma_x^2 = 0.01$, a case beyond which the standard EKF was able to track a transition.

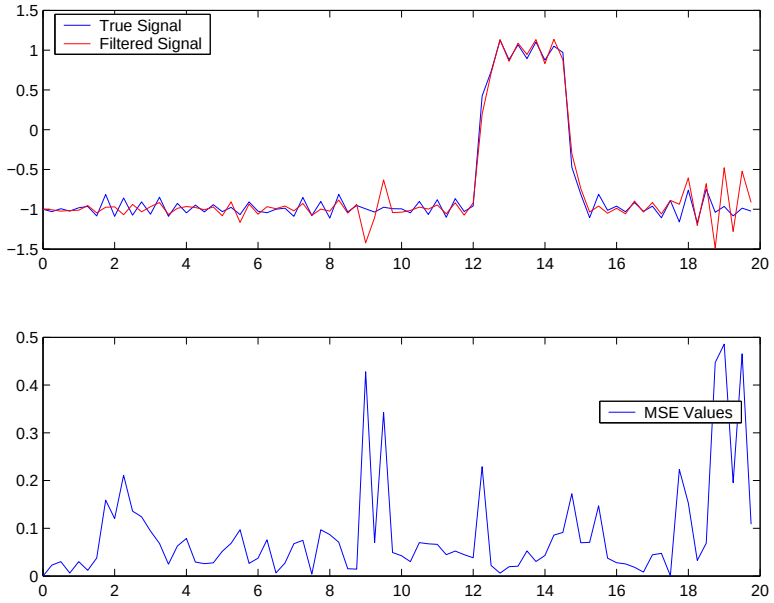


Figure B.16: Case 4 - The GM-EKF tracks a rapid transition even if the observation noise variance is $\sigma_z^2 = 0.07$ and $\sigma_x^2 = 0.01$, a case beyond which the standard EKF was able to track a transition.

Case 5: Effects of Observation Outlier

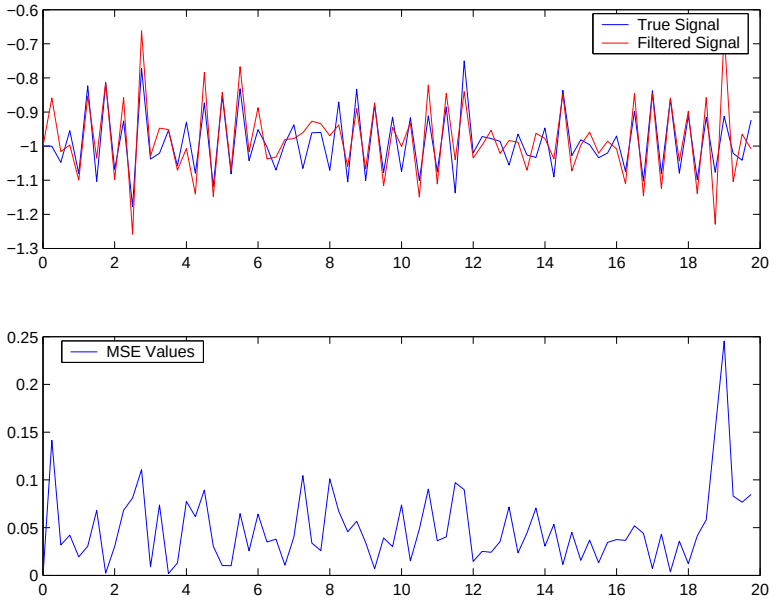


Figure B.17: Case 5 - The GM-EKF accurately recognizes that one of the two observations being processed is an outlier, and does not output a state transition.

Case 6: Varying Sampling Frequency and Observation Noise Variance

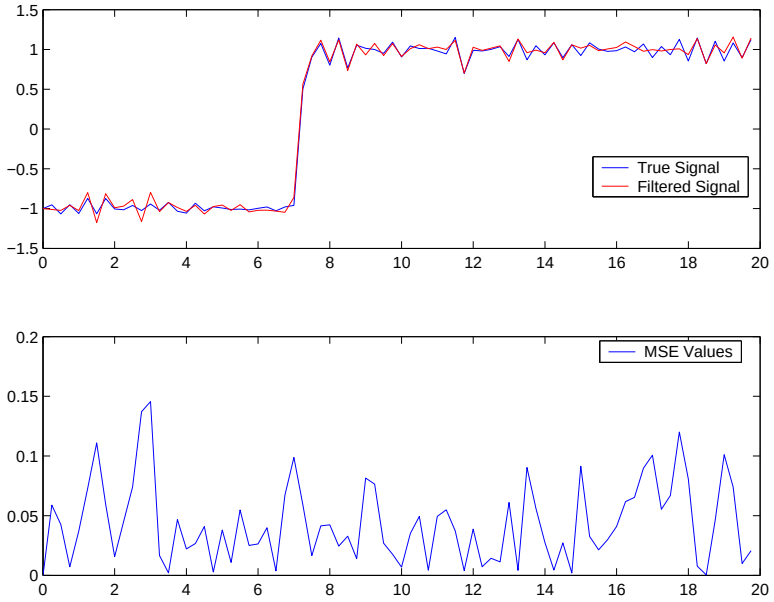


Figure B.18: Case 6 - The EKF tracks the system transition with $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and 4 Hz sampling frequency.

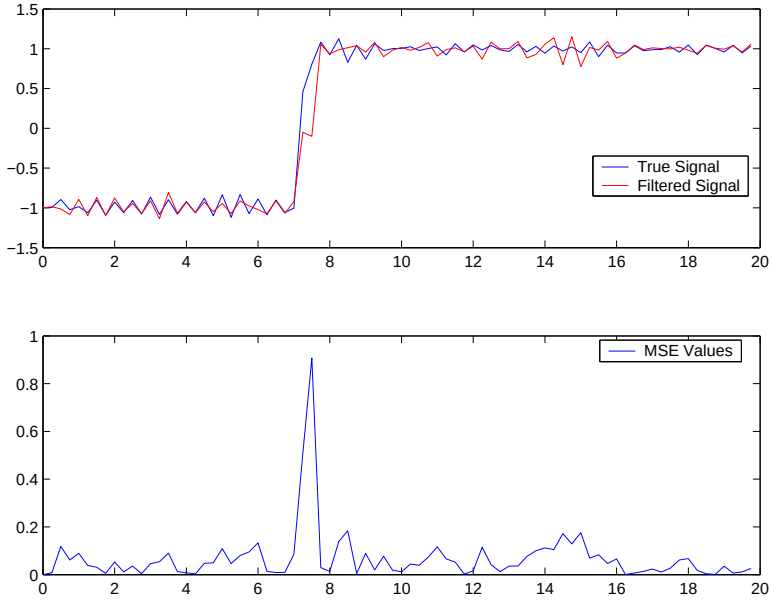


Figure B.19: Case 6 - The GM-EKF tracks the system transition with multiple effects $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and sampling frequency is 2 Hz.

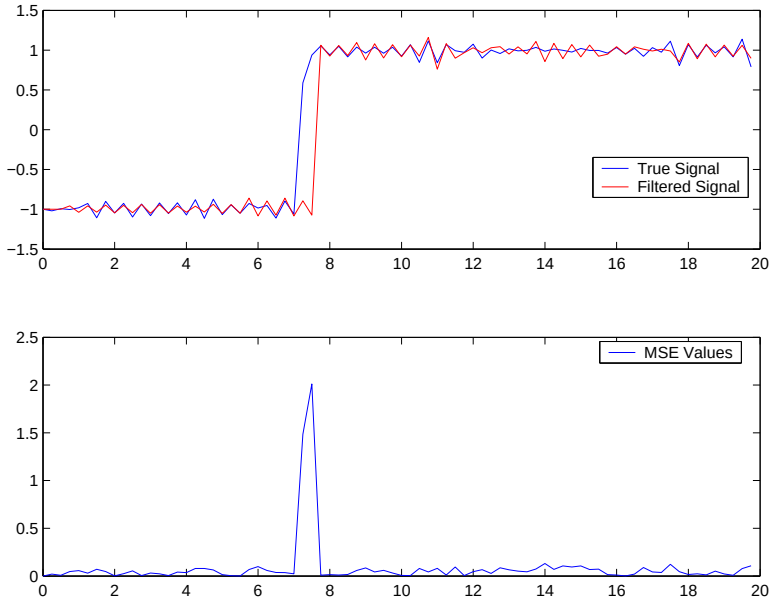


Figure B.20: Case 6 - The GM-EKF tracks the system transition with multiple effects $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and sampling frequency is 1 Hz.

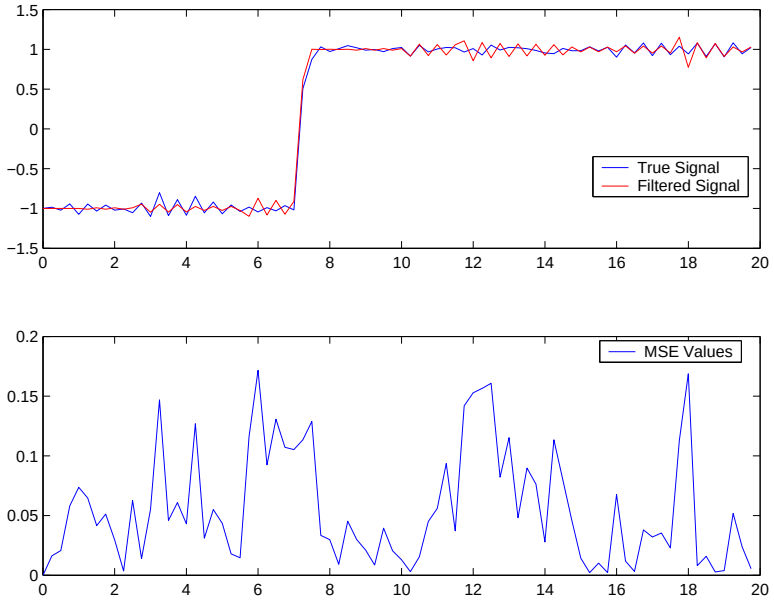


Figure B.21: Case 6 - The GM-EKF tracks the system transition with multiple effects $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and sampling frequency is 0.67 Hz.

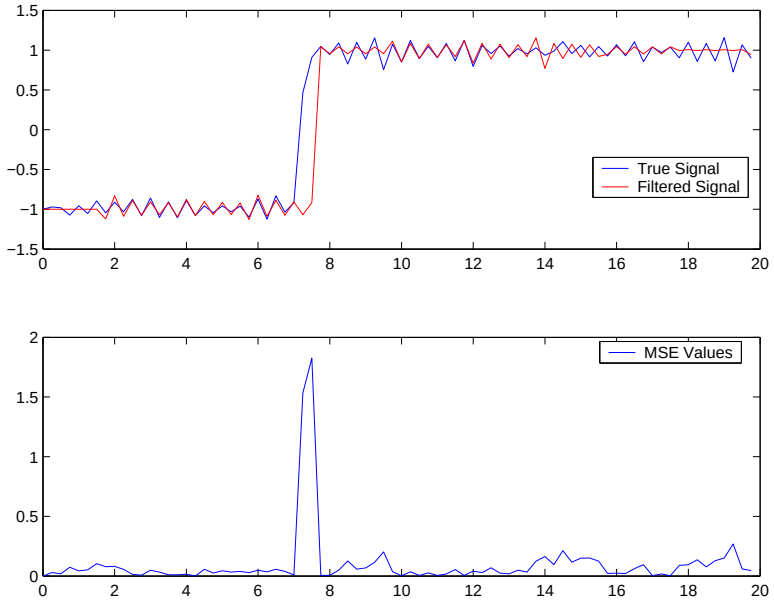


Figure B.22: Case 6 - The GM-EKF tracks the system transition with multiple effects $\sigma_z^2 = 0.02$, $\sigma_x^2 = 0.01$, and sampling frequency is 0.50 Hz.

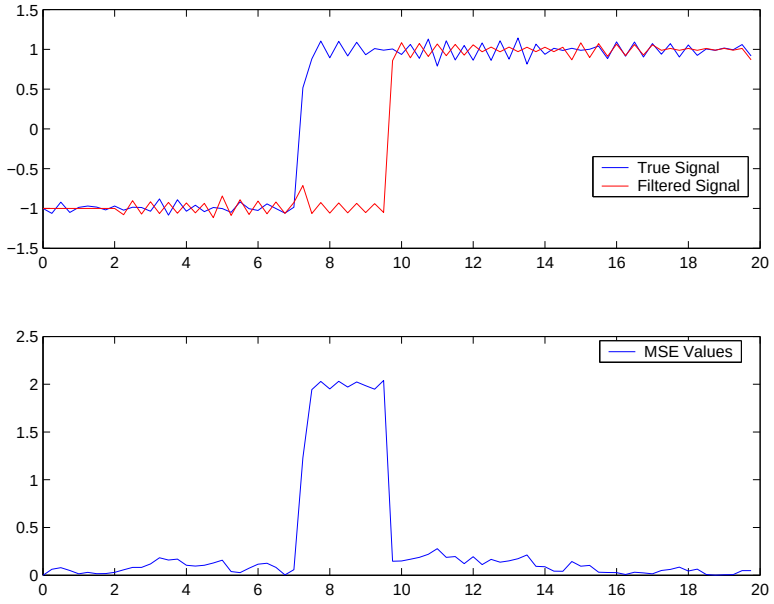


Figure B.23: Case 6 - The GM-EKF shows signs of deterioration only when $\sigma_z^2 = 0.10$, $\sigma_x^2 = 0.01$, and sampling frequency is 0.4 Hz, which indicates that much more extreme cases of observation sampling frequency and noise variance can be handled by the new filter in comparison to the standard EKF.

Bibliography

[1] John Aldrich. R. A. Fisher and the making of maximum likelihood, 1912-1922. *Statistical Science*, 12(3):162–176, 1997.

[2] B.D.O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, Inc., New Jersey, 1979.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, New York, 2003.

[4] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[5] A. C. Atkinson. Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):1–36, 1982.

[6] S. Ayat, T.M. Manzuri, R. Dianat, and J. Kabudian. An improved spectral subtraction speech enhancement system by using an adaptive spectral estimator. In *Canadian Conference on Electrical and Computer Engineering*, pages 261–264, May 2005.

[7] R. K. Bansal and P. Papantoni-Kazakos. Outlier-resistant algorithms for detecting a change in a stochastic process. *IEEE Transactions on Information Theory*, 35(3):521–535, 1989.

[8] O. Bar-Shalom and A. J. Weiss. Doa estimation using one-bit quantized measurements. *IEEE Trans. On Aerospace and Electronic Systems*, 38(3):868–884, 2002.

[9] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, Chichester, 1978.

[10] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, New York, 1994.

[11] J. R. Barry, D. G. Messerschmitt, and E. A. Lee. *Digital Communication*. Springer, 2003.

[12] Jounghoon Beh and Hanseok Ko. A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. *IEEE International Conference on Multimedia and Expo*, 3:633–636, 2003.

[13] R. Benzi, G. Parisi, A. Sutera, and A. Vulpiani. Stochastic resonance in climatic change. *Tellus*, 34:10–16, 1982.

[14] R. Benzi, G. Parisi, A. Sutera, and A. Vulpiani. A theory of stochastic resonance in climatic change. *SIAM Journal of Applied Mathematics*, 43:565–578, 1983.

[15] N. M. Blachman. *Noise and its Effect on Communication*. McGraw-Hill, New York, 1966.

[16] G.E.P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and control*. Holden-Day, San Francisco, 1970.

[17] Robert Grover Brown and Patrick Y. C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering with Matlab Exercises and Solutions*. Wiley, 1996.

[18] R. S. Bucy and K. D. Senne. Realization of optimum discrete-time nonlinear estimators. In *Proceedings of the Symposium on Nonlinear Estimation Theory and Its Applications*, volume 83, pages 15–108, 1970.

[19] G. Burger and M. Cane. Interactive kalman filtering. *Journal of Geophysical Research*, 90(C4):8015–8031, 1994.

[20] R. W. Butler, P. L. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *Annals of Statistics*, 21:1385–1400, 1985.

[21] C. Chen and L. M. Liu. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88:284–297, 1993.

[22] G. S. Christensen and S. A. Soliman. Optimal filtering of linear discrete dynamic systems based on least absolute value approximations. *Automatica*, 26(2):389–395, 1990.

[23] Philip Anthony D'Ambrosio. A differential geometry-based algorithm for solving the minimum hellinger distance estimator. Master's thesis, Virginia Polytechnic Institute and State University, 2008.

[24] Laurie Davies. The asymptotics of rousseeuw's minimum volume ellipsoid estimator. *Annals of Statistics*, 20(4):1828–1843, 1992.

[25] Piet de Jong and Jeremy Penzer. Diagnosing shocks in time series. *Journal of the American Statistical Association*, 93(442):796–806, 1998.

[26] G. Doblinger. Adaptive kalman smoothing of ar signals disturbed by impulses and colored noise. In *Proceedings of IEEE Symposium in Digital Filtering and Signal Processing*, pages 72–76, 1998.

[27] D. L. Donoho. Breakdown properties of multivariate location estimators. PhD Qualifying Paper, 1982.

[28] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[29] J. Doyle, K. Glover, P. Khargonekar, and B. Francis. State-space solutions to standard h_2 and h_∞ control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, 1989.

[30] G. Durgaprasad and S. S. Thakur. Robust dynamic state estimation of power systems based on m-estimation and realistic modeling of system dynamics. *IEEE Transactions on Power Systems*, 13(4):1331–1336, 1998.

[31] Z. M. Durovic and B. D. Kovacevic. Robust estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 44(6):1292–1296, 1999.

[32] A. J. Efron and H. Jeen. Pre-whitening for detection in correlated plus impulsive noise. *Proceedings of IEEE ICASSP*, II:469–472, 1992.

[33] A. J. Efron and H. Jeen. Detection in impulsive noise based on robust whitening. *IEEE Transactions on Signal Processing*, 42(6):1572–1576, 1994.

[34] A. J. Efron, P. F. Swaszek, and D. W. Tufts. Insight into detection of deterministic and gaussian signals in correlated plus impulsive noise environments. *IEEE Transactions on Signal Processing*, 39(3):603–611, 1991.

[35] G. Eyink and J. Restrepo. Most probable histories for nonlinear dynamics: Tracking climate transition. *Journal of Statistical Physics*, 101(1-2):459–472, 2000.

[36] Ashil Farahmand. Cooperative decentralized intersection collision avoidance using extended kalman filtering. Master's thesis, Virginia Polytechnic Institute and State University, 2008.

[37] L. Fernholz. Von mises calculus for statistical functionals, lecture notes in statistics. Technical report, Virginia Polytechnic Institute and State University, New York, 1983.

[38] R. A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700–725, 1925.

[39] Jason Ford. Non-linear and robust filtering: From the kalman filter to the particle filter. Technical Report DSTO-TR-1301, Defense Science and Technology Organization, Australian Department of Defense, 2002.

[40] A. J. Fox. Outliers in time series. *Journal of the Royal Statistical Society*, 34:350–363, 1972.

[41] Sharon Gannot, David Burshtein, and Ehud Weinstein. Iterative and sequential kalman filter-based speech enhancement algorithms. *IEEE Transactions on Speech and Audio Processing*, 6(4):373–385, 1998.

[42] M. Gasko and D. Donoho. Influential observation in data analysis. *American Statistical Association, Proceedings of the Business and Economic Statistics Section*, pages 104–110, 1982.

[43] M. Gasko and D. Donoho. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

[44] U. Gather and C. Becker. Outlier identification and robust methods. *Handbook of Statistics: Robust Inference*, 15:123–143, 1997.

[45] Arthur Gelb. *Applied Optimal Estimation*. MIT Press, Cambridge, 1974.

[46] C.F. Gerald and P.O. Wheatley. *Applied Numerical Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1997.

[47] M. Ghil and P. Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in Geophysics*, 33:141–266, 1991.

[48] N. Gordon, D. Salmond, and A. Smith. A novel approach to non-linear and non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

[49] Andrew Green. *Orbit Determination and Prediction Processes for Low Altitude Satellites*. PhD thesis, Massachusetts Institute of Technology, 1979.

[50] Michael Green. An introduction to h_∞ control. *Control*, 1992.

[51] M. Grimble and M. Johnson. h_∞ robust control design - a tutorial review. *Computing and Control Engineering Journal*, 2(6):275–282, 1991.

[52] Michael Grimble. Solution of the h_∞ optimal linear filtering problem for discrete-time systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1092–1104, 1990.

[53] F. R. Hampel. A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896, 1971.

[54] F. R. Hampel. The influence curve and its role in robust estimation. *JASA: Theory and Methods*, 69:383–393, 1974.

[55] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Inc., New York, 1986.

[56] J. E. Handschin and D. Q. Mayne. Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9(5):547–559, 1969.

[57] D. M. Hawkins. A feasible solution algorithm for the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 17:197–210, 1994.

[58] D. M. Hawkins and D. J. Olive. Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis*, 30:1–11, 1999.

[59] Heikki Hella. *On Robust ESACF Identification of Mixed ARIMA Models*. PhD thesis, Bank of Finland Studies, 2003.

[60] I. Hoteit, D. Pham, G. Triantafyllou, and G. Korres. Particle kalman filtering for data assimilation in meteorology and oceanography. In *Third WCRP International Conference on Reanalysis*, 2008.

[61] P. J. Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[62] Peter J. Huber. *Robust Statistics*. John Wiley & Sons, Inc., New York, 1981.

[63] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1972.

[64] B. H. Juang. The past, present, and future of speech processing. *IEEE Signal Processing Magazine*, 1053:24–28, 1998.

[65] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, New Jersey, 2000.

[66] R. Kaiser and A. Maravall. Seasonal outliers in time series. *Working Paper No 9915*, 1999.

[67] R. E. Kalman. A new approach to linear filtering and prediction theory. *Transactions of the ASME. Journal of Basic Engineering.*, 82(0):35–45, 1960.

[68] R. E. Kalman. New methods in wiener filtering theory. *Proceedings of the Symposium on Eng. Appl. Random Function Theory and Probability*, pages 270–388, 1963.

[69] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Transactions of the ASME. Journal of Basic Engineering.*, 83(D):15–108, 1961.

[70] S. Kosanam and D. Simon. Kalman filtering with uncertain noise covariances. In *Intelligent Systems and Control*, pages 375–379, 2004.

[71] B. Kovacevic, Z. Durovic, and S. Glavaski. On robust kalman filtering. *International Journal of Control*, 56(3):547–562, 1992.

[72] W. S. Krasker and R. E. Welsch. Efficient bounded influence regression estimation. *Journal of the American Statistical Association*, 77:595–604, 1982.

[73] Y. H. Lee and S. A. Kassam. Generalized median filtering and related nonlinear filtering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33:672–683, 1985.

[74] E. L. Lehmann. *Theory of Point Estimation*. John Wiley & Sons, Inc., New York, 1983.

[75] Jae S. Lim. *Speech Enhancement*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.

[76] Stephen Linder. *Robust Qualitative and Quantitative Methods for Disturbance Rejection and Fault Accommodation*. PhD thesis, Northeastern University, 1998.

[77] G. S. Maddala and Y. Yin. Outliers, unit roots and robust estimation of nonstationary time series. *Handbook of Statistics: Robust Inference*, 15:237–266, 1997.

[78] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Applications*. John Wiley, West Sussex, England, 2006.

[79] R. A. Maronna and V. J. Yohai. The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995.

[80] R. A. Maronna and R. H. Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.

[81] D. Martin. Robust preprocessing for kalman filtering of glint noise. *IEEE Transactions on Aerospace and Electronic Systems*, AES-23(1):120–128, 1987.

[82] R. D. Martin. Influence functionals for time series. *Annals of Statistics*, 14(3):781–818, 1986.

[83] R.D. Martin, V. J. Yohai, and R. H. Zamar. Min-max bias robust regression. *Annals of Statistics*, 17:1608–1630, 1989.

[84] C. J. Masreliez. Approximate non-gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control*, 20:107–110, 1975.

[85] C. J. Masreliez and R. D. Martin. Robust bayesian estimation for the linear model and robustifying the kalman filter. *IEEE Transactions on Automatic Control*, AC22(3):361–371, 1997.

[86] John Mathews and Kurtis Fink. *Numerical Methods*. Prentice-Hall Pub. Inc., Upper Saddle River, New Jersey, 2004.

[87] Richard J. Meinhold and Nozer D. Singpurwalla. Understanding the kalman filter. *The American Statistician*, 37(2):123–127, 1983.

[88] Jerry Mendel. *Lessons in Digital Estimation Theory*. Prentice-Hall, Inc, Englewood Cliffs, New Jersey, 1987.

[89] L. Mili. Robust filtering and estimation, class notes. Technical report, Virginia Polytechnic Institute and State University, 2005.

[90] L. Mili, M. Cheniae, N. Vichare, and P. Rousseeuw. Robust state estimation based on projection statistics. *IEEE Transactions on Power Systems*, 11(2):11181127, 1996.

[91] L. Mili and C. W. Coakley. Robust estimation in structured linear regression. *Annals of Statistics*, 24(6):25932607, 1996.

[92] R. Miller, E. Carter, and S. Blue. Data assimilation into nonlinear stochastic models. *Tellus*, 51A(2):167–194, 1999.

[93] R. Miller, M. Ghil, and F. Gauthiez. Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of the Atmospheric Sciences*, 51(8):1037–1056, 1994.

[94] D. S. Moore and G. P. McCabe. *Introduction to the Practice of Statistics*. W.H. Freeman and Company, New York, 1989.

[95] C. R. Muirhead. Distinguishing outlier types in time series. *Journal of the Royal Statistical Society (B)*, 48:39–47, 1986.

[96] Eran Naftali and Nicholas Makris. Necessary conditions for a maximum likelihood estimate to become asymptotically unbiased and attain the cramerrao lower bound. *Journal of the Acoustical Society of America*, 110(4):1917–1930, 2001.

[97] K.S. Narendra and S.S. Tripathi. Identification and optimization of aircraft dynamics. *Journal of Aircraft*, 10(4):193–199, 1973.

[98] Shawn Neugebauer. Robust analysis of m-estimators of nonlinear models. Master's thesis, Virginia Polytechnic Institute and State University, 1996.

[99] C. Nicolis and G. Nicolis. Stochastic aspects of climate transitions - additive fluctuations. *Tellus*, 33:225–234, 1981.

[100] Russell J. Niederjohn and James A. Heinen. Understanding speech corrupted by noise. In *Proceedings of IEEE International Conference on Industrial Technology*, pages P1–P5, 1996.

[101] L. Onsager. Reciprocal relations in irreversible processes. *Physics Review*, 37:2265–2279, 1931.

[102] L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Physics Review*, 91:1505–1512, 1953.

[103] K. K. Paliwal and A. Basu. A speech enhancement method based on kalman filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 177–180, 1987.

[104] A. Papavasiliou. *Adaptive Particle Filters with Applications*. PhD thesis, Princeton University, 2002.

[105] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., New York, 1965.

[106] Christopher Pesch. Fast computation of the minimum covariance determinant estimator. Technical Report MIP-9806, Fakultat fur Mathematik und Informatik, Universitat Passau, 1998.

[107] D. Pham. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, 129(5):1194–1207, 1999.

[108] Jean Picard. Efficiency of the extended kalman filter for nonlinear systems with small noise. *SIAM Journal of Applied Mathematics*, 51(3):843–885, 1991.

[109] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–??, 1999.

[110] H. V. Poor. On robust wiener filtering. *IEEE Transactions on Automatic Control*, 25:531–536, 1980.

[111] H. V. Poor. Signal detection in the presence of weakly dependent noise - part ii: robust detection. *IEEE Transactions on Information Theory*, 28:744–752, 1982.

[112] S. Portnoy and X. He. A robust journey in the new millennium. *Journal of the American Statistical Association*, 95(451):1331–1335, 2000.

[113] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[114] L. R. Rabiner and J. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978.

[115] C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley-Interscience, New York, 1973.

[116] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice-Hall, New Jersey, 1996.

[117] H. Risken. *The Fokker-Planck Equation: Methods of Solutions and Applications*. Springer, New York, 1996.

[118] P. J. Rousseeuw. Optimally robust procedures in the infinitesimal sense. In *42nd Session of the ISI*, pages 467–470, 1979.

[119] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[120] P. J. Rousseeuw and C. Croux. The bias of k-step estimators. *Statistics and Probability Letters*, 20:411–420, 1994.

[121] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley, 1987.

[122] P. J. Rousseeuw and E. Ronchetti. Influence curves for general statistics. *Journal of Computational Appl. Math.*, 7:161–166, 1981.

[123] P. J. Rousseeuw and B. C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–651, 1990.

[124] P. J. Rousseeuw and B. C. Van Zomeren. Robust distances: Simulations and cutoff values. *Directions in Robust Statistics and Diagnostics, Part II*, 1991.

[125] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[126] Y. Ruan and P. Willett. Fusion of quantized measurements via particle filtering. *IEEE Aerospace Conference*, page 19671978, 2003.

[127] B. F. La Scala, R. R. Bitmead, and M.R. James. Conditions for stability of the extended kalman filter and their applications to the frequency tracking problem. *Mathematics of Control, Signals, and Systems*, 8:1–26, 1995.

[128] L. L. Scharf. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley, Reading, MA, 1991.

[129] G. Schoenig. *Contributions to Robust Adaptive Signal Processing with Application to Space-Time Adaptive Radar*. PhD thesis, Virginia Polytechnic Institute and State University, 2007.

[130] Fred Schweppe. *Uncertain Dynamic Systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1973.

[131] U. Shaked and Y. Theodor. h_∞-optimal estimation: a tutorial. In *IEEE Conference on Decision and Control*, pages 2278–2286, 1992.

[132] Dan Simon. From here to infinity. *Embedded Systems Programming*, 14(11):20–32, 2001.

[133] Dan Simon. *Optimal State Estimation*. John Wiley, 2006.

[134] S. K. Sinha, C. A. Field, and B. Smith. Robust estimation of nonlinear regression with autoregressive errors. *Statistics and Probability Letters*, 63:49–59, 2003.

[135] H. Sorenson. *Kalman Filtering: Theory and Application*. IEEE Press, 1985.

[136] W. A. Stahel. Breakdown of covariance estimators. Technical Report No. 31, Fachgruppe fur Statistik, ETH, Zurich, 1981.

[137] J. Stewart, K. Sandberg, and B. Pirtle. *Multivariable Calculus: Early Transcendentals*. Brooks/Cole Pub Co, New York, 2003.

[138] S. M Stigler. Simon newcomb, percy daniell, and the history of robust estimation 1885 - 1920. *Journal of the American Statistical Association*, 68:872–879, 1973.

[139] S. M. Stigler. *The History of Statistics*. Belknap Press, Cambridge, 1986.

[140] A. Stromberg and D. Ruppert. Breakdown in nonlinear regression. *Journal of the American Statistical Association*, 87(420), 1992.

[141] D. W. Stroock. *A Concise Introduction to the Theory of Integration*. Birkhäuser, Boston, 1994.

[142] A. Sutera. On stochastic perturbation and long-term climate behavior. *Quarterly Journal of the Royal Meteorological Society*, 137-152, 1981.

[143] O. Talagrand. Assimilation of observations, an introduction. *Journal of Meteorological Society of Japan*, 75:191–209, 1997.

[144] H. L. Van Trees. *Detection, Estimation, and Modulation Theory*. John Wiley & Sons, New York, 1949.

[145] L. Thomas and L. Mili. A robust gm-estimator for the automated detection of external defects on barked hardwood logs and stems. *IEEE Transactions on Signal Processing*, 55(7):3568–3576, 2007.

[146] J. Tsai, Y. Lee, P. Cofie, L. Shieh, and X. Chen. Active fault tolerant control using state-space self-tuning control approach. *International Journal of Systems Science*, 37(11), 2006.

[147] R. S. Tsay. Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7:1–20, 1988.

[148] J. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics, Ed. I*, 1960.

[149] John W. Tukey. *Exploratory Data Analysis*. Addison Wesley, 1977.

[150] S. Vaseghi and P. Rayner. Detection and suppression of impulsive noise in speech communications. In *IEE Proceedings of Communications, Speech, and Vision*, pages 38–46, 1990.

[151] Saeed V. Vaseghi. *Advanced Signal Processing and Digital Noise Reduction*. John Wiley & Sons and B.G. Teubner, Chichester, England and Stuttgart, Germany, 1996.

[152] Q. F. Wei, W. P. Dayawansa, and P. S. Krishnaprasad. h_∞ control for impulsive disturbances: A state-space solution. In *Proceedings of the American Control Conference*, pages 4379–4383, 1995.

[153] D. Williamson. *Digital Control and Implementation*. Prentice-Hall, New Jersey, 1991.

[154] Alan S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12:601–611, 1976.

[155] L. S. Wu, J.R.M. Hosking, and N. Ravishanker. Reallocation outliers in time series. *Applied Statistics*, 42:301–313, 1993.

[156] Y.M. Zhang and J. Jiang. Active fault-tolerant control system against partial actuator failures. *IEE Proceedings Control Theory and Applications*, 149(1):95–104, Jan 2002.