Iterative Memoryless Non-linear Estimators of Correlation for Complex-Valued Gaussian Processes That Exhibit Robustness to Impulsive Noise

Philip M. Tamburello

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
A.A. Beex
Charles Clancy
Jeffrey Reed
Konstantinos Triantis

December 7, 2015
Falls Church, Virginia

Keywords: Robust Estimation, Spectral Estimation, Filtering

Iterative Memoryless Non-linear Estimators of Correlation for Complex-Valued Gaussian Processes That Exhibit Robustness to Impulsive Noise

Philip M. Tamburello

ABSTRACT

The autocorrelation function is a commonly used tool in statistical time series analysis. Under the assumption of Gaussianity, the sample autocorrelation function is the standard method used to estimate this function given a finite number of observations. Non-Gaussian, impulsive observation noise following probability density functions with thick tails, which often occurs in practice, can bias this estimator, rendering classical time series analysis methods ineffective.

This work examines the robustness of two estimators of correlation based on memoryless nonlinear functions of observations, the Phase-Phase Correlator (PPC) and the Median-of-Ratios Estimator (MRE), which are applicable to complex-valued Gaussian random processes. These estimators are very fast and easy to implement in current processors. We show that these estimators are robust from a bias perspective when complex-valued Gaussian processes are contaminated with impulsive noise at the expense of statistical efficiency at the assumed Gaussian distribution. Additionally, iterative versions of these estimators named the IMRE and IPPC are developed, realizing an improved bias performance over their non-iterative counterparts and the well-known robust Schweppe-type Generalized M-estimator utilizing a Huber cost function (SHGM).

An impulsive noise suppression technique is developed using basis pursuit and *a priori* atom weighting derived from the newly developed iterative estimators. This new technique is proposed as an alternative to the robust filter cleaner, a Kalman filter-like approach that relies on linear prediction residuals to identity and replace corrupted observations. It does not have the same initialization issues as the robust filter cleaner.

Robust spectral estimation methods are developed using these new estimators and impulsive noise suppression techniques. Results are obtained for synthetic complex-valued Guassian processes and real-world digital television signals collected using a software defined radio.

Acknowledgments

I would like to thank my advisor Dr. Lamine Mili for encouraging me to apply to the graduate program. This research was shaped by the many hours of discussions during our weekly meetings on Friday nights. The path to completion of a part-time student is much different than that of a full-time student. I appreciate your patience. Additionally, I must thank Professors A.A. Beex, Charles Clancy, Jeffrey Reed, and Konstantinos Triantis for participating on my committee. The time you spent reading my work, and attending my exams was greatly appreciated.

Finally, I would like to thank my family and friends. Thank you for understanding my absence from gatherings over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In general, discrete-time time series analysis attempts to extract the temporal relationship between a collection of observations uniformly spaced in time. This temporal relationship can be used to model the underlying system, commonly resulting in a discrete-time difference equation, relating the output of a system to current and past inputs and outputs. These models are useful when attempting to predict future observations, identify trends or characterizing spectral content.

If the observed random process is a wide-sense stationary, circularly-symmetric, complex-valued Gaussian random process [64], $\{X_t, t = 0, \pm 1 \pm 2, \ldots\}$, it can be completely defined by its first and second-order statistics. That is, given a random vector, $\mathbf{x} = \{X_{t_1}, \ldots, X_{t_n}\}^T$, the n-dimensional multivariate complex-valued normal distribution is defined by the joint probability density function [30] expressed as

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{e^{\left\{-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}}}{\pi^n |\boldsymbol{\Sigma}|}, \tag{1.1}$$

where $\boldsymbol{\mu} = E\{\mathbf{x}\}$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}, \mathbf{x})$ is the covariance matrix. For the time series employed in this work, it is assumed that $\boldsymbol{\mu} = \mathbf{0}$ and the first row of the toeplitz covariance matrix is

the autocovariance function given by

$$\gamma_x(\tau) = E\left\{X_t X_{t+\tau}^*\right\}, \tau = 0, 1, \ldots, N. \tag{1.2}$$

Arguably, the single most important analysis tool for stationary, linear, time-invariant systems is the autocorrelation function, $\rho_x(\tau) = \gamma_x(\tau)/\gamma_x(0)$, which captures the linear relationship of a stationary random process as a function of time offset [9]. The importance of this function is evidenced by its utilization in many statistical signal processing applications. For instance, the Wiener-Khinchin theorem [42] relates the power spectral density of a stationary random process through the Fourier transform of its autocorrelation function. It also has a significant role in optimal linear filters such as the Wiener and Kalman filters [64]. The Yule-Walker equations [9], capture the relationship between the autocorrelation function and the coefficients of an autoregressive random process. Under the Gaussian assumption, optimal linear filters and estimators, in the mean-squared-error sense, will often be a function of the autocorrelation function.

Typically, the autocorrelation function of an observed random process is not known. Therefore, it must be estimated from a finite number of observations, often in the presence of observation noise. The sample autocorrelation function is the estimator of choice due to its simplicity, and under certain assumptions, exhibits very good asymptotic properties. One of these assumptions is that the random process and the observation noise follow a Gaussian probability density function. However, impulsive observation noise characterized by thick-tailed distributions, can severely bias this estimator, frequently invalidating the estimates. In these cases, estimators that are robust to the thick-tailed observation noise must be used.

The subject of this work is to initiate and to analyze four robust alternatives to the sample autocorrelation function, namely the PPC and the MRE and their iterative versions. These estimators of correlation are based on memoryless, nonlinear transforms of complex-valued Gaussian processes that are robust to non-Gaussian observation noise. They are attractive due to their excellent robustness properties to impulsive noise along with their relative

simplicity and computation speed. Indeed, their computational cost grows approximately linearly with the number of observations. Iterative versions of these estimators are developed, sacrificing some speed for better performance from a bias perspective. As a baseline, the performance of these estimators are compared to the well-known robust generalized M-estimators (GM-estimators), which can be used to estimate correlations in linear and nonlinear regression. Finally, we develop an impulsive-noise suppression technique based on these estimators and a weighted multi-bases signal representation. This approach avoids some issues at the edges of data records that occur with some of the more well-known prediction-based data cleaners. Another important application of these methods that is investigated is robust parametric and non-parametric spectral estimation.

1.1 Literature Review

1.1.1 Sources of Impulsive Noise

Thermal noise is an inevitable source of observation noise in electronic systems. It is usually accurately modeled as an additive Gaussian process. In general, the Gaussian distribution has several attractive properties leading to tractable analysis that stems from the stability of the Gaussian distribution and fast processing algorithms using linear operators. Unfortunately, the observed noise in many applications includes other sources, commonly following distributions with thicker tails than those of the Gaussian distribution.

Impulsive noise is a general name given to sporadic noise with sudden short durations and large magnitudes. The time delay between these events can also be random in nature. Consequently, this noise is spectrally flat, or white and is modeled with probability density functions characterized by thick tails. There are many sources of impulsive noise that routinely impact the effectiveness of signal processing applications. These noise sources are related to shot noise associated with discrete events.

Sources of shot noise in wireless communication systems include electric motors, power lines, heavy current switches, fluorescent lights, etc. [62]. Impulsive noise has also been problematic in power line communication systems, elevating segments of the background noise power spectral density by 10 to 15 dB, and in extreme cases as much as 50 dB [78]. Lightning is another source, and is problematic for over-the-horizon radars, especially in regions where large thunderstorms are common, elevating the background noise in range-Doppler maps masking targets [33].

Impulsive noise is not limited to the electromagnetic spectrum. A common source of impulsive noise affecting sonar and underwater telemetry systems in littoral ocean environments is the snapping shrimp [43]. Ice also has a significant role, creating many types of noise. Particularly, acoustic impulsive noise is created when ice is under stress and cracks [58].

1.1.2 Robust Nonlinear Correlation Estimators

In practice, knowledge of the cross- or autocorrelation functions of processes are not known. Therefore, they must be estimated given a finite number of observations from these processes. Classical parametric estimation theory initiated by Fisher lead to methods to estimate the parameters of the assumed statistical model describing the stochastic process, including the correlation coefficient. For computational easiness, the assumed model has been the Gaussian distribution. However, as even Gauss noted, the tails of the Gaussian distribution decay much too rapidly to describe observed data accurately. It is these thick tails and skewness that can cause the linear estimators to suffer statistical inefficiency and from severe biases.

The sample autocorrelation function is a normalized inner product between two $N-\tau$ length vectors, yielding a linear estimator. A well-known fact is that this estimator is very sensitive to outliers defined as data points that deviate from Gaussianity. For example, replacing one observation with a single uncorrelated large magnitude sample will bias the sample-correlation coefficient toward zero. Consequently, conventional time-series analysis based on that linear estimator may produce unacceptable biases. Additionally, as we will see in 6.3.2,

results from eigenanalysis techniques relying on estimates of the autocorrelation matrix, such as principal component analysis, may be distorted because of these outliers [46].

Nonlinear estimators, many of which are based on the early work of Tukey, Huber and Hampel [36, 38, 34, 46], are often employed when the bulk of the observed data follows a Gaussian distribution, and a fraction of the observations, the outliers, follow a non-Gaussian thick-tailed distribution. With the advances in modern computing technology, robust nonlinear estimators are now more accessible to the practicing data analyst. Kassam and Poor [41] authored a survey of robust techniques for signal processing that discusses many nonlinear estimators. The class of M-estimators is one such example. In this case, nonlinearity is introduced by bounding the influence of the impulsive noise by minimizing non-quadratic cost functions of the residuals [46]. For instance, Zou et al. [80], [79] have incorporated M-estimators into robust system identification algorithms. The cost of this robustness is a sacrifice of efficiency at the Gaussian model.

Another method of introducing nonlinearity is to first transform the observed data via some nonlinear function, and then apply the classical linear estimator. Several authors have analyzed the impact nonlinear functions have on estimating the correlation coefficient between Gaussian processes. Hardware limitations, both digital and analog, have motivated much of this research. One such nonlinearity, signal quantization, was analyzed by Widrow [77]. Hard-limiting of real-valued Gaussian signals led to the Bussgang theorem [12].

Researchers in robust statistical signal processing have extended these previous studies and benefitted from the robustness to impulsive noise these nonlinear functions exhibit. Galin and Messer [26], and Ben Mâad et al. [5] have incorporated nonlinear functions into signal processing applications to estimate AR parameters and parity check codes for wireless communication systems respectively. In [63], Santamaria applies a Gaussian kernel to transform observations into a nonlinear feature space. Correlation estimates are then performed and are incorporated into a robust wireless channel equalizer. Chakhchoukh et al. [14] developed a highly robust nonlinear estimation method that is used to forecast the electric power con-

sumption in France. All of the above have realized a benefit over traditional linear estimators when the data deviated from Gaussianity.

One disadvantage that impacted real-time implementation of nonlinear estimators is their increased computational requirements. However, with specialized hardware and the massively parallel computer architectures of today, this obstacle is becoming less of an issue. This is evidenced by active research utilizing robust techniques for signal processing applications in harsh noise environments, such as DARPA's Advanced Speech Encoding (ASE) program [52]. Raytheon BBN Technologies developed an enhanced Mixed Excitation Linear Prediction (MELPe) vocoder that must operate in environments where the noise is not Gaussian, and where conventional encoders produce audio of unacceptable quality.

1.2 Research Objective

Impulsive observation noise can severely degrade the performance of classical time series analysis methods derived for Gaussian random processes, often rendering them ineffective. Linear techniques used to identify and suppress impulsive noise must rely on robust estimators of correlation. For instance, robust Mahalanobis distances or forward and backward prediction residuals can be used to identify observations that do not follow the underlying time series model. The abnormal observations are then replaced with predicted values. Other techniques that do not specifically use the correlation of the desired signal, such as multiple basis signal representation, attempt to separate the desired signal and impulsive noise into multiple orthogonal bases. These techniques can be improved with *a priori* information obtained from correlation-based outlier detection methods to indicate which observations contain the impulsive noise.

The goal of this research is to perform an asymptotic global robustness analysis [34] of two nonlinear estimators described by memoryless envelop distorting filters applied to complex-valued Gaussian random processes. These estimators execute much faster and are more

appropriate for time series analysis than the Schweppe-type GM-estimator using the Huber cost function (SHGM) with projection statistics-based bad leverage point downweighting. Iterative versions of these estimators will be developed, realizing an improved bias performance over the SHGM and faster execution for larger data sets. These robust estimators will be used to add robustness to both parametric and non-parametric spectral estimators. Either of these types of spectral estimators could be used in the presence of impulsive observation noise commonly found in the RF environment.

1.3 Summary of Achievements

This work provides the global robustness analysis of four nonlinear estimators of correlation for complex-valued Gaussian processes. Two of these estimators, the phase-phase correlator and median-of-ratios estimator have been related to the same general result involving memoryless envelop distorting filters. We demonstrate that they are both robust to impulsive observation noise using standard robust-statistics analysis metrics. As part of this analysis, a definition of estimator breakdown for complex-valued time series estimators of correlation in the presence of uncorrelated impulsive noise was established. This definition is applicable to both the isolated replacement or additive outlier model.

Additionally, iterative versions of these robust estimators were developed. These new estimators exhibit improved bias performance over their non-iterative versions and the well-known SHGM estimator. These estimators are faster than the SHGM which relies on projection statistics to identify and downweight both vertical outliers and bad leverage points.

A data cleaning procedure using multiple-bases signal representation was developed that incorporates these new iterative estimators and basis pursuit, an l_1 minimization technique. This technique eliminates the issue encountered by other prediction-based data cleaners when corrupted samples occur near the edge of the observation data record. Data cleaned using this procedure can be processed by standard non-robust algorithms with very little bias.

This research produced two publications [71, 72] in the IEEE Transactions on Signal Processing. Robust parametric and non-paramteric spectral estimators have been developed using these estimators.

1.4 Applications and Results

An application that was investigated that used our new estimators was spectral estimation in the presence of impulsive noise. Both parametric and non-parametric spectral estimators were developed. These techniques were applied to both synthetically generated complex-valued Gaussian processes and collected digital television signals. These techniques were tested for stressing situations, where both strong and weak signals were simultaneously present in the data. Simulation results show very good performance of these techniques for both uncontaminated and contaminated data observations, revealing both the strong and weak signals.

1.5 Dissertation Organization

The remainder of this dissertation is organized as follows. Classical time series analysis, such as the Yule-Walker equations, Levinson recursions and the Burg method, is reviewed in Chapter 2. Classical maximum likelihood estimation is reviewed in Chapter 3. Several outlier models are discussed and their impact on estimating the autocorrelation function is explained. During this discussion, the concept of estimator bias and breakdown point for time series is developed. Robust Huber estimation is also introduced, and we show how to use this method to estimate the autocorrelation function of a time series. In Chapter 4 both the phase-phase correlator and median of ratios estimator for complex-valued Gaussian processes are developed and analyzed. This development is finished with the introduction of iterative versions of both estimators. A natural direct application of the robust estimators to

parametric spectral estimation using the Burg method is introduced in Chapter 5. An impulsive noise suppression technique based on multiple-bases signal representation is improved utilizing *a priori* weighting obtained from the newly developed iterative robust estimators of correlation in Chapter 6. This method of data cleaning is followed by a non-parametric spectral estimator called the multi-taper method, resulting in a robust non-parametric spectral estimator. The new robust spectral estimators are applied to collected signals from the digital television bands in Chapter 7, indicating that these techniques are applicable to signals that are approximately Gaussian. Finally, several interesting future research topics unveiled during this research effort are discussed in Chapter 8.

Chapter 2

Classical Time Series Analysis

2.1 Time Series Analysis for Discrete-Time Data

2.1.1 Discrete-time Difference Equation

The discrete-time difference equation [9] can be used to model an observed time series. If we assume a linear, time-invariant model, then the observed random process, X_t, is the result of a linear combination of current and past inputs and outputs following the equation given by

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \theta_1 U_{t-1} + ... + \theta_q U_{t-q} + U_t. \tag{2.1}$$

This model produces what is often referred to as an autoregressive moving average, ARMA(p,q), process. If U_t and X_t are the input and output of (2.1), then $\boldsymbol{\phi} = [\phi_1, ..., \phi_p]^T$ and $\boldsymbol{\theta} = [\theta_1, ..., \theta_q]^T$ determine the poles and zeros of the system's transfer function. For this system to be stable, that is the output is bounded given a bounded input, the poles of the system must lie inside the unit circle. Often, the input to this model is called the innovations process. In this work it is a complex-valued, zero-mean, white Gaussian random process with finite variance. The output of the model can be interpreted as a filtered version of the

input. By definition, X_t is a Gaussian process since it is assumed to be a linear combination of Gaussian processes. The statistics of X_t follows the probability density function given in (1.1), which highlights the importance of knowing $\gamma_x(\tau)$ or the normalized equivalent, $\rho_x(\tau)$. The zero-mean model is assumed for our analysis and is reflected by the absence of a constant term in (2.1). Additionally, some definitions of the ARMA process also include an additional term, θ_0, which multiplies the innovations process U_t. This term has also been omitted since it is simply a gain term and does not affect the location of the zeros or poles of the ARMA model. To equate a model with this term to (2.1), simply divide the other terms of $\boldsymbol{\theta}$ by θ_0 and absorb θ_0 into the variance of U_t.

The ARMA model is useful in many areas of signal processing. A well-known application of a real-valued version of this model is the compression of speech in voice encoders. This is accomplished by modeling voice as a white noise sequence that is filtered by the vocal tract, which is modeled as an all-pole filter, $\boldsymbol{\theta} = \mathbf{0}$. Specifically, voice encoding can be cast as a spectral estimation problem of an autoregressive process of order-p, AR(p). In this approach, frames of the voice signal are encoded by estimating the pole locations that best fit the spectrum of the observed process X_t.

Often, the observed time series does not strictly follow (2.1). For example, X_t may be the combination of several independent time series, each created from a unique finite difference equation. Another possibility is that some or all of the individual time series may not necessarily follow a Gaussian distribution. However, if each of the combined time series are second-order stationary with a finite variance following Lindeberg's condition [6], then by the central limit theorem, the resulting combination will result in a time series that in the limit follows a Gaussian distribution. An example of this would be the combination of radio signals, some of which could be digitally modulated such as wireless modems, while other might be frequency modulated (FM) such as common analog radio stations. Another modification of (2.1) that is required for most analysis applications is the addition of observation noise. A common source of observation noise is the collection sensor's internally generated thermal noise. This is typically modeled as a white Gaussian process over most observation

bands.

Given these modifications to the original model, it is still possible to model and analyze the observed time series using (2.1) with only one slight modification given by

$$X_{ot} = X_t + E_t. \tag{2.2}$$

Now, a new random process representing the observation noise, E_t, is included. Under ideal conditions, E_t is defined to be a white Gaussian process, uncorrelated to X_t.

2.2 The Role of the Autocorrelation Function in Time Series Analysis

In this section we give a generic definition of the correlation coefficient and give examples of how the autocorrelation function is useful in time series analysis. The Levinson recursions are developed and the Burg method of estimating AR parameters is introduced.

2.2.1 Definition of Correlation Coefficient

The most useful statistical quantity that captures the normalized linear dependence between two complex-valued random processes, X_{1,t_1} and X_{2,t_2}, and allows us to estimate the parameters of (2.1) is the correlation coefficient [9] defined generically as

$$\rho_{x_1 x_2}(t_1, t_2) = \frac{E\left\{\left(X_{1,t} - \bar{X}_{1,t}\right)\left(X_{2,t+\tau} - \bar{X}_{2,t+\tau}\right)^*\right\}}{\sigma_{x_1}\sigma_{x_2}}. \tag{2.3}$$

Here, $E\{\}$ denotes the expected value operator, $\bar{X}_t$ denotes the mean value of X_t, σ_{x_t} is the standard deviation of X_t, and the superscript * indicates the complex conjugate operator. The magnitude of the correlation coefficient for complex-valued random processes is upper

bounded by one, and the angle can take any value from 0 to 2π.

For stationary time series applications, (2.3) becomes

$$\rho_x(\tau) = \frac{E\left\{\left(X_t - \bar{X}_t\right)\left(X_{t+\tau} - \bar{X}_t\right)^*\right\}}{\sigma_x^2}. \tag{2.4}$$

and is commonly called the autocorrelation function. We note a few properties of the autocorrelation function of a stationary time series [9]:

1. The autocorrelation function is one at lag zero, $\rho_x(0) = 1$
2. The autocorrelation function cannot exceed unity, $|\rho_x(\tau)| \leq 1$
3. The autocorrelation function is Hermitian symmetric, $\rho_x(\tau) = \rho_x^*(-\tau)$
4. The autocorrelation function is non-negative definite, $0 \leq \sum_{i,j=1}^{n} a_i^* \rho(t_i - t_j) a_j$
5. For all $\rho_x(\tau)$ there exists a stationary Gaussian time series with $\rho_x(\tau)$ as its autocorrelation function.

2.2.2 Linear Prediction and the Yule-Walker Equations

Often, it is useful to perform a one-step prediction of an observation using past observations, that is we wish to predict X_t given L past observations $X_{t-1}, ..., X_{t-L}$, which we will denote as $\hat{X}_{t|t-1,...,t-L}$. If we constrain the predictor to be a linear operator, we can solve this problem by minimizing with repsect to $\mathbf{w}_f$, the quadratic cost function given by

$$J(\mathbf{w}_f) = E\left\{\left|X_t - \mathbf{x}^T\mathbf{w}_f\right|^2\right\}, \tag{2.5}$$

where $\mathbf{w}_f = \{w_{f_1}, w_{f_2}, \cdots, w_{f_L}\}^T$ represents the forward prediction filter weights and $\mathbf{x} = \{X_{t-1}, ..., X_{t-L}\}^T$. Evaluating the expectation, and setting the complex-valued derivative [56] of $J(\mathbf{w}_f)$ with respect to $\mathbf{w}_f^H$ equal to zero after normalizing by σ_x^2 produces $\mathbf{P}\mathbf{w}_f = \boldsymbol{\rho}_f$,

or in expanded form

$$\begin{bmatrix} \rho_x(0) & \rho_x(1) & \cdots & \rho_x(L-1) \\ \rho_x^*(1) & \rho_x(0) & \cdots & \rho_x(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x^*(L-1) & \rho_x^*(L-2) & \cdots & \rho_x(0) \end{bmatrix} \begin{bmatrix} w_{f_1} \\ w_{f_2} \\ \vdots \\ w_{f_L} \end{bmatrix} = \begin{bmatrix} \rho_x^*(1) \\ \rho_x^*(2) \\ \vdots \\ \rho_x^*(L) \end{bmatrix}. \tag{2.6}$$

$\mathbf{P}$ is the Toeplitz Hermitian autocorrelation matrix and $\boldsymbol{\rho}_f$ is the L-length vector where $\rho_{f_l} = \rho_x^*(l), 1 \leq l \leq L$. At this point, we see that the optimal, least-squares weights are equal to

$$\mathbf{w}_f = \mathbf{P}^{-1}\boldsymbol{\rho}_f. \tag{2.7}$$

It is assumed that $\mathbf{P}$ is invertible because of our assumption of Gaussianity and the definition of the joint probability density function (1.1).

The generic result given by 2.6 is now examined for the purely autoregressive model of order p, AR(p). Consider the discrete-time difference equation describing the system given by

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = U_t. \tag{2.8}$$

We know from the projection theorem [9] that the optimal linear predictor error is orthogonal to the data used to perform the prediction. A simple one-step predictor is shown pictorially in Fig. 2.1 using a Hilbert space vector representation of the random processes X_t and X_{t-1} [9]. In this case we are predicting X_t using X_{t-1}, to get $\hat{X}_{t|t-1}$. The difference between X_t and the linearly predicted estimate $\hat{X}_{t|t-1}$ is perpendicular to the direction of X_{t-1} [48, 74], that is $E\left\{\left(X_t - \hat{X}_{t|t-1}\right) X_{t|t-1}^*\right\} = 0$. In the special case of an AR(p) model, we will define the one-step predictor to be of length p. The resulting predictor error associated with this order-p predictor is $E_t = X_t - \sum_{n=1}^{p} X_{t-n} w_n$. It is apparent that by choosing $w_n = \phi_n, n = 1, \cdots, p$ the error results in $E_t = U_t$ from the definition in (2.8), which has been defined to be white. Therefore, the prediction error is necessarily orthogonal, or uncorrelated, to all of the previous values $X_{t-n}, n = 1, \ldots, p$ used in the prediction. Thus, we can now use this result

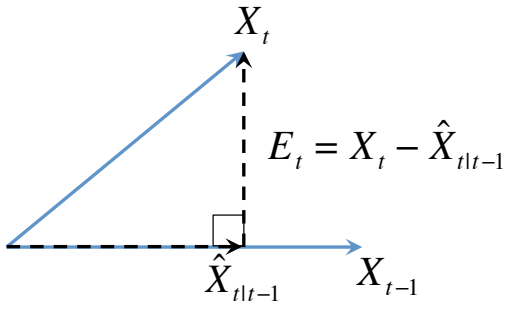

Figure 2.1: Graphical interpretation of orthogonal linear prediction error. The error between X_t and it's estimate using all previous values of X_t is orthogonal. No additional information can be extracted from the previous data to improve the estimate.

with (2.7) to produce the Yule-Walker equations

$$\begin{bmatrix} \rho_x(0) & \rho_x(1) & \cdots & \rho_x(p-1) \\ \rho_x^*(1) & \rho_x(0) & \cdots & \rho_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x^*(p-1) & \rho_x^*(p-2) & \cdots & \rho_x(0) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \rho_x^*(1) \\ \rho_x^*(2) \\ \vdots \\ \rho_x^*(p) \end{bmatrix}, \tag{2.9}$$

which are a set of nonlinear equations that can be solved using a linear system of equations.

Interestingly, if this exercise is repeated for a generic order-L *backward* linear predictor, or a smoothing estimator, where X_t is predicted from future observations $X_{t+1}, \cdots, X_{t+L}$, the optimal linear predictor is found to be a length L linear filter, $\mathbf{w}_b = \left\{w_{f_L}^*, w_{f_{L-1}}^*, \cdots, w_{f_1}^*\right\}^T$, which is the time-reversed and complex-conjugated forward predictor. For an AR(p) model, the optimal linear backward prediction filter is then $\mathbf{w}_b = \left\{\phi_p^*, \phi_{p-1}^*, \cdots, \phi_1^*\right\}^T$. These results will be used later when discussing the Levinson recursions and the Burg method for estimating the parameters of the AR(p) model.

Finally, we investigate the error of the optimal order-L one-step predictor, which leads to another useful set of equations. The expected residual power of the predictor can be expressed

as

$$E\left\{E_t E_t^*\right\} = \sigma_e^2. \tag{2.10}$$

It can be rewritten as

$$E\left\{\left(X_t - \sum_{l=1}^{L} X_{t-l} w_l\right) E_t^*\right\} = \sigma_e^2. \tag{2.11}$$

By the projection theorem, $E\left\{X_{t-l} E_t^*\right\} = 0$, $l = 1, \ldots, L$. Therefore, (2.11) becomes

$$E\left\{\left(X_t X_t^* - X_t \sum_{l=1}^{L} X_{t-l}^* w_l^*\right)\right\} = \sigma_e^2. \tag{2.12}$$

Simplifying produces

$$\sigma_x^2 - X_t \sum_{l=1}^{L} X_{t-l}^* w_l^* = \sigma_e^2, \tag{2.13}$$

and

$$\sigma_x^2 - \sum_{l=1}^{L} \rho_x^*(l) w_l^* = \sigma_e^2, \tag{2.14}$$

which gives us the linear system of equations

$$\begin{bmatrix} \rho_x(0) & \rho_x(1) & \cdots & \rho_x(L) \\ \rho_x^*(1) & \rho_x(0) & \cdots & \rho_x(L-1) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x^*(L) & \rho_x^*(L-1) & \cdots & \rho_x(0) \end{bmatrix} \begin{bmatrix} 1 \\ -w_1 \\ -w_2 \\ \vdots \\ -w_L \end{bmatrix} = \begin{bmatrix} \sigma_e^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.15}$$

For the AR(p) model, the optimal linear predictor error (2.15) is obtained from a length-p

linear predictor and becomes

$$\begin{bmatrix} \rho_x(0) & \rho_x(1) & \cdots & \rho_x(p) \\ \rho_x^*(1) & \rho_x(0) & \cdots & \rho_x(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x^*(p) & \rho_x^*(p-1) & \cdots & \rho_x(0) \end{bmatrix} \begin{bmatrix} 1 \\ -\phi_1 \\ -\phi_2 \\ \vdots \\ -\phi_p \end{bmatrix} = \begin{bmatrix} \sigma_u^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.16}$$

where the minimal prediction error is equal to the variance of the innovations process, σ_u^2. We observe that the variance of the forward and backward linear predictors are the same.

2.2.3 Levinson Recursions

The Yule-Walker equations allow us to compute the optimal one-step predictor coefficients using (2.9). Typically, this requires $O(n^3)$ operations. An interesting perspective can be gained if we examine how to create one-step predictors of increasing length, which yields a recursive algorithm known as the Levinson recursions. An additional benefit is a reduced number of computations and increased numerical stability over standard matrix inversion techniques. Suppose we wish to predict X_t using an order-L linear one-step forward predictor. Instead of solving (2.9) directly, we start with a predictor of order one, and would like to successively update the predictor $L-1$ more times to reach the goal of an order-L predictor.

Simple manipulation of (2.9) will reveal the recursive structure of the Yule-Walker equations. Before this development, we will first define notation to make the derivation easier. First, the $L \times L$ autocorrelation matrix in (2.9) is a Hermitian-symmetric Toeplitz matrix containing the autocorrelation coefficients $\rho_x(\tau)$ corresponding to lags $\tau = 0, \cdots, L-1$. This matrix will be written as $\mathbf{P}^{(L)}$. Related to this matrix is the vector $\boldsymbol{\rho}_f^{(L)}$. It is a lag-shifted version of the first column of $\mathbf{P}^{(L)}$. That is, $\boldsymbol{\rho}_f^{(L)} = \{\rho_x^*(1), \rho_x^*(2), \cdots, \rho_x^*(L)\}^T$. Time reversing and conjugating $\boldsymbol{\rho}_f^{(L)}$ gives us $\boldsymbol{\rho}_b^{(L)} = \{\rho_x(L), \rho(L-1), \cdots, \rho_x(1)\}^T$. We will represent the order-L forward and backward linear predictor weight vectors using $\mathbf{w}_f^{(L)}$ and $\mathbf{w}_b^{(L)}$, where

$\mathbf{w}_b^{(L)} = \left\{ w_{f_L}^{*(L)}, w_{f_{L-1}}^{*(L)}, \cdots, w_{f_1}^{*(L)} \right\}$.

First, treating (2.9) as a set of L equations, we subtract the last term from both sides of each equation. This allows us to capture the Yule-Walker relationships with one $(L-1) \times (L-1)$ system of equations,

$$\mathbf{P}^{(L-1)} \begin{bmatrix} w_{f_1}^{(L)} \\ w_{f_2}^{(L)} \\ \vdots \\ w_{f_{L-1}}^{(L)} \end{bmatrix} = \boldsymbol{\rho}_f^{(L-1)} - w_{f_L}^{(L)} \boldsymbol{\rho}_b^{(L-1)}, \tag{2.17}$$

and an additional equation containing $L-1$ unknowns,

$$\boldsymbol{\rho}_b^{(L-1)^H} \begin{bmatrix} w_{f_1}^{(L)} \\ w_{f_2}^{(L)} \\ \vdots \\ w_{f_{L-1}}^{(L)} \end{bmatrix} = \rho_x^*(L) - w_{f_L}^{(L)} \rho_x(0). \tag{2.18}$$

Multiplying (2.17) by the inverse of $\mathbf{P}^{(L-1)}$ results in

$$\begin{bmatrix} w_{f_1}^{(L)} \\ w_{f_2}^{(L)} \\ \vdots \\ w_{f_{L-1}}^{(L)} \end{bmatrix} = \mathbf{P}^{(L-1)^{-1}} \boldsymbol{\rho}_f^{(L-1)} - w_{f_L}^{(L)} \mathbf{P}^{(L-1)^{-1}} \boldsymbol{\rho}_b^{(L-1)} \tag{2.19}$$

Recognizing that the matrix-vector operations in the right-hand side of (2.19) are the forward and reverse order-$L-1$ linear predictors, we can rewrite (2.19) as

$$\begin{bmatrix} w_{f_1}^{(L)} \\ w_{f_2}^{(L)} \\ \vdots \\ w_{f_{L-1}}^{(L)} \end{bmatrix} = \mathbf{w}_f^{(L-1)} - w_{f_L}^{(L)} \mathbf{w}_b^{(L-1)} \tag{2.20}$$

Substituting (2.20) into (2.18) and recognizing that $\rho_x(0) = 1$ provides us with the following relationship

$$w_{f_L}^{(L)} = \frac{\rho_x^*(L) - \boldsymbol{\rho}_b^{(L-1)^H} \mathbf{w}_f^{(L-1)}}{1 - \boldsymbol{\rho}_b^{(L-1)^H} \mathbf{w}_b^{(L-1)}} \tag{2.21}$$

With knowledge of the autocorrelation function, one can iteratively calculate the weights of an order-L one-step prediction filter. For the special case of a pure AR(p) process, the $L = p$ weights obtained from the Levinson recursions are in fact the coefficients of the AR(p) process, that is $\mathbf{w}_f^{(p)} = [\phi_1, ..., \phi_p]^T$. The computational complexity of this algorithm is proportional to $O(n^2)$.

We can interpret (2.20) as a method to update the weights using the previous order weights. The term $w_{f_L}^{(L)}$ can be thought of as the update step and is often called the reflection coefficient. The reflection coefficient, which we will denote as κ_L, is critical to a very important algorithm called the Burg method [11], which is a very stable method to estimate the AR(p) coefficients, exhibiting very good properties when the coefficients are close to the unit circle. Other algorithms often produce inaccurate results in these cases [10]. Therefore, a better understanding of this reflection coefficient is warranted. We will see that it is related to the partial correlation coefficient.

2.2.4 Partial Correlation Coefficient

Another important statistical quantity used in time series analysis is the partial correlation coefficient. This quantity captures the linear dependence between X_t and X_{t-L} after first removing all correlation between X_t and $X_{t-L+i}, t < i < L$, and second the correlation between X_{t-L+i} and $X_t, t < i < L$. Another useful interpretation can be made of this quantity. In fact, the partial correlation coefficient is the correlation coefficient between the residual error resulting from the forward and backward linear predictions of X_t and X_{t-L} respectively using an order-$(L-1)$ linear predictor. The partial correlation coefficient will be denoted $\rho_{X_t X_{t-L} \cdot \{X_{t-1}, \cdots, X_{t-L+1}\}}$.

At this point we formally define the new random processes associated with the forward and backward order-$(L-1)$ linear prediction errors,

$$E^f_{t,L-1} = X_t - \hat{X}_{t|t-1,\ldots,t-L+1}, \tag{2.22}$$

and

$$E^b_{t,L-1} = X_{t-L} - \hat{X}_{t-L|t-1,\ldots,t-L+1}. \tag{2.23}$$

The correlation between these processes is written as

$$\frac{E\left\{E^f_{t,L-1}\left(E^b_{t,L-1}\right)^*\right\}}{\sigma_{e_f}\sigma_{e_b}} = \frac{E\left\{\left(X_t - \mathbf{x}^T\mathbf{w}_f^{(L-1)}\right)\left(X_{t-L} - \mathbf{x}^T\mathbf{w}_b^{(L-1)}\right)^H\right\}}{\sigma_{e_f}\sigma_{e_b}}, \tag{2.24}$$

where $\mathbf{x} = \{X_{t-1}, X_{t-2}, \ldots, X_{t-L+1}\}^T$.

Expanding, we have

$$\frac{E\left\{E^f_{t,L-1}\left(E^b_{t,L-1}\right)^*\right\}}{\sigma_{e_f}\sigma_{e_b}} = \frac{\sigma_x^2\left[\rho_x^*(L) - \mathbf{w}_b^{(L-1)^H}\boldsymbol{\rho}_f^{(L-1)} - \boldsymbol{\rho}_b^{(L-1)^H}\mathbf{w}_f^{(L-1)} + \mathbf{w}_b^{(L-1)^H}\mathbf{P}^{(L-1)}\mathbf{w}_f^{(L-1)}\right]}{\sigma_{e_f}\sigma_{e_b}}. \tag{2.25}$$

We recognize that $\sigma_{e_f} = \sigma_{e_b}$ and that $\mathbf{w}_f = \mathbf{P}^{-1}\boldsymbol{\rho}_f$. Therefore,

$$\frac{E\left\{E^f_{t,L-1}\left(E^b_{t,L-1}\right)^*\right\}}{\sigma_{e_f}\sigma_{e_b}} = \frac{\sigma_x^2\left[\rho_x^*(L) - \boldsymbol{\rho}_b^{(L-1)^H}\mathbf{w}_f^{(L-1)}\right]}{\sigma_{e_b}^2}, \tag{2.26}$$

which simplifies to

$$\frac{E\left\{E^f_{t,L-1}\left(E^b_{t,L-1}\right)^*\right\}}{\sigma_{e_f}\sigma_{e_b}} = \frac{\rho_x^*(L) - \boldsymbol{\rho}_b^{(L-1)^H}\mathbf{w}_f^{(L-1)}}{1 - \boldsymbol{\rho}_b^{(L-1)^H}\mathbf{w}_b^{(L-1)}}. \tag{2.27}$$

We observe that (2.27) is equivalent to (2.21), which means that the partial correlation coefficient is equivalent to the reflection coefficient. Note that some derivations result with the reflection coefficient being equal to the negative of the partial correlation coefficient.

Obviously this notation will impact the derivation of the Burg method developed next in Section 2.2.5.

2.2.5 The Burg Method

Estimating the unknown parameters of an autoregressive process is a common occurrence in time series analysis. As mentioned previously, there are certain situations when the method proposed by Burg is preferred.

Burg proposed to estimate the AR parameters by using the reflection coefficients and the Levinson recursions. It is assumed that the estimates of the previous order, forward and backward, linear one-step prediction filter coefficients exist. Therefore, the updated weights are given by

$$\begin{bmatrix} \hat{w}_{f_1}^{(L)} \\ \hat{w}_{f_2}^{(L)} \\ \vdots \\ \hat{w}_{f_{L-1}}^{(L)} \end{bmatrix} = \hat{\mathbf{w}}_f^{(L-1)} - \hat{\kappa}_L \mathbf{w}_b^{(L-1)}, \tag{2.28}$$

and $\hat{w}_L^{(L)} = \hat{\kappa}_L$.

In order to perform this update, the reflection coefficient must be estimated for this iteration. To perform this estimation, Burg proposed to jointly minimize the power of the forward and reverse prediction errors (2.22) and (2.23) with respect to the reflection coefficient. This is accomplished by first taking the complex-valued derivative [67] of the average of the expected forward and backward error with respect to the complex-conjugate of the reflection coefficient. Then, after equating to zero we solve

$$\frac{\partial}{\partial \kappa_L^*} \frac{1}{2} \left(|E_{t,L}^f|^2 + |E_{t,L}^b|^2 \right) = 0, \tag{2.29}$$

for the reflection coefficient. Substituting (2.20) into (2.22) and (2.23) reveals the recursive

formulas for the forward and backward prediction errors,

$$E^f_{t,L} = E^f_{t,L-1} - \kappa_L E^b_{t-1,L-1}, \tag{2.30}$$

and

$$E^b_{t,L} = E^b_{t-1,L-1} - \kappa^*_L E^f_{t,L-1}, \tag{2.31}$$

where $E^f_{t,0} = E^b_{t,0} = X_t$.

Inserting (2.30) and (2.31) into (2.29) produces the relationship between the L^{th} reflection coefficient and the prediction errors associated with order-$L-1$ which is given by,

$$\kappa_L = \frac{2E\{E^f_{t,L-1}E^{b^*}_{t-1,L-1}\}}{E\{|E^f_{t,L-1}|^2\} + E\{|E^b_{t-1,L-1}|^2\}}. \tag{2.32}$$

To implement the Burg method, the above quantities are replaced with estimates based on the sample average estimators. This will be revisited later, in the parametric spectral estimation section.

2.2.6 Wiener-Khinchin Theorem

An important result from stochastic process theory relates the autocorrelation function to the power spectral density. The power spectral density can be thought of as the power occupied by the random process as a function of frequency. Formally, the power spectral density is defined in terms the autocorrelation function as

$$P_x(f) = \sigma^2_x \sum_{k=-\infty}^{k=\infty} \rho_x[k]e^{-j2\pi fk} \qquad -\frac{1}{2} \leq f \leq \frac{1}{2}, \tag{2.33}$$

assuming the autocorrelation function decays sufficiently quickly. This result is known as the Wiener-Khinchin theorem. The normalized power spectral density is just (2.33) divided by the variance, σ^2_x, of the random process.

Insight is provided by (2.33) into the relationship between the autocorrelation function and the power spectral density. Essentially, the power spectral density is the Fourier transform of the autocorrelation function. For a random process to have a non-flat power spectral density, its autocorrelation function must be non-zero for some values of $k \neq 0$.

2.2.7 Gaussian Maximum Likelihood Estimator of the Correlation Coefficient

The previous sections provide insight into the importance of the correlation coefficient, and the partial correlation coefficient, for time series analysis. Estimating the unknown parameters of an observed process requires us to estimate the correlation coefficient, or partial correlation, from a finite number of observations.

Given N observations of two zero-mean Gaussian random processes, $X_{1,t}$ and $X_{2,t}$, the Gaussian Maximum Likelihood Estimate (MLE) of the correlation coefficient is called the sample correlation coefficient and is given by

$$\hat{\rho}_{x_1 x_{2_{\mathrm{MLE}}}} = \frac{\sum_{n=1}^{N} x_{1,n} x_{2,n}^{*}}{\left(\sum_{n=1}^{N} |x_{1,n}|^2 \sum_{n=1}^{N} |x_{2,n}|^2\right)^{1/2}}. \tag{2.34}$$

Under general conditions, it is asymptotically unbiased, Fisher consistent and efficient [76]. We must emphasize that the random variables $X_{1,t}$ and $X_{2,t}$ are independent but correlated in time sequences.

For time series analysis, (2.34) can be modified and used to estimate the correlation coefficient as a function of time separation for a single random process. This results in the sample autocorrelation function defined by

$$\hat{\rho}_x(\tau) = \frac{\sum_{n=1}^{N-\tau} x_n x_{n+\tau}^{*}}{\left(\sum_{n=1}^{N-\tau} |x_n|^2 \sum_{n=\tau+1}^{N} |x_{n+\tau}|^2\right)^{1/2}}. \tag{2.35}$$

It is important to understand that (2.35) is not necessarily the MLE for the autocorrelation coefficient. An explanation of why this is so will be discussed later in Section 3.4.1 (it will be the MLE for our test example). However, due to its simplicity and favorable properties in most applications, it is often the choice for estimating the autocorrelation function. Given Gaussian observations, (2.35) is asymptotically unbiased due to the law of large numbers, consistent, and its variance is given by the well-known Bartlett's formula [9, 44].

Chapter 3

Maximum Likelihood and Robust Huber Estimation

In this section we will describe the maximum likelihood estimation approach, and the desirable asymptotic properties of the resulting estimators. We then describe the well-known family of Maximimum Likelihood-type estimators (M-estimators) developed by Huber, and a generalization of this class, called the generalized M-estimators (GM-estimators). These estimators are useful when the observed random process deviates from the assumed model. This can occur when the observations are corrupted by outliers.

Robustness is obtained by GM-estimators by identifying and bounding the influence of outliers through the use of non-quadratic cost functions of residuals and proper suppression of bad leverage points. This formulation results in estimators that are a nonlinear function of the observations, a requirement for robustness when the observations deviate from Gaussianity.

3.1 Definition of Stationarity

Before discussing the theory and properties of various estimators, we first describe a necessary condition of the random processes on which the estimators operate. In order to produce estimates of the parameters of an observed random process, we must require that the process is stationary.

A random process is said to be wide-sense, or autocovariance stationary if its first and second order statistics exist, and do not change over time. Quite often, wide-sense stationarity is implied when discussing the stationarity of a Gaussian random process since it can be completely defined by the first and second order statistics. A more strict and general requirement is complete stationarity. This implies that all of the higher order statistics of the random process (not necessarily Gaussian) do not change over time. For a Gaussian time series, X_t, the requirements for wide-sense stationarity are listed in Table 3.1.

Table 3.1: Stationarity Requirements for Gaussian Processes

1. $E\{X_t\} = E\{X_{t+\tau}\} = \mu < \infty,$
2. $E\{X_t X_t^*\} = \sigma_x^2 < \infty,$
3. $E\{X_t X_{t+\tau}^*\} = \gamma(\tau), \forall t$

The ARMA model defined by 2.1 must be stable for X_t to be finite variance. This requirement forces the poles of the ARMA model to lie inside the unit circle in the complex plane. Additonally, for stationarity, all of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are assumed fixed. This condition is often violated in real-world data sets over longer periods of time. However, approximate stationarity is typically obtained over short time intervals. In this work, it will be assumed that all observations used for analysis are obtained from a wide-sense stationary complex-valued Gaussian random process.

3.2 Definition of Maximum Likelihood Estimators

Maximum likelihood estimation is the benchmark for classic estimation theory. We briefly discuss the principles behind maximum likelihood estimator. We start by making observations contained in $\mathbf{x}_i$. It is assumed that the distribution, $f(\mathbf{x};\theta)$, is known, for instance Gaussian. However, the parameter, θ, of the distribution is unknown and must be estimated from the observations. This is accomplished by finding the parameter of the distribution that is most likely based on the observed data. To this end, we introduce the likelihood function defined as

$$L(\theta;\mathbf{x}) = cf_{\mathbf{x}}(\mathbf{x};\theta), \tag{3.1}$$

where c is a constant and usually set to one. It is a function of θ given $\mathbf{x}$, our observations. It is also the conditional probability density function given the fixed, deterministic parameter θ. The objective is to maximize the likelihood function with respect to the parameter θ. This is accomplished by setting the derivative of that function with respect to θ equal to zero and solving for θ, yielding the maximum likelihood estimator. Formally we have

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta;\mathbf{x}). \tag{3.2}$$

We note that when the observations are independent and identically distributed, the likelihood function becomes a product of densities $L(\theta;\mathbf{x}) = c_i\Pi_{i=1}^{N} f_x(x_i;\theta)$. If the observations are correlated, one simply uses the joint probability density function. Additionally, since the natural logarithm is monotonically increasing, one can maximize the log of the likelihood function. This is often much easier than maximizing the likelihood function.

3.3 Desirable Properties of Estimators

We now discuss several desirable properties of estimators.

3.3.1 Fisher Consistency

To understand Fisher consistency, we first discuss the idea of the empirical distribution function. Given n observations $\mathbf{x} = \{x_1, \cdots, x_n\}^T$, the empirical cummulative probability distribution function is given by

$$F_n(u) = \frac{1}{n} \sum_{i=1}^{n} \Delta(u - x_i), \tag{3.3}$$

where

$$\Delta(u - x_i) = \begin{cases} 0 & u < x_i, \\ 1 & u \leq x_i. \end{cases} \tag{3.4}$$

In the limit as $n \to \infty$, $F_n(u) \to F(u)$ with probability one according to the Glivenko-Cantelli theorem for both independent or correlated observations [54, 73]. An estimator is said to be Fisher consistent if $T_n(F_n) = T(F) = \theta$ as $n \to \infty$ where $T(F)$ is the asymptotic functional form of the estimator. In other words, if the estimator operates on the entire sample population, then to be Fisher consistent, the estimator must produce the true parameter.

3.3.2 Unbiasedness

We define the bias of our estimator, T, operating on a distribution F as the difference between the expected value of the estimator and the true value of the parameter θ being estimated, that is

$$b = E\{T(F)\} - \theta. \tag{3.5}$$

Suppose our estimator performs m estimates using m distinct observation records of length n where both m and n are large. We define,

$$b_m = E\{T(F_n)\} - \theta, \tag{3.6}$$

to be the bias for each estimate. If $E\{b_m\} = 0$, than the estimator is said to be unbiased, or mean unbiased. In other words, on average, an unbiased estimator produces the correct value when operating on a large number of observations.

3.3.3 Efficiency

Efficiency is another measure of estimator performance. It quantifies how well the estimator uses the available information contained in the random process to produce the estimate. The measure of utilization is the resulting variance of the estimator. Estimators with lower variances are more efficient.

The variance of an unbiased estimator is theoretically lower bounded by the Cramer-Rao lower bound. Or, alternatively, lower bounded by the reciprocal of the Fisher information, $I(\theta)$. The Fisher information is the amount of information contained in a random process about the parameters of the defining model (our parameters to be estimated).

Estimator efficiency is defined by

$$\eta = \frac{I(\theta)^{-1}}{\mathrm{VAR}(T)}. \tag{3.7}$$

If a maximum likelihood estimator exists, its variance is known to be the lowest possible, that is its variance is equal to the Cramer-Rao lower bound. Therefore, all MLE have an efficiency of one.

Often, we wish to compare estimators using the efficiency metric. We call this the relative

efficiency of one estimator with respect to another. Relative efficiency is defined by

$$\eta_{\mathrm{A,B}} = \frac{\mathrm{VAR}(T_B)}{\mathrm{VAR}(T_A)}. \tag{3.8}$$

A relative efficiency that is less than one indicates that T_B is better from a variance perspective than T_A.

3.3.4 Consistency

It is desirable, if not expected, for an estimator to yield better results if more observations are used. In the limit, as the number of observations increases, a consistent estimator converges in probability to the true value with probability equal to one,

$$\operatorname*{plim}_{n\to\infty} T_n = \theta \tag{3.9}$$

This can be thought of as the variance of the estimator becoming smaller (tending to zero) as the estimator uses more observations.

3.3.5 Asymptotic Properties of Maximum Likelihood Estimators

If a maximum likelihood estimator (MLE) exists for the parameter of interest, asymptotically, it is efficient and consistent. This means that no other estimator outperforms it with respect to these measures of performance. Additionally, if the random process is a member of the single parameter exponential family (Gaussian, exponential, chi-squared, etc.), then the MLE is also asymptotically unbiased (assuming the observations are independently and identically distributed). Because of these nice properties, MLE estimators are often used in practice.

3.4 The Gaussian Maximum Likelihood Estimator of the Correlation Coefficient

Maximum likelihood estimation given independent Gaussian observations usually results in simple linear operators. We will see that temporal correlation (sometimes called serial correlation) in time series analysis can complicate the MLE of the correlation coefficient. This may result in difficult optimization problems requiring numerical solvers.

In this section, we will develop the maximum likelihood estimator for the solution to the linear regression problem assuming that the errors are complex-valued uncorrelated Gaussian random variables. For complex-valued stochastic signals, proper problem formulation will yield the correlation coefficient. The case where this result is equal to the autocorrelation coefficient with be highlighted. Asymptotically, both forms of this estimator and the sample autocorrelation coefficient are equal. The usefulness of this effort will be seen when we develop the GM-estimator, which is based on the linear regression formulation.

3.4.1 Definition and Linear Regression Form

Let us first consider a multiple linear regression model defined by

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{3.10}$$

where $\mathbf{H}$ is an $(m \times n)$ matrix containing the explanatory variables, and $\mathbf{z}$ and $\mathbf{e}$ are the m-dimensional response and observation noise process vectors. The $(n \times 1)$ vector $\mathbf{x}$ contains the parameters that we wish to estimate. The residuals are defined as

$$r_i = z_i - \hat{z}_i = z_i - \mathbf{h}_i^T \hat{\mathbf{x}}, \tag{3.11}$$

where $\mathbf{h}_i$ is the i^{th} row of $\mathbf{H}^T$ and $\hat{\mathbf{x}}$ is the estimate of the regression coefficient vector. We will assume that the observation noise process in (3.10), $\mathbf{e}$, is a zero-mean complex-valued Gaussian process with covariance matrix $2\sigma_e^2\mathbf{I}$ and $\mathbf{H}$ is full rank. This leads to the conditional Gaussian probability distribution function given by

$$p\left(\mathbf{z};\mathbf{x}\right) = \frac{1}{\pi^N\left(2\sigma_e^2\right)^N} e^{\left[-\frac{1}{2\sigma_e^2}(\mathbf{z}-\mathbf{H}\mathbf{x})^H(\mathbf{z}-\mathbf{H}\mathbf{x})\right]}, \tag{3.12}$$

where the superscript ^{H} indicates the complex transpose, or the Hermitian operator. We recognize (3.12) as the conditional likelihood equation $L(\mathbf{x};\mathbf{z})$.

The Gaussian MLE, $\hat{\mathbf{x}}_{\text{MLE}}$, is obtained from the solution that maximizes (3.12), or alternatively the natural log of (3.12), with respect to $\mathbf{x}$. After taking the natural log, and discarding all terms that are not a function of $\mathbf{x}$ produces the cost function that must be minimized given by

$$J\left(\mathbf{x}\right) = \frac{1}{2}\left(\mathbf{z}-\mathbf{H}\mathbf{x}\right)^H\left(\mathbf{z}-\mathbf{H}\mathbf{x}\right). \tag{3.13}$$

Differentiating with respect to the vector, $\mathbf{x}$ and equating to zero produces the well-known least-squares solution,

$$\hat{\mathbf{x}}_{\text{MLE}} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{x}, \tag{3.14}$$

and it is the MLE for the linear regression model described by (3.10) given our assumptions.

It is worth making a few comments at this point. First, the cost function defined by (3.13) can be rewritten in terms of the residuals as

$$J\left(\mathbf{x}\right) = \frac{1}{2}\mathbf{r}^H\mathbf{r}. \tag{3.15}$$

Thus, the Gaussian conditional MLE for the regression model is the "least-squares", which minimizes the sum of the square of the residuals.

Additionally, the least-squares estimator defined by (3.14) is not the MLE if the observation error is not zero-mean Gaussian with covariance matrix $2\sigma_e^2\mathbf{I}$. Finally, the least-squares estimator is not necessarily the MLE if $\mathbf{H}$ is a random matrix [42].

The generic problem of estimating the normalized autocorrelation function for time series can be cast as the solution to the general linear regression equations. To compute the lag-one autocorrelation coefficient we solve,

$$\mathbf{x}_2 = \mathbf{x}_1\rho + \mathbf{e}. \tag{3.16}$$

Comparing (3.16) with (3.10) we infer that the explanatory and response variables are now $\mathbf{x}_1 = [x_{t+\tau}, x_{t+\tau-1}, ..., x_{t+\tau-N+1}]^T$ and $\mathbf{x}_2 = [x_t, x_{t-1}, ...x_{t-N+1}]^T$, respectively. We notice that observations of the time series x_t can show up in both the explanatory and response variables since they are both time-shifted subsets of x_t. Now, (3.14) can be used to estimate the autocorrelation coefficient. If we expand the matrix-vector math, we see that

$$\hat{\rho}_{\text{MLE}} = \frac{\sum_{n=1}^{N-\tau} x_n x_{n+\tau}^*}{\sum_{n=1}^{N-\tau} x_n^* x_n}. \tag{3.17}$$

In the limit, as N tends to infinity, (3.17) and (2.35) are asymptotically equivalent.

In general, applying (3.10) to time series analysis violates our assumption of uncorrelated observations errors. However, it is often done in practice due to the simple form of the least-squares solution. Additionally, constructing the MLE of correlation for correlated time series analysis must account for this temporal correlation in X_t, not necessarily in the additive noise E_t which can be white. Often the MLE solution is quite involved, requiring numerical solvers. Another issue is that creating the likelihood function requires one to know the exact order of the ARMA process, something generally not known *a priori*.

Fortunately, (3.16) is the exact form of an autoregressive system of order one. The finite variance, uncorrelated error term equates to the innovations process. This means that the sample autocorrelation function is the Gaussian MLE for the correlation coefficient of a

Gaussian AR(1) system. We will use this model to assess our new estimators from a bias and variance perspective. This allows us to compare the performance of the new estimators to the MLE (the sample autocorrelation function) in a time series application without the added complication of implementing a non-trivial MLE.

3.4.2 Asymptotic Properties

Before moving on to discussions about other estimators, it is useful to discuss some of the important asymptotic properties of the sample autocorrelation function for our test case, the AR(1) model. For analysis purposes, we will estimate the lag-one autocorrelation coefficient of the AR(1) model, which is equivalent to estimating the reflection coefficient. For this test case then, $\phi_1 = \rho(1)$. Throughout this analysis, will will denote this estimator as T_{MLE}, with the understanding that $T_{\text{MLE}} = \hat{\phi} = \hat{\rho}(1)$.

Under our assumptions, the sample autocorrelation coefficient is unbiased. Asymptotically, $|E\{T_{\text{MLE}}\} - \rho| = 0$, meaning that the expected value of the MLE for our example is the true value of the reflection coefficient ϕ. The asymptotic variance of the MLE estimator for our example is given by Bartlett's formula [9] and is

$$\text{var}\left(\sqrt{N}|T_{\text{MLE}} - \rho|\right) = 1 - |\rho|^2, \tag{3.18}$$

where N is the number of observations used by the estimator.

Examining (3.18) we observe that for the AR(1) model, the variance is dependent on the true value of the magnitude of the lag-one correlation coefficient. This is demonstrated in Figure 3.1 where the variance of the MLE is plotted as a function of the magnitude of the correlation coefficient.

Due to the properties of the MLE, Figure 3.1 displays the minimum achievable variance of all unbiased estimators of the AR(1) reflection coefficient of a Gaussian process. We may accept greater variance at Gaussianity for other unbiased estimators. We will use this variance result

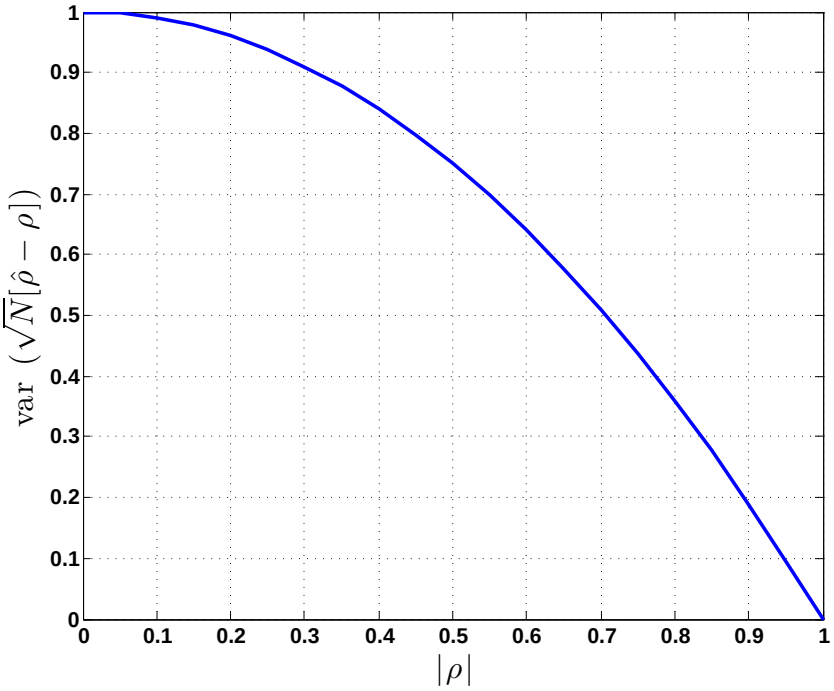

Figure 3.1: The variance of the sample correlation coefficient estimator for an AR(1) time series model is a function of the magnitude of the lag-one correlation coefficient given by Bartlett's formua. The variance decreases as the correlation increases, approach zero when the data is completely correlated.

later when we compare the relative efficiency of our proposed robust estimators.

3.4.3 Robustness to Non-Gaussian Observations

The major disadvantage of the sample autocorrelation function is that it is not robust against outliers. A well-known fact is that when the underlying distributions deviate from Gaussianity, (2.4) can be very sensitive to outliers. In fact, a single outlier, can render this estimator useless. An outlier is defined as a data point that does not follow the main pattern of the bulk of the data points. In particular, in this analysis, the outliers deviate from Gaussianity. Examination of (2.34) reveals that in an extreme case, a single observation with a large magnitude can drive the sample autocorrelation function to zero.

For the general linear regression model given by (3.10), outliers can occur in both $\mathbf{z}$ and $\mathbf{H}$. Outliers in z_i and not in $\mathbf{h}_i^T$ are called vertical outliers and can adversely affect the residuals, r_i. Outliers in $\mathbf{h}_i^T$ and not in z_i and not z_i create *good* leverage points while outliers in both $\mathbf{h}_i^T$ and z_i create *bad* leverage points. In other words, a good leverage point fits the trend of the data and can possibly improve the precision of an estimator [46]. On the other hand, a bad leverage point occurs when the projection of the data point into the factor space is away from the bulk of the data and does not fit the trend, producing deleterious effects to the general class of the M-estimators.

As a consequence, applying the Gaussian MLE to our observations in the presence of outliers produces severely biased results. Several naturally occurring sources of outliers that are common in wireless communication systems are electrical motors, power lines, heavy current switches, fluorescent lights, etc [62]. These sources cause what is referred to as impulsive noise and are characterized by probability density functions with thick tails typical of non-Gaussian noise. In this analysis, we define impulsive noise as isolated, random-phase instantaneous magnitude spikes in the observed signal. Additionally, we assume the impulsive-noise process is white, similar to the examples in [51] [1] and [29]. In some cases, the magnitude of these outliers may be relatively small compared to the valid data.

As the Gaussian MLE and least-squares estimators of the correlation coefficient are asymptotically equivalent, we will show why this estimator is not robust. We will also show how to gain robustness by using non-quadratic cost-functions of the residuals.

3.5 Outliers in Time Series Analysis

Outliers in time series analysis are abnormal observations, deviating from the so-called core process. Following Huber [38], we use the epsilon contaminated model [46] represented by

$$G = (1 - \epsilon)F + \epsilon H. \tag{3.19}$$

The contaminated distribution, G, consists of the assumed distribution of the core process, F, and the thick-tailed outlier distribution, H, which is unknown. In the robust time series literature, three types of outliers are defined, namely the *additve*, *replacement*, and *innovative* outlier models [25, 60]. Typically, these outliers are modeled as short-duration disturbances or single-sample impulses with uncorrelated random magnitude and phase while following a probability distribution that is independent of the core process. It is common to model the outlier process as a white Gaussian process with a larger variance than the core process, which can simplify the analysis. Outliers can be modeled as *isolated* or *patchy* events. The different models and outlier distributions affect the sample autocorrelation function differently. We briefly review the three types of outliers discussed in [46] and illustrate the worst case effects that they can have on the sample autocorrelation function.

3.5.1 Additive and Replacement Outliers

An ε-contaminated time series containing additive or replacement outliers [46] follows the general process:

$$X_{\varepsilon t} = X_t(1 - Z_t) + Z_t W_t. \tag{3.20}$$

The contamination rate is determined by the zero-one process, Z_t, defined by $P(Z_t = 1) = \varepsilon$. Replacement outliers are modeled by defining W_t as a noise process that is uncorrelated with X_t. By contrast, additive outliers are modeled by defining $W_t = X_t + V_t$, where V_t is the noise distribution. In this model, each outlier is contained to only one sample. When the magnitude of the outliers are very large with respect to the core process, the replacement and additive outliers have essentially the same effects on the sample autocorrelation function. However, when this is not the case, replacement outliers can affect the sample autocorrelation function more than the additive outliers.

First, we consider how a single, isolated, replacement or additive outlier with a magnitude equal to A, can affect the sample autocorrelation function. For instance, we replace a single

observation $x_j, 1 < j < N-1$ with A and compute the lag-one correlation coefficient. Re-evaluating (2.35), we see that A appears in the numerator linearly, and quadratically in the denominator.

$$\hat{\rho}_x(1)|_{x_j=A} = \frac{\sum_{n=1}^{j-2} x_n x_{n+1}^* + \sum_{n=j+1}^{N-1} x_n x_{n+1}^* + x_{j-1}A^* + Ax_{j+1}^*}{\left(\left[\sum_{n=1}^{j-1} |x_n|^2 + \sum_{n=j+1}^{N-1} |x_n|^2 + |A|^2\right]\left[\sum_{n=2}^{j-1} |x_{n+1}|^2 + \sum_{n=j+1}^{N} |x_{n+1}|^2 + |A|^2\right]\right)^{1/2}}. \tag{3.21}$$

If A is allowed to become very large, then (3.21) becomes small, approaching zero. It should be apparent that the same effect occurs for both the replacement and additive outlier models.

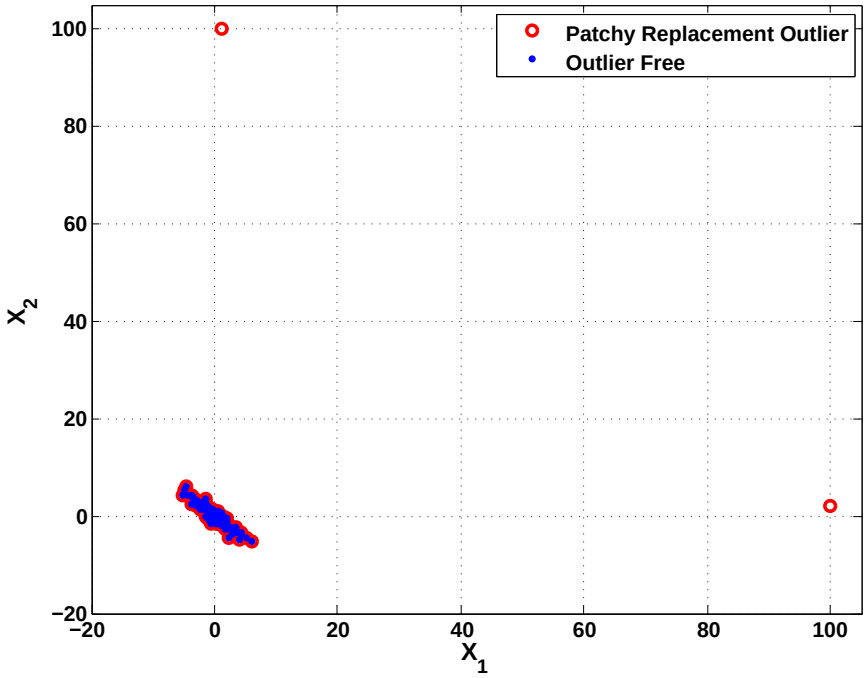

Figure 3.2: Scatter plot of correlated time series data without (blue dots) and with (red circle) patchy outliers. The patchy outliers are fixed in magnitude and appear at $X_1 = 100$ and $X_2 = 100$.

We extend this example by examining the effect of k consecutive patchy replacement or additive outliers. If the distribution of the outliers is chosen to be white with a large variance, again, the sample autocorrelation function will be biased towards zero as the variance of the outlier process becomes large. However, if the distribution of the outliers is a point-mass

distribution with a large magnitude, the result is different. In this case, as the number of patchy outliers increases, the sample autocorrelation is biased towards one. In fact, the output of the sample autocorrelation will be $(k-1)/k$ where k is the number of patchy outliers. Figs. 3.2 and 3.3 present two ways of visualizing the impact of these replacements outliers. Fig. 3.2 shows the scatter plot of X_t versus X_{t-1} for both the uncontaminated and contaminated cases. Notice that the outliers appear at $X_1 = 100$ and $X_2 = 100$. The same outliers are observed Fig. 3.3 at sample numbers 21 through 25 having a fixed value of 100. Note that the outliers in the scatter plot appear in pairs of corrupted variables. Also note that if we were to plot the data containing only one outlier, the scatter plot would look similar, however only one pair of observations would be affected. The time series plot would obviously appear different, with only one abnormal observation.

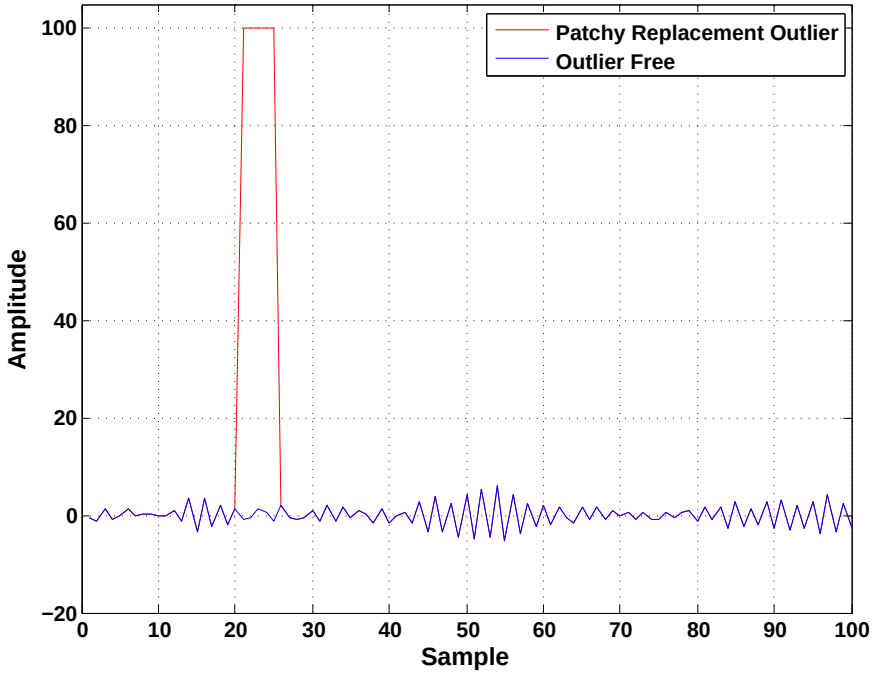

Figure 3.3: Another view of patchy replacement outliers for a time series. Notice that the patchy outliers occupy consecutive sample positions.

These examples highlight several important remarks. First, a single outlier can affect the sample autocorrelation function biasing it towards zero. Second, isolated and patchy outliers

can also have different effects, causing the sample autocorrelation function to be biased differently. Finally, when choosing an outlier model for robustness analysis, we should take care to make sure it accurately represents the application and outlier process.

3.5.2 Innovative Outliers

Referring back to (2.1), outliers occurring in the innovations process, U_t, are appropriately called innovative outliers. The innovative outlier model does not describe non-Gaussian observation noise, but is examined for completeness. In fact, it is possible for a single innovative outlier to theoretically impact all subsequent observations if the time series model contains feedback. For examle, Fig. 3.4 shows a scatter plot of an AR(1) system with and without an innovative outlier. The uncontaminated output in blue (dots) lies along the line with a slope of -0.95. In red (circles), we observe that the scatter plot gets stretched but still lies along the same line.

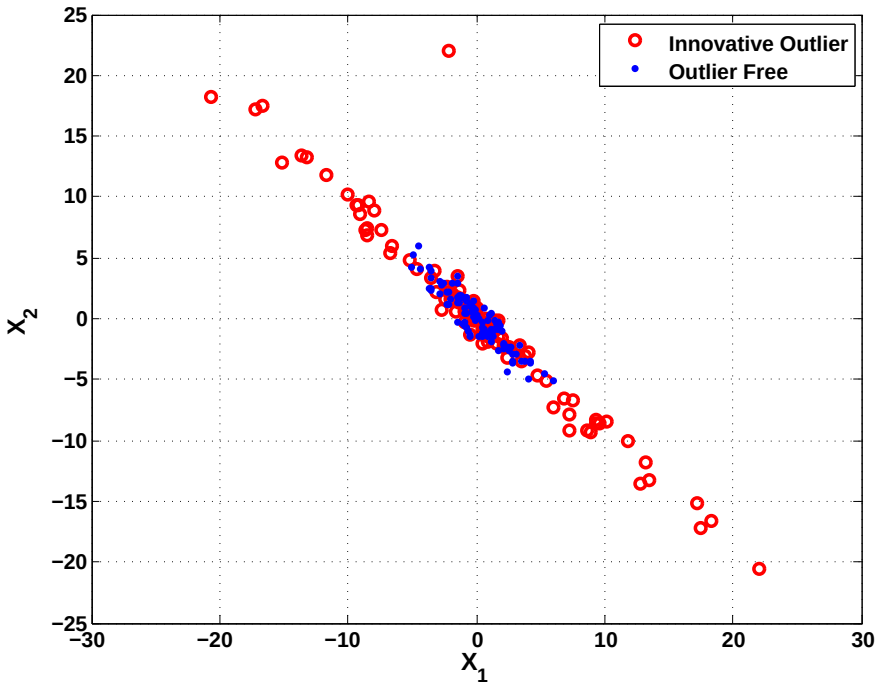

Figure 3.4: Scatter plot of AR(1) time series without (blue dots) and with (red circles) an innovative outlier. Notice how the scatter plot is stretched by the innovative outlier, but its orientation is not changed.

Surprisingly, innovative outliers do not degrade the sample autocorrelation function. In fact, ARMA processes with infinite innovations variance actually improve the precision of the

sample autocorrelation function [46]. We can think of the output of an ARMA process due to impulsive innovations as being very similar to the impulse response of the system. Fig. 3.5 shows the same example but now as a function of sample number. The effect of the innovative outlier starts at sample 20. The red curve shows that the impact lasts for several samples until it decays at about sample 80. The only issues that arise from an estimation point of view, is that the mean and variance of the ARMA process will be difficult to estimate.

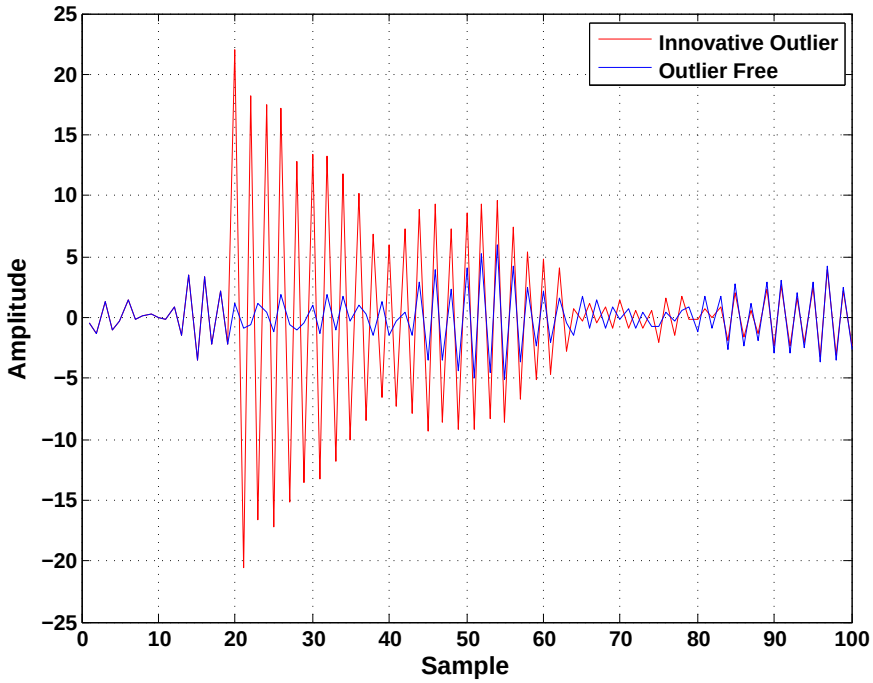

Figure 3.5: The linear time series plot of an AR(1) time series shows that many samples are affected after a single innovative outlier is inserted into a time series. The sample autocorrelation function is not adversely affected by innovative outliers.

3.5.3 Outlier Detection Using Robust Mahalanobis Distances

Distance metrics are commonly used to identify statistical outliers. Consider independent observations of a univariate Gaussian random process with a known mean and variance. The distribution of the square of the distances of the observations from the mean, normalized by the variance, follows the chi-squared distribution with one degree of freedom, that is $d_n^2 = \frac{(x_n-\mu)^2}{\sigma_x^2} \sim \chi_1^2$. Outliers can be declared if their distance exceeds some percentile, such as the 97.5 percentile. For instance, an outlier is flagged for the n^{th} observation if $d_n^2 > \chi_{1,0.975}^2$.

As long as the observations are independent, we can extend this example to include vectors of observations. For instance, suppose $\mathbf{x}_i = \{x_{i,1}, x_{i,2}\}^T$ and $\mathbf{x}_j = \{x_{j,1}, x_{j,2}\}^T$. Then $d_{ij}^2 = \frac{(\mathbf{x}_j-\boldsymbol{\mu}_j)^T}{\sigma_{x_j}^2}\frac{(\mathbf{x}_i-\boldsymbol{\mu}_i)}{\sigma_{x_i}^2} \sim \chi_2^2$. In general, the distribution of the square of the distances follows a χ_n^2 distribution with $n = N$ degrees of freedom where N is the length of the data vector.

When the observations are dependent with covariance matrix $\boldsymbol{\Sigma}$, we must first whiten the demeaned data. Assuming $\boldsymbol{\Sigma}$ is positive definite, then a matrix $\boldsymbol{\Sigma}^{1/2}$ exists and is invertible. Whitening is accomplished by multiplying the demeaned data vector $(\mathbf{x} - \boldsymbol{\mu})$ by $\boldsymbol{\Sigma}^{-1/2}$. The resulting squared distances are the squared Mahalanobis distances [45],

$$MD_{ij}^2 = (\mathbf{x}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_j). \tag{3.22}$$

The squared Mahalanobis distances for real-valued Gaussian random variables follow a χ_n^2 distribution with $n = N$ degrees of freedom where again, N is the length of the vectors. If the vectors consist of complex-valued Gaussian random variables, the squared Mahalanobis distances follow a χ_{2N}^2 distribution with $2N$ degrees of freedom. Notice that the transpose in (3.22) is replaced with the complex-conjugate transpose ^{H}.

When dealing with observations from an unknown Gaussian distribution, we must estimate both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. However, we must do so using a robust estimator, otherwise the resulting distances may suffer from the masking effect. Therefore, we will have to replace $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

in (3.22) with robust estimates. Mahalanobis distances calculated with robust mean and covariance estimates will be called robust Mahalanbobis distances.

3.5.4 Robust Projection Statistics

An alternate, robust formulation of the Mahalanobis distance was developed by both Stahel [69] and Donoho [19]. To start, an equivalent from of (3.22) can be expressed as

$$MD_i^2 = \max_{||\mathbf{v}||=1} \frac{|\mathbf{x}_i^T\mathbf{v} - \frac{1}{m}\sum_{j=1}^m \mathbf{x}_j^T\mathbf{v}|^2}{\frac{1}{m-1}\sum_{k=1}^m \left(\mathbf{x}_k^T\mathbf{v} - \frac{1}{m}\sum_{j=1}^m \mathbf{x}_j^T\mathbf{v}\right)^2}, \tag{3.23}$$

where $\mathbf{v}$ is a unit-norm direction vector.

An obvious method to make (3.23) more robust is to replace the sample means with medians. We can scale the denominator to ensure it is Fisher consistent at the Gaussian distribution. This results in a robust version of (3.22) given by

$$PS_i^2 = \max_{||\mathbf{v}||=1} \frac{\left|\mathbf{x}_i^T\mathbf{v} - \underset{\mathrm{j}}{\mathrm{med}}\left(\mathbf{x}_j^T\mathbf{v}\right)\right|^2}{\left(1.4826 \times c \times \underset{\mathrm{k}}{\mathrm{med}}\left[\mathbf{x}_k^T\mathbf{v} - \underset{\mathrm{j}}{\mathrm{med}}\left(\mathbf{x}_j^T\mathbf{v}\right)\right]\right)^2}, \tag{3.24}$$

where $c = 1 + 15/(N-n)$ is scaling factor accounting for the dimension [59], $\dim\{\mathbf{x}_i\} = n$, and number of observation vectors $\mathbf{x}_i$, $i = 1, \cdots, N$. Note that when n is small and N is large, $c \approx 1$.

We note that this form requires all search directions $\mathbf{v}$, which in practice is not possible. A practical implementation proposed by Gasko and Donoho [20] restricts $\mathbf{v}$ to a finite number. For example, the method defined by Donoho and Stahel establishes K normalized vectors where K is the number of n-dimensional observation vectors. Each vector starts at the coordinate-wise median of all observations and points in the direction of the k^{th} observation. All data observations are then projected onto each directional vector $\mathbf{v}_k$. For a given direction, all projections are normalized, and the greatest normalized projection for each

direction vector is maintained. This maximum projection is the projection statistic for that observation vector.

The properties of the resulting projection statistics have been studied in detail [59, 49]. It was found that when the number of observations is greater than five times the dimensionality of the data, the square of the projection statistics roughly follow a χ^2_n distribution with $n = N$ degrees of freedom for data vectors containing N elements when $N > 5n$. Otherwise, Mili et al. [49] demonstrated that the projection statistics themselves, not their square, follow a χ^2_n distribution with $n = N$ degrees of freedom . A projection statistic algorithm was developed for complex-valued data by Schoenig and Mili [66, 65], and will be used in this analysis.

We observe that projection statistics are quite involved computationally. We notice that K^2 projection operations are required. Therefore, as the number of observations increases, the computational time grows at least quadratically. Additionally, the median operation is performed multiple times per projection statistic calculation. Fast sorting algorithms can be used for this purpose [7], exhibiting a cost that grows approximately linearly.

3.6 Definition of Breakdown Point and Maximum Bias for Complex-Valued Time Series

Two important and related concepts in global robustness analysis are estimator breakdown point and maximum bias. These concepts were originally developed for estimators of location and scale given independent and identically distributed observations. However, these concepts cannot be directly applied to estimators of temporal correlation. We first review the original concepts and build on them to yield useful definitions for estimators of correlation.

In the most generic sense, an estimator's breakdown point, ε^*, is the largest asymptotic contamination level of data that an estimator can tolerate and still provide useful information about the true parameter [46]. Consider the case where we are trying to estimate the

location parameter, for example the mean of a normally distributed random variable. Any real number is a valid value for the true value of the mean. Also, it is reasonable to assume that the observations used by the estimator may also be any real number. Therefore, both the domain and range of the estimator is all real numbers. Using this example, breakdown is defined to be the fraction of observations required to "pull" the estimator to a value that is arbitrarily far away from the parameter's true value [46]. For example, the sample average can be pulled infinitely away from the true value with one positive or negative valued outlier with an arbitrarily large magnitude. In this case, the outlier is said to have an *unbounded* influence on the estimator. The sample average has an *asymptotic* breakdown point of $\varepsilon^* = 0$, and the maximum bias is infinite. The same analysis can be performed for a robust estimator of location, such as the sample median. Huber [36, 38] showed the median to have the smallest maximum bias of all translation equivariant location estimators and an asymptotic breakdown point of 0.5. At which point, the maximum bias is also infinite.

Unlike the location and scale parameters, the range of the correlation coefficient is bounded. For complex-valued data, this range contains all points on or inside the unit circle in the complex plane. Therefore, outliers will not have an unbounded influence on the estimator (it is assumed that only valid correlation values are produced by an estimator for this discussion).

As a result, the definition of estimator breakdown must be modified for our analysis, as it is not possible to pull the estimator arbitrarily "far away" from the true value. Additionally, as we have discussed in Section 3.5.1, the structure of the additive or replacement outliers can influence how the sample correlation coefficient is biased.

For our analysis, we will adopt the definition proposed by Genton and Lucas [28], where an estimator breaks down when increasing the contamination rate no longer increases its range. This contamination rate will be called the breakdown point. For example, the MLE defined in (2.34) and the sample autocorrelation function in (2.35) will breakdown if a single isolated observation is replaced with an infinitely large value. The range in this case collapses to a point, namely the origin. The asymptotic breakdown point for this estimator is zero.

Adding additional isolated replacement outliers does not change the range. However, as we have discussed, changing the outlier model can cause this estimator to break down differently. Recall, a point-mass patchy outlier model with arbitrarily large values causes the estimator's range to collapse to one. Therefore, an outlier model must be chosen and held fixed to use this definition of breakdown point for our time series analysis.

In this analysis, we are interested in the robustness properties that the PPC and MRE display for impulsive noise. Since the isolated replacement outlier model resembles impulsive noise, we will follow the convention and consider that a breakdown has occurred when the estimator's range reaches the origin. Our signal model follows (3.20) where X_t are realizations of a complex-valued Gaussian random process following the distribution defined by (1.1) and Z_t is a zero-one process with $P(Z_t = 1) = \epsilon$, setting the contamination rate. The impulsive noise process is W_t.

An effective measure of robustness is the asymptotic maximum bias defined as

$$b_{max}(\epsilon) = \sup_{H} |T(F) - T(G)|, \tag{3.25}$$

where the supremum is taken over all H. It is assumed that the estimator is Fisher consistent at F, that is $T(F) = \theta, \forall\theta \in \Theta$. Based on our isolated replacement outlier model and the associated definition of breakdown, we choose H such that $T(G)$ in (3.25) tends to zero from the direction of $T(F)$ as depicted in Fig. 3.6. Therefore, the maximum bias is upper bounded by $|T(F)|$, that is

$$b_{max}(\epsilon) \leq |T(F)|. \tag{3.26}$$

The value of ϵ where the bias reaches the upper bound is called the *breakdown point*.

The cumulative probability distribution function H defines the oultier process, W_t. This design parameter is chosen to yield the worst-case bias by driving the estimator to zero. We note that the definition of H may depend on the contamination level ϵ. The choice of H is also dependent on the estimator under analysis. The asymptotic breakdown point for the

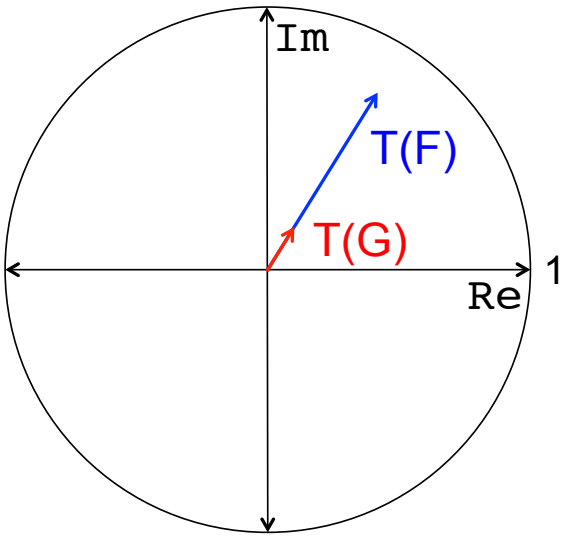

Figure 3.6: Uncorrelated additive or replacement outliers will bias estimators of correlation toward zero from the direction of the unbiased estimator. The maximum obtainable bias is defined to be the magnitude of the true correlation coefficient.

sample autocorrelation function is zero. The maximum bias of this estimator is equal to the magnitude of the true correlation coefficient to be estimated.

Finally, observe that we have modified the definition of bias given by (3.5) to discard phase. This was chosen primarily for ease when plotting results. However, since we have forced $T(G)$ to the origin from the direction of $T(F)$, the phase of $T(G)$ will be the same as $T(F)$. Thus, we do not lose any information using this convention.

3.7 Desirable Properties of Robust Estimators

The ε-contaminated outlier model introduced in (3.20) affects estimates of univariate parameters, such as the mean and variance, differently than estimates of correlation for a time series. In the univariate examples, a single outlier only corrupts one observation. However, estimating the correlation coefficient for a time series requires the estimator to consider time dependence. Thus, pairwise groupings of the observations are processed. This requirement causes

a single outlier to impact two pairwise groupings, thereby increasing the contamination rate observed by the estimator of correlation. Fig. 3.7 shows the probability that two randomly chosen observations from an ε-contaminated time series contain an outlier, that is the pairwise contamination rate. This probability is calculated by $p = \sum_{j=1}^{n} \binom{n}{n-j+1} \varepsilon^{n-j+1}(1-\varepsilon)^{j-1}$, where n is in general the number of observations randomly chosen, two in this case.

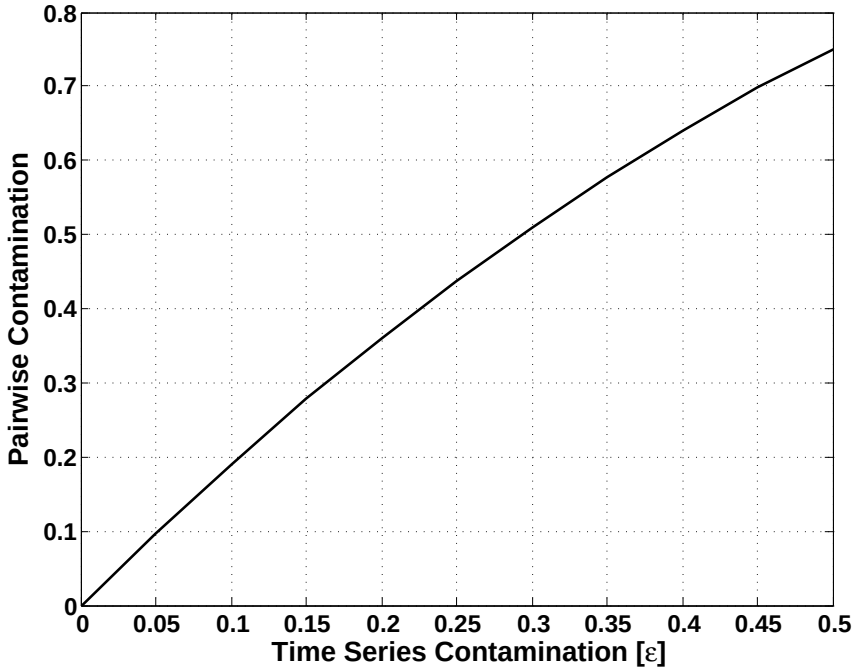

Figure 3.7: Effective pairwise contamination rate of an ε-contaminated time series. Fifty percent of the pairwise observations will be contaminated with at least a single outlier when the time series contamination rate is approximately 0.3

Traditionally, robust estimators are designed to withstand a fraction of the observations deviating from the assumed distribution model. For instance, robustness analysis of equivariant estimators of location parameters stops at an asymptotic contamination rate of $\varepsilon^* \leq 0.5$, beyond which their bias becomes unbounded and the estimator is said to breakdown. Fig. 3.7 highlights the difficulty encountered in time series analysis, as we see that half of the pairwise observations will be contaminated at a time series contamination rate slightly below $\varepsilon = 0.3$.

3.8 Huber Estimation

The theory of the maximum likelihood estimators assumes exact knowledge of the probability density function conditioned on the unknown parameters. However, we have seen that even a slight deviation from the assumed model can severely degrade their performance. Huber considered a different approach [36], one that provides robustness to model approximations. For example, Huber postulated that the probability distribution of the observations is the ε-contaminated model given by (3.19). While the resulting distribution, G is in the neighborhood of F, it is unknown. Therefore, an alternate to the maximum likelihood approach is needed.

Realizing that the least-squares estimator is sensitive to a slight deviation from Gaussianity, Huber [37] proposed alternative estimators that offer robustness to abnormal observations while maintaining most of the efficiency under the assumed distribution. This new class of estimators, termed ML-type estimators or M-estimators for short, minimize the sum of a function of the residuals over all the observations. Under this framework, the maximum likelihood estimators fall under the general category of the Maximum Likelihood- Type estimators. An example is developed for the regression model in the next section.

3.9 Generalized Maximum Likelihood-Type Estimators

For the general linear regression model given by (3.10), outliers can occur in both $\mathbf{z}$ and the factor space spanned by the row vectors of $\mathbf{H}$. Outliers only in $\mathbf{z}$ are called *vertical* outliers and can adversely affect estimators with unbounded influence of residuals, such as the least-squares estimator. Outliers in $\mathbf{H}$ can create either *good* or *bad* leverage points. A good leverage point is associated with an outlier in the factor space only. Therefore it fits the trend of the data and typically improves the precision of an estimator. In the case of an autoregressive model, a good leverage point can be thought of as an innovative outlier.

On the other hand, a bad leverage point occurs when the projection of the data point into the factor space is away from the bulk of the data and does not fit the trend, producing deleterious effects for estimators with unbounded influence of position.

A simple two-dimensional regression problem is now used to illustrate vertical outliers and leverage points. Figure 3.8 depicts a linear regression where the observations marked with circles fit a linear model exactly. Using the least-squares estimator, a linear model was estimated with (3.14) and a line was fit to the data with slope equal to one and intercept equal to zero. In this case, all of the residuals equal zero, minimizing the squared error.

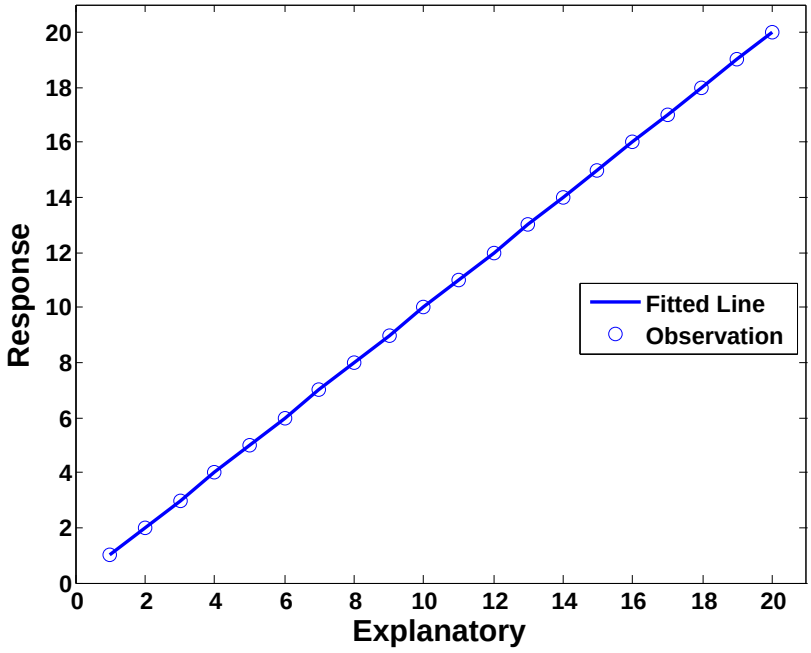

Figure 3.8: Linear regression model where the response and explanatory data are related to each other with a unit slope line and zero y-intercept. The least-squares estimator correctly estimates the slope and intercept of an exact fit model.

3.9.1 Vertical Outliers

We now introduce our first type of regression outlier and see the effect on the least-squares solution. In Figure 3.9, we observe a vertical outlier located at (2,60). Vertical outliers appear in the residuals defined in (3.11) and not in the explanatory variable. Recall that vertical outliers were seen in the time series example in Fig. 3.2 along the X_2 axis. A robust estimator is one that downweights this observation. By contrast, the least-squares estimator is not robust. In this example, it fits a line that has a slope of 0.2586 and an intercept of 10.6842 which are very different from the values (1,0). In fact, a single outlier can change the parameter values arbitrarily. It breaks down because it weights all observations equally.

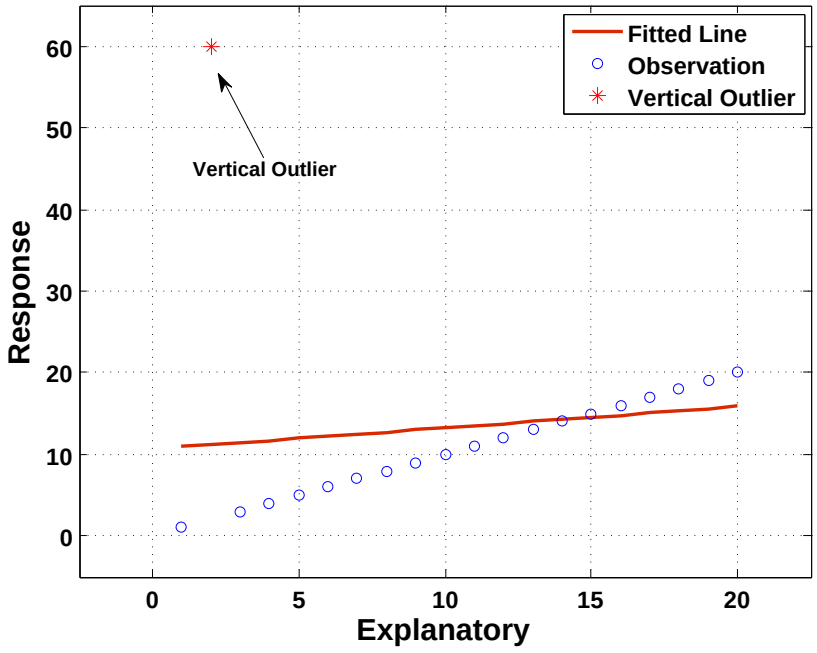

Figure 3.9: Linear regression with a single vertical outlier. The least-squares estimate is biased by the vertical outlier.

3.9.2 Good and Bad Leverage Points

We now discuss good and bad leverage points, the latter producing biased results on non-robust estimators including the M-estimators. In Figure (3.10), we show the effects these two different types of leverage points have on our linear regression. First, notice that the majority of the observations, once again indicated by blue circles, fall between 1 and 19. Now, we introduce two observations that do not fit the bulk of the data. The projection of these two leverage points onto the explanatory axis lies at 70 far away from [1,19], the range of the other data points. However, the black triangle located at (70,70) actually fits the trend of the bulk of the data, indicating it is a good leverage point. In fact, using the least-squares solution, the line fitted to the bulk of the data (blue circles) including this good leverage point (green asterisk) produces a line passing through the bulk of the observations and this good leverage point. The other leverage point, indicated by the red asterisk, does not fit the trend associated with the bulk of the data. This bad leverage point will drive the parameters of a non-robust estimator to arbitrary values. Observe that bad leverage points were seen in Fig. 3.2 along the X_1 axis.

3.9.3 Estimating the Correlation Coefficient With the GM-Estimator

Isolated time series replacement outliers appear as both vertical outliers and bad leverage points. Thus, if we are to use the linear regression model to estimate the correlation coefficient, we must examine an estimator that is robust to both. M-estimators are known to be sensitive to bad leverage points since they have an unbounded influence of position [60]. This motivates us to compare the PPC and MRE with the Schweppe-type Generalized M-estimator using the Huber function (SHGM), which is robust to all types of outliers, including bad leverage points. The SHGM [46] introduces mechanisms to downweight both vertical outliers and bad leverage points.

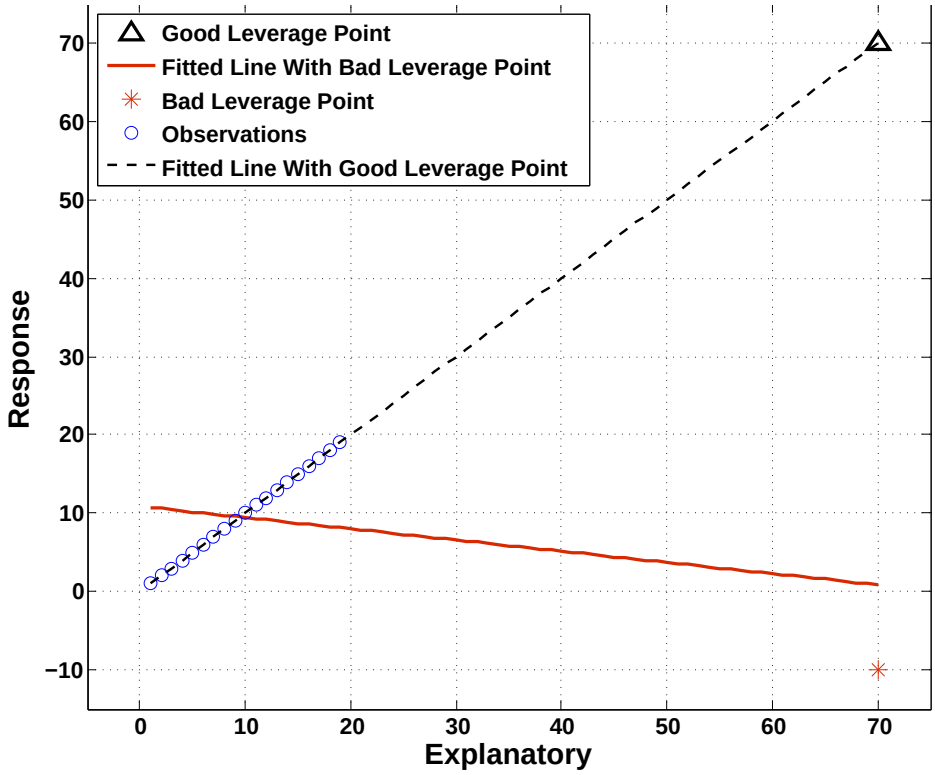

Figure 3.10: Linear regression estimates are affected by leverage points. Bad leverage points will bias the estimate (red line) while good leverage points improve the estimate (dashed black line). Robust estimators will simultaneously retain good leverage points while suppressing bad ones.

To see this, we start with the objective function that is used to derive the SHGM,

$$J(\mathbf{x}) = \sum_{i=1}^{m} w^2(\mathbf{h}_i)\, \rho\left(\frac{r_i}{\sigma_i w(\mathbf{h}_i)}\right). \tag{3.27}$$

The estimator is the solution to the minimization of (3.27) with respect to $\mathbf{x}$

$$\left.\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}} = \sum_{i=1}^{m} \frac{w(\mathbf{h}_i)}{\sigma_i} \frac{\partial \rho\left(\frac{r_i}{\sigma_i w(\mathbf{h}_i)}\right)}{\partial\left(\frac{r_i}{\sigma_i w(\mathbf{h}_i)}\right)} \left(\frac{\partial r_i}{\partial \mathbf{x}}\right) = \mathbf{0}. \tag{3.28}$$

We define

$$\psi(u) = \frac{d\rho(u)}{du}, \tag{3.29}$$

and observe that

$$\frac{\partial r_i}{\partial \mathbf{x}} = -\mathbf{h}_i{}^T. \tag{3.30}$$

Simplifying (3.28) and using a robust measures of dispersion (to replace σ_i), yields the implicit equation that defines the SHGM given by

$$\sum_{i=1}^{n} w\left(\mathbf{h}_i\right) \mathbf{h}_i \psi\left(\mathbf{r}_{Si}\right) = \mathbf{0}, \tag{3.31}$$

where $\mathbf{r}_{Si} = \mathbf{r}_i / \left(s_r w\left(\mathbf{h}_i\right)\right)$. In (3.31), we observe two mechanisms that bound the influence of residuals and of position. First, the Huber psi-function, $\psi\left(r_{S_i}\right)$, is the derivative of the rho-function, a nonlinear function of the standardized residuals, with respect to its argument. Here, the residuals are divided by a robust measure of scale, s_r, commonly determined by the normalized median absolute deviation (MADN) [46] defined by

$$\text{MADN}(\mathbf{r}) = 1.4826 \beta_m \text{Med}\left(|\mathbf{r}|\right). \tag{3.32}$$

The median of $\mathbf{r} = 0$ under the null hypothesis, $E\left\{\mathbf{r}\right\} = \mathbf{0}$ The scaling factor of 1.4826 ensures a Fisher consistent estimate at the *normal* or Gaussian assumption, hence the 'N' in the name MADN. An additional correction factor β_m was tabulated by Hampel et al. [34] for sample sizes $m \leq 9$ and

$$\beta_m = \frac{m}{m - .8}, \tag{3.33}$$

for larger sample sizes, quickly approaching 1 as m becomes large. A bounded *psi-function* is employed to bound the influence that vertical outliers in $\mathbf{z}$ have on the estimator.

Since the weighting used in the SHGM is a function of $\mathbf{x}$, (3.31) must be solved iteratively. A common solver is the Iteratively Reweighted Least-Squares (IRLS) algorithm. The IRLS

algorithm is expressed as

$$\mathbf{x}^{k+1} = \left(\mathbf{H}^T\mathbf{Q}^k\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Q}^k\mathbf{z}, \tag{3.34}$$

where $\mathbf{Q}$ is a diagonal matrix of weights, that is $\mathbf{Q} = \text{diag}\left[\psi\left(\mathbf{r}_{Si}\right)/\mathbf{r}_{Si}\right]$.

Relating the SHGM to the least-squares estimator provides some additional insight into the robust mechanisms incorporated into this estimator. First, if the weight function $w\left(\mathbf{h}_i\right)$ in (3.31) is set to $\mathbf{1}$, the SHGM reduces to an M-estimator. This makes it vulnerable to bad leverage points. Second, if the rho-function of the residuals is set to $\rho(r_{Si}) = r_{Si}^2/2$ and minimized with respect to the normalized residuals we obtain $\psi\left(r_{Si}\right) = r_{si}$. This produces the least-squares estimator. The unbounded least-squares psi-function fails to bound the influence of vertical outliers. Thus, the least-squares estimator lacks robustness to both bad leverage points and vertical outliers, which was discussed previously.

To aid in the understanding of how the residuals are used to weight the observations, we will explicitly define and plot the Huber *rho-function*, *psi-function*, and the *weight* function with tuning parameter k, and compare them to associated functions of the least-squares estimator.

The Huber function is defined as

$$\rho_{\text{Huber}_k}\left(r\right) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq k; \\ k|r| - \frac{1}{2}k^2 & \text{if } |r| > k, \end{cases} \tag{3.35}$$

and is quadratic for residuals less than the tuning parameter, k. Above this value, the Huber function becomes linear. The parameter can be adjusted to trade efficiency with robustness. For instance, if k is set arbitrarily large, the Huber rho-function becomes the same as the least-squares rho-function, approaching the efficiency of the Gaussian MLE, but lacking all robustness.

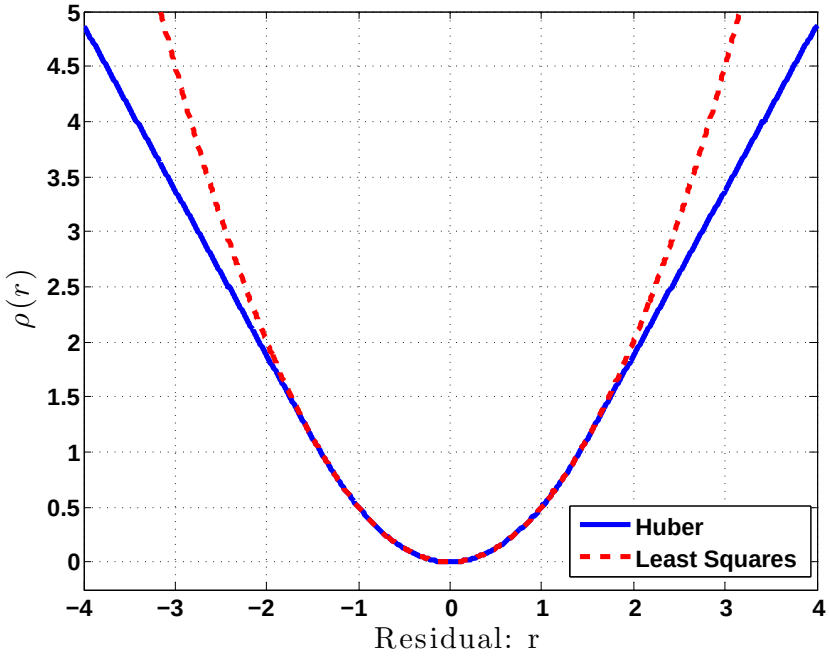

Figure 3.11: The least-squares and Huber rho-function with tuning parameter $k = 1.5$. The Huber function becomes linear when the magnitude of r is greater than k.

The psi-function is the derivative of the rho-function. The least-squares,

$$\psi_{\text{LS}}(r/s) = r/s, \tag{3.36}$$

and Huber psi-functions

$$\psi_{\text{Huber}_k}(r) = \begin{cases} r/s & \text{if } |r/s| \leq k; \\ \text{sign}(r/s) & \text{if } |r/s| > k, \end{cases} \tag{3.37}$$

are illustrated in Figure 3.12. The least-squares psi-function is linear. Similarly, we see that when the magnitude of the residual is less than the tuning parameter k, the Huber function is identical to the least-squares psi-function. However, we see that for magnitudes greater than k, the function is constant and limited in magnitude. As we will see, this is where we

gain the robustness to vertical outliers.

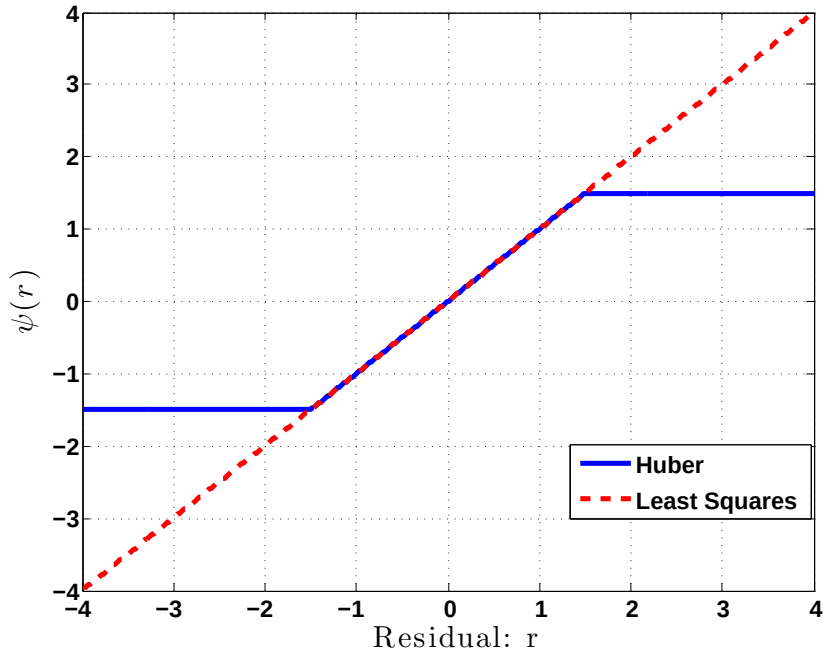

Figure 3.12: Least-squares and Huber psi-function, magnitude of Huber function is limited to k as magnitude exceeds $k = 1.5$.

Finally, we observe the weighting function, $q(r) = \psi(r)/r$. This function is used to weight the observations in the iteratively reweighted least-squares solution in (3.34). The least-squares rho-function yields equal weights

$$q_{\mathrm{LS}}(r) = 1, \tag{3.38}$$

to all observations regardless the value of the residual. However, the Huber weight function,

$$q_{\mathrm{Huber}_k}(r) = \begin{cases} 1 & \text{if } |r| \leq k; \\ \mathrm{sign}(r)k/r & \text{if } |r| > k, \end{cases} \tag{3.39}$$

downweights observations with residuals greater than k. This weighting function is how the

SGHM gains robustness to vertical outliers. These weighting functions are depicted in Fig. 3.13, decreasing in value as the residual gets larger.

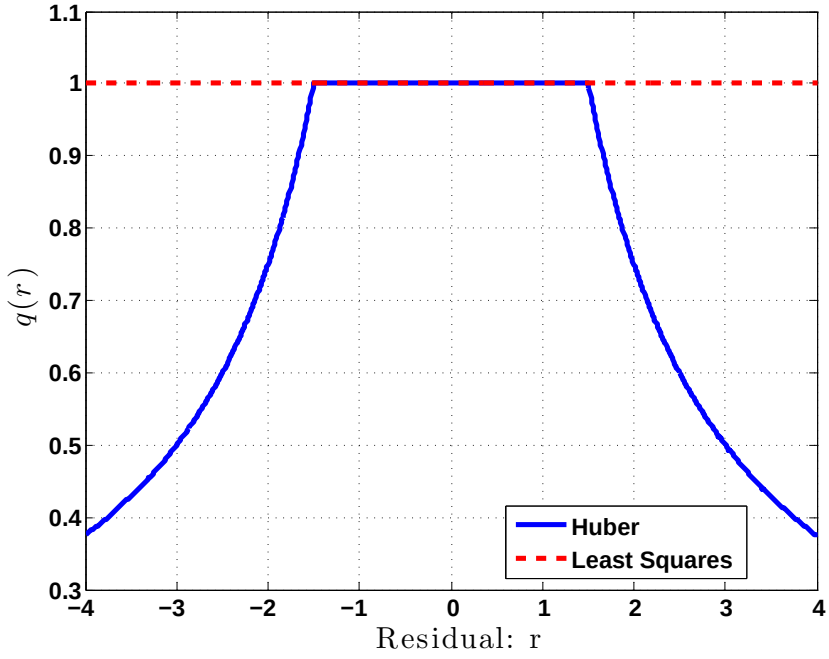

Figure 3.13: The Least-squares and Huber weighting functions with tuning parameter $k = 1.5$. The least-squares estimator weights all residuals equally (dashed line), while the Huber weight function assigns a decreasing weight as the magnitude of the residual increases.

The second downweighting mechanism, $w(\mathbf{h}_i)$, is the weight function which assigns weights decreasing in magnitude to outliers in the factor space spanned by the row vectors of $\mathbf{H}$. The SHGM relies on this weighting mechanism to downweight bad leverage points while at the same time preserves good leverage points; notice that it also appears in the denominator of robustly scaled residuals. In this analysis, the weight function is defined by

$$w(h_i) = \min\left(1, \frac{b}{PS_i^2}\right), \text{ where } b = \chi^2_{n,0.975}, \tag{3.40}$$

and is inversely proportional to the squared projection statistics [66] [27] which can be

thought of as robust Mahalanobis distances. This weighting function is a function of the square of the projection statistics and is shown in Figure 3.14 with a tuning parameter, b, set to $\sqrt{\chi^2_{n,0.975}} = 2.71$. This tuning parameter is chosen since Projection Statistics roughly follow the chi-square distribution (see Section 3.5.4) with n degrees of freedom, which in our example is equal to two.

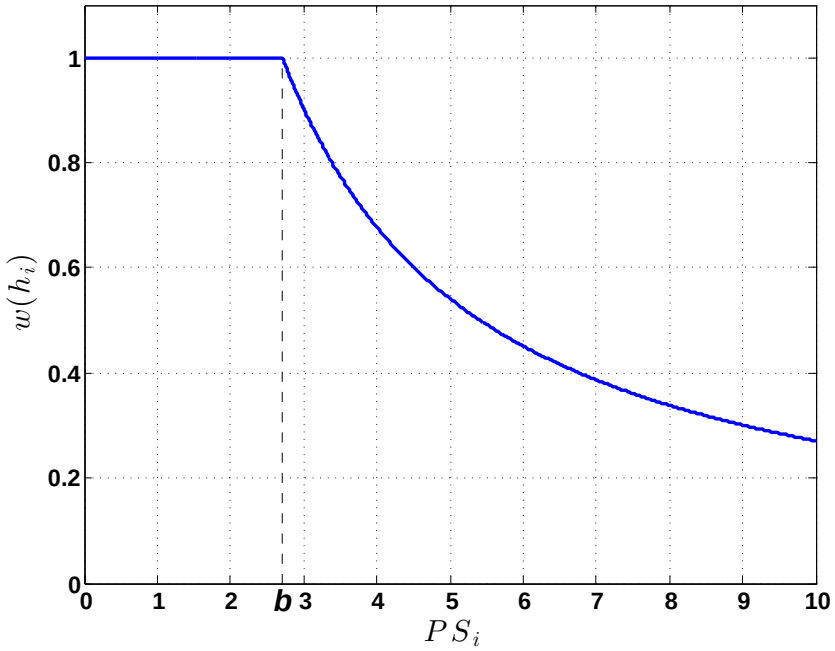

Figure 3.14: Leverage point weighting as a function of the robust projection statistics parameterized by $b = 2.71$ (dashed line). Observations with increasing projection statistics values are downweighted.

We finally revisit the linear regression example, this time applying the least-squares and the SHGM estimators. The resulting "fitted" lines are graphed along with the observations and outliers in Figure 3.15. The line determined by the least-squares estimator has a slope of 0.3657 and an intercept of 7.8181. Clearly, the different types of outliers bias this estimator. Even with the good leverage point, the vertical outlier and bad leverage point provide unbounded influence on the estimator. In contrast, we see that the line fitted by the SHGM

very closely matches the bulk of the data. The line resulting from this estimator has a slope of 0.9991 and an intercept of 0.0372. The SHGM has produced an estimate with very small biases in the presence of very large outliers in an exact-fit setting.

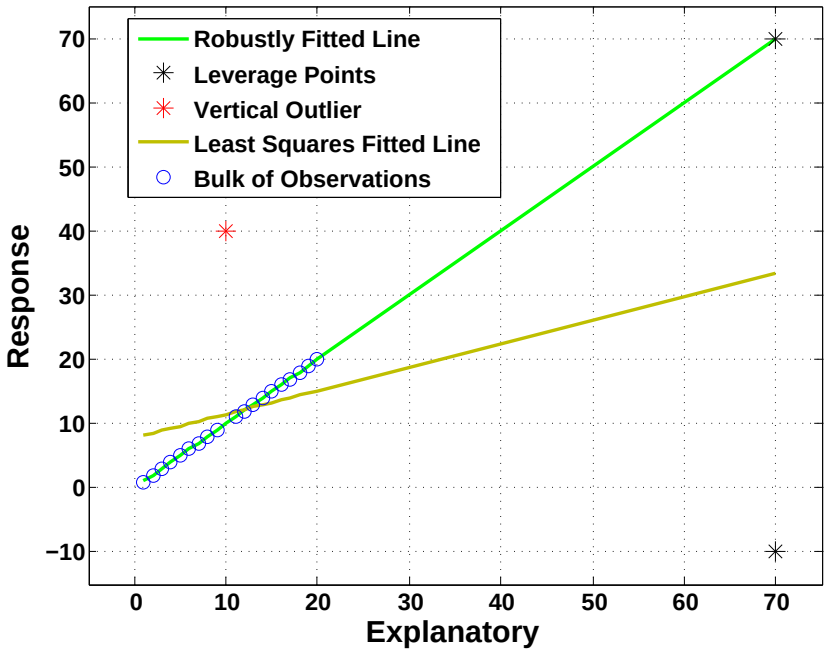

Figure 3.15: Linear regression result using both least-squares and robust SHGM. The bad leverage point and vertical outliers bias the least-squares estimator. The SHGM correctly downweights both the vertical outlier and bad leverage point while preserving the good leverage point. The SHGM correctly fits a line to the bulk of the data.

In general, there are a few implementation issues one must consider when implementing a GM-estimator. First, if we choose to use the IRLS algorithm, we must understand that it is iterative. To guarantee convergence, the cost function must be strictly convex. Additionally, strongly convex cost functions are desired to avoid slow convergence. However, non-convex rho-functions with redescending psi-functions, Tukey's bisquare function for example, can be used if one uses a robust initial condition to help guarantee convergence. Convergence is not a concern when using the Huber cost function.

When we specifically consider the SHGM, we notice that it is a tunable estimator. This is seen in the two weighting functions. The tuning parameters are used to trade efficiency for robustness. The values chosen in this dissertation are fairly standard and are well known to exhibit high efficiency without significantly sacrificing robustness.

An issue with the SHGM, as presented, is that the leverage point weighting function can take quite a long time to compute. As presented, projection statistics are used as a robust Mahalanobis distance. For the time series application, these calculations can prove to be impractical, as the computational complexity grows with the square of the number of observations.

Related to the issue of leverage point identification is the fact that this estimator is designed on the assumption that the residual errors are uncorrelated. For time series applications, this assumption is often violated. However, due to simplicity this is often ignored. Additionally, when estimating the correlation coefficient, the explanatory matrix is actually a column vector. Outliers are only identified by examining the magnitude of the observation. We realize that this is not the best method when temporal correlation exists. We therefore will propose two variants of the SHGM. The first method examines only the magnitude of the data, and does not exploit the temporal correlation. The second method exploits the pairwise correlations using robust projection statistics. The new PPC and MRE will be compared to both variants of the SHGM.

It is worth noting that the SHGM is only robust to outliers if there are redundant observations. This requires that the number of observations, which is the dimension of $\mathbf{z}$, must be much greater than the number of unknowns, which is the dimension of $\mathbf{x}$, in (3.10). When using this estimator to estimate the correlation coefficient, this should not be an issue since $\mathbf{H}$ is a vector and $\mathbf{x}$ is a scalar.

Finally, the SHGM can be implemented for complex-valued data in one of two ways. First, the cost functions can be formulated such that they are real-valued functions of complex-valued inputs such as was done in [2, 53] for M-estimators. Additionally, we must modify the

IRLS algorithm, using Hermitian transposes in place of the transpose operator. The second option is to rewrite the linear regression equation (3.10) as

$$\begin{bmatrix} \Re\{\mathbf{z}\} \\ \Im\{\mathbf{z}\} \end{bmatrix} = \begin{bmatrix} \Re\{\mathbf{H}\} & -\Im\{\mathbf{H}\} \\ \Im\{\mathbf{H}\} & \Re\{\mathbf{H}\} \end{bmatrix} \begin{bmatrix} \Re\{\mathbf{x}\} \\ \Im\{\mathbf{x}\} \end{bmatrix} + \begin{bmatrix} \Re\{\mathbf{e}\} \\ \Im\{\mathbf{e}\} \end{bmatrix}, \tag{3.41}$$

where $\Re\{\}$ and $\Im\{\}$ are the real and imaginary operators. We can then use the formulation previously developed. The latter method is chosen for implementation in this analysis, following the work of Schoenig [65]. When implementing the projection statistics, we will adopt the method for complex-valued data developed by Schoenig as mentioned earlier.

Chapter 4

Robust Correlation Estimators for Complex-Valued Gaussian Processes Based on Memoryless Nonlinearities

We have seen that properly constructing the regression problem allows us to estimate the autocorrelation coefficients of a time series. Outliers, additive or replacement, can create both vertical outliers and bad leverage points, potentially biasing this estimator towards zero. We can use a robust GM-estimator to downweight both the vertical outliers and bad leverage points. Robustness is gained by sacrificing efficiency at the assumed Gaussian distribution.

A necessary step for the robust GM-estimator is the identification and downweighting of the bad leverage points. Often, the suggested methods to identify these bad leverage points require the use of the covariance matrix, or autocovariance (autocorrelation) in the case of time series analysis, that describes the underlying non-corrupted process of the observed data. With this knowledge, robust distance metrics such as robust Mahalanobis distances can be used to identify the bad leverage points. In our case, the estimator requires the autocovariance (or autocorrelation), which is the exact quantity that it is trying to estimate.

If this method is to be pursued, robust covariance estimates that can be used to accomplish this include Rousseeuw's Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE).

A potential substitute to this approach is the calculation of robust projection statistics as mentioned earlier. The latter directly compute distance metrics which can be thought of as robust Mahalanobis distances. Unfortunately, this method is very computationally burdensome. Its computing time is a nonlinear function of the number of samples used to compute the estimate, becoming prohibitive for real-time applications as the number of observations becomes very large. Additionally, the breakdown point of the projection statistics are approximately $\epsilon = 0.3$ for time series applications, due to contaminated pairwise observations as illustrated in Fig. 3.7.

One might question why the GM-estimator is even needed if it requires a robust estimate of the autocovariance, to create appropriate weighting for the corrupted observations. Perhaps, fast estimators that exhibit robustness to outliers in some other manner, instead of non-quadratic cost functions of the residuals, may be used instead.

To meet these requirements, we explore alternative methods for complex-valued time series. For instance, instead of non-quadratic cost functions of residuals, we explore applying simple nonlinear functions directly to the observations. These are the PPC and the MRE and their iterative versions. We propose a new methodology to carry out robustness analysis of complex-valued correlation indices in time series.

At first, one might suspect that limiting the random processes to Gaussian would preclude these estimators from practical applications. However, this limitation is not as restrictive as one might initially think, as there are many instances where the observed complex-valued time series under analysis is Gaussian, or very close to Gaussian. We know from the central limit theorem that the sum of independent identically distributed finite variance random processes tend to Gaussian. However, under certain conditions, this theorem can be relaxed. For example, the Lindeberg-Feller central limit theorem does not require the individual

random processes to be identically distributed as long as the variance of any single process does not dominate the variance of the sum of processes.

4.1 General Equation For The Cross-Correlation of Two Complex-Valued Gaussian Processes at the Output of a Memoryless Nonlinearity

In [55], Reed uses Laguerre polynomials to derive the cross-corrleation function of two zero-mean complex-valued Gaussian random processes at the output of two envelope distorting filters. We will use this result to develop two nonlinear robust estimators, namely the PPC and MRE that can be applied to complex-valued Gaussian processes.

In this analysis, consider (1.1) to be a two- dimensional probability density function $f_{\mathbf{X}}(x) = f_{x_1x_2}(x_1, x_2)$ with covariance matrix $\mathbf{\Sigma}$ defined by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{x_1}^2 & \rho_{x_1x_2}\sigma_{x_1}\sigma_{x_2} \\ \rho_{x_2x_1}\sigma_{x_1}\sigma_{x_2} & \sigma_{x_2}^2 \end{bmatrix}, \tag{4.1}$$

where σ_x is the standard deviation of X_t, and $\rho_{x_1x_2}$ is the correlation coefficient given by

$$\rho_{x_1x_2}(\tau) = \frac{E\left\{X_{1,t}X_{2,t+\tau}^*\right\}}{\sigma_{x_1}\sigma_{x_2}}. \tag{4.2}$$

Note that $X_{2,t}$ may simply be a time-shifted version of $X_{1,t}$.

Let $R_{1,t}, \theta_{1,t}$ and $R_{2,t}, \theta_{2,t}$ be the magnitude and angle of the random processes $X_{1,t}$ and $X_{2,t}$ respectively. Reed [55] considers the distorting filters $y_t = G(R_t) = R_t^\alpha$ where α may be complex with real part greater than -2. If $G_1(R_{1,t}) = R_{1,t}^\alpha$ and $G_2(R_{2,t}) = R_{2,t}^\beta$, it is shown

that

$$E\left\{G_1(R_1)G_2(R_2)e^{jm(\theta_{2,t+\tau}-\theta_{1,t})}\right\} = e^{jm\mathrm{Arg}[\rho_{x_1x_2}(\tau)]}|\rho_{x_1x_2}(\tau)|^m\left(\sigma_1\sqrt{2}\right)^{\alpha}\left(\sigma_2\sqrt{2}\right)^{\beta}$$
$$\cdot\frac{\Gamma\left(\frac{m+\alpha}{2}+1\right)\Gamma\left(\frac{m+\beta}{2}+1\right)}{m!}\cdot{}_2F_1\left(\frac{m-\alpha}{2},\frac{m-\beta}{2};m+1;|\rho_{x_1x_2}(\tau)|^2\right), \tag{4.3}$$

where ${}_2F_1(a,b;c;z)$ is the Gaussian Hypergeometric function formally expressed as

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{n=0}^{\infty}\frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)}\frac{z^n}{n!}. \tag{4.4}$$

4.2 The Phase-Phase Correlator

The nonlinearity used by the PPC is the complex sign function for complex-valued data. It exploits the fact that the time correlation of a complex-valued signal is contained completely in its phase. The PPC then performs a correlation estimate using only the phase information of the observations. The robustness of this approach stems from the fact that all observations have the same impact on the estimator, regardless of their amplitude. Therefore, the influence of any particular observation is equal to all other observations.

4.2.1 Definition of the Phase-Phase Correlator

We now define the PPC, which obtains its robustness to impulsive noise by discarding the amplitude information, relying only on the relative phases of the complex-valued Gaussian processes to measure correlation. Formally, the PPC for complex-valued Gaussian processes is defined by

$$\rho_{\mathrm{PPC}}(\tau) = E\left\{e^{j(\theta_{t+\tau}-\theta_t)}\right\}. \tag{4.5}$$

The PPC is derived from (4.3) by setting $X_{2,t} = X_{t+\tau}$, α and $\beta = 0$ and $m = 1$. Doing so,

(4.3) becomes

$$\begin{aligned} & E\left\{e^{j(\theta_{t+\tau}-\theta_t)}\right\} \\ & = e^{j\mathrm{Arg}[\rho_x(\tau)]}|\rho_x(\tau)| \cdot \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{3}{2}\right)}{1!} \cdot {}_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right) \\ & = e^{j\mathrm{Arg}[\rho_x(\tau)]}\left(\frac{\pi}{4}\right)|\rho_x(\tau)|{}_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right). \end{aligned} \tag{4.6}$$

We emphasize that (4.6) is a phase preserving function of the correlation coefficent, however the magnitude is altered in a non-linear manner. Highlighting this, (4.6) can be written as

$$E\left\{e^{j(\theta_{t+\tau}-\theta_t)}\right\} = e^{j\mathrm{Arg}[\rho_x(\tau)]}f[|\rho_x(\tau)|], \tag{4.7}$$

where

$$f[|\rho_x(\tau)|] = \left(\frac{\pi}{4}\right)|\rho_x(\tau)|{}_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right). \tag{4.8}$$

Fig. 4.1 illustrates the nonlinear mapping between the magnitude of the correlation coefficient and the magnitude of the phase-phase correlator.

To implement the PPC-based estimator of correlation, we define the new random process

$$Y_{\mathrm{S}_t} = e^{j\mathrm{Arg}[X_t]} = \mathrm{Sign}(X_t), \tag{4.9}$$

which is the output of the complex-valued sign function applied to the original process X_t. Given a finite-length observation vector $\mathbf{x} = \{X_t, t = 1, 2, \ldots, N\}^T$, we define a new vector $\mathbf{y}_\mathrm{s}$ such that $\mathbf{y}_\mathrm{s} = \{Y_{\mathrm{S}_t}, t = 1, 2, \ldots, N\}^T$. Using the sample average, we estimate (4.7)

$$\hat{\rho}_{y_\mathrm{s}}(\tau) = \frac{1}{N-\tau}\sum_{n=1}^{N-\tau} y^*_{\mathrm{s}_n} y_{s_{n+\tau}}. \tag{4.10}$$

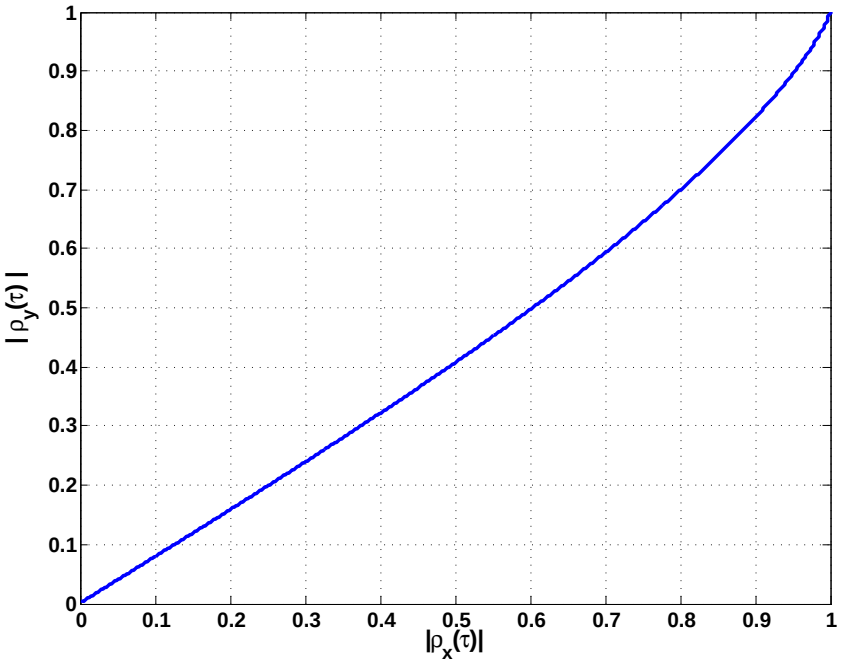

Figure 4.1: The nonlinear mapping $|\rho_y(\tau)| = f[|\rho_x(\tau)|]$ is based on the Gaussian Hypergeometric function. There is no closed-form expression for the inverse of this mapping, requiring the use of a look-up table.

Finally, the PPC estimator is defined as

$$\hat{\rho}_{x_{\mathrm{PPC}}}(\tau) = e^{j\mathrm{Arg}[\hat{\rho}_{y_{\mathrm{s}}}(\tau)]} f^{-1}\left[|\hat{\rho}_{y_{\mathrm{s}}}(\tau)|\right], \tag{4.11}$$

where we resort to a lookup table for $f^{-1}[\cdot]$ since no closed-form solution exists.

4.3 Asymptotic Maximum Bias Curve and Breakdown Point of the Phase-Phase Correlator

The asymptotic maximum bias and breakdown point of the PPC is interesting from a theoretical perspective. We will observe that the worst-case outlier model is unrealistic, but it provides insight as to why this estimator is so robust to uncorrelated impulsive noise.

Let us now perform a robustness analysis of the phase-phase correlator. To this end, we derive the worst-case outlier mechanism for the impulsive noise model. Due to the nonlinear nature of the PPC, we choose to model the impulsive noise with the replacement outlier model, providing a tractable analysis. This leads us to the asymptotic maximum bias curve and breakdown point of the PPC-based estimator as a function of the normalized correlation coefficient.

This robustness analysis is carried out by putting the estimator into the asymptotic functional form [68], $T(G)$, where G is the epsilon contaminated model from (3.19). Distributional robustness is exhibited by $T(G)$ if it produces meaningful estimates when the bulk of the data follows a given distribution F, typically supposed to be Gaussian, and a fraction of the data is contaminated with outliers following a different unknown distribution H. The worst case for estimators of location is obtained when H is a point mass distribution. However, as will be seen, the least favorable H can be defined differently in time series analysis.

Because the estimate $\hat{\rho}_x(\tau)$ is based on the PPC is dependent on $\hat{\rho}_y(\tau)$, we analyze the

robustness of the latter. Under epsilon contamination, we find that asymptotically we have

$$\begin{aligned} E\{\hat{\rho}_y(\tau)\} =& E\{Y_t Y_{t+\tau}^*\} \\ =& (1-\epsilon)^2 E\left\{\text{Sign}(X_t)\text{Sign}^*(X_{t+\tau})\right\} \\ &+ \epsilon(1-\epsilon) E\left\{\text{Sign}(W_t)\text{Sign}^*(X_{t+\tau})\right\} \\ &+ (1-\epsilon)\epsilon E\left\{\text{Sign}(X_t)\text{Sign}^*(W_{t+\tau})\right\} \\ &+ \epsilon^2 E\left\{\text{Sign}(W_t)\text{Sign}^*(W_{t+\tau})\right\}. \end{aligned} \tag{4.12}$$

Assuming that the outlier process W_t is uncorrelated with the Gaussian process X_t, and substituting the expectation term containing X_t and $X_{t+\tau}$ in (4.12) with $E\left\{X_t X_{t+\tau}^*\right\} = \frac{\pi}{4}\rho_x(\tau)_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right)$ yields

$$\begin{aligned} E\{\hat{\rho}_y(\tau)\} =& (1-\epsilon)^2 \frac{\pi}{4}\rho_x(\tau)_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right) \\ &+ \epsilon^2 E\left\{\text{Sign}(W_t)\text{Sign}^*(W_{t+\tau})\right\}. \end{aligned} \tag{4.13}$$

The maximum bias is obtained when the outliers force (4.13) as close to zero as possible. This is achieved when

$$\begin{aligned} &E\left\{\text{Sign}(W_t)\text{Sign}^*(W_{t+\tau})\right\} = \\ &\quad -\frac{(1-\epsilon)^2}{\epsilon^2}\frac{\pi}{4}\rho_x(\tau)_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right). \end{aligned} \tag{4.14}$$

We note that the magnitude of $E\left\{\text{Sign}(W_t)\text{Sign}^*(W_{t+\tau})\right\}$ cannot exceed unity because the sign-function transforms w_t into a unit-variance process. Therefore, the outliers are unable to drive the PPC to zero when the magnitude of the right-hand side of (4.14) exceeds one. Thus, after separating the correlation coefficient into its magnitude and phase, we see that

the outliers must be distributed such that

$$E\left\{\text{Sign}(W_t)\text{Sign}^*(W_{t+\tau})\right\} = -e^{\{jArg[\rho_x(\tau)]\}} \times min\left[1, \frac{(1-\epsilon)^2}{\epsilon^2}\frac{\pi}{4}|\rho_x(\tau)|_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right)\right], \tag{4.15}$$

in order to achieve the maximum asymptotic bias. This outlier distribution defines H in (3.25) and is a function of both ϵ and $\rho_x(\tau)$. Interestingly, the magnitude of the outliers does not affect the definition of H. This is explained by the normalization performed by the complex sign function on all observations to unit magnitude, including the replacement outliers. Worst case outliers are easily generated by choosing w_t as Gaussian random processes satisfying $\rho_w(\tau) = -e^{\{jArg[\rho_x(\tau)]\}} \times f^{-1}\left(min\left[1, \frac{(1-\epsilon)^2}{\epsilon^2}\frac{\pi}{4}|\rho_x(\tau)|_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right)\right]\right)$, where $f[\cdot]$ is defined as in (4.8).

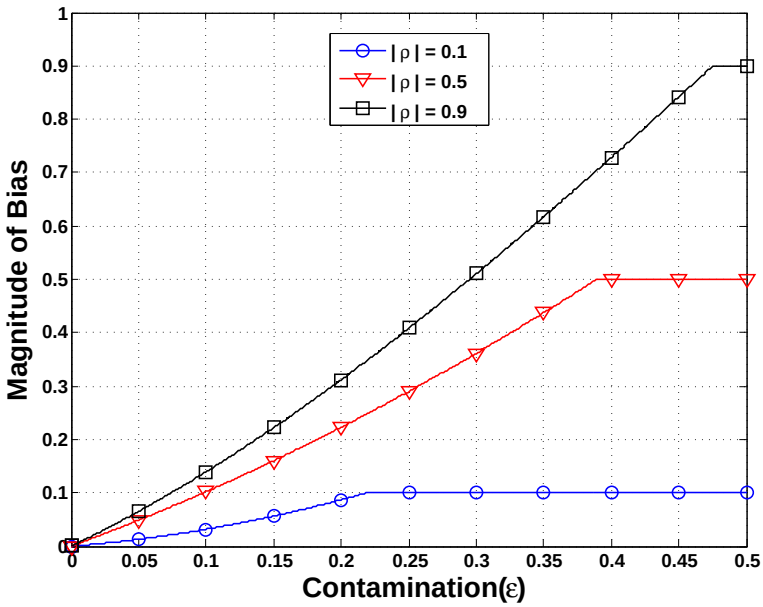

Figure 4.2: Maximum bias curves for phase-phase correlator with correlation coefficient magnitudes of 0.1, 0.5, and 0.9. The replacement outliers are worst case and occur in correlated pairs, driving the estimator to zero for the given contamination level.

Figure 4.2 shows the theoretical maximum asymptotic bias curves of the PPC for the cases where the magnitude of the correlation coefficient, $|\rho_x(\tau)|$, is 0.1, 0.5, and 0.9. The breakdown

of the PPC occurs when the asymptotic maximum bias curve reaches the upper bound defined by (3.26). From Fig. 4.2 we note that the bias, and thus the breakdown point of the PPC is a function of the magnitude of the normalized correlation coefficient.

The maximum asymptotic bias curves displayed in Fig. 4.2 are now approximated using a Monte Carlo simulation. The simulations perform one hundred cases for each contamination level, generating correlated sequences of 2000 samples using fixed correlation coefficient values with magnitude 0.1, 0.5, and 0.9. Worst-case, isolated replacement outliers are introduced satisfying the condition given by (4.14). The normalized correlation coefficient is estimated using (4.10) and (4.11) and averaged over the 100 cases. The bias for each contamination level is then approximated by computing the magnitude of the difference between the true value and the averaged experimental estimates. We notice excellent agreement between

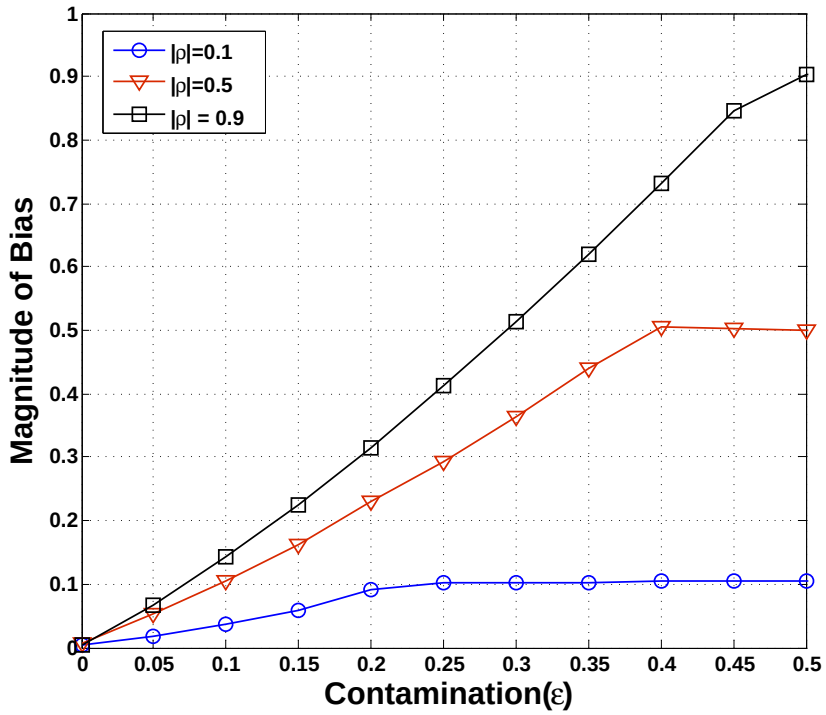

Figure 4.3: Approximated maximum asymptotic bias curves for phase-phase correlator with correlation coefficient magnitudes of 0.1, 0.5, and 0.9 using Monte Carlo simulations. Very good agreement is observed between theoretical and simulation based curves.

the theoretical and Monte Carlo-generated curves.

From (4.14) we can numerically determine the breakdown point for a given magnitude of the correlation coefficient. In Fig. 4.4 we show that this is an increasing function reaching 0.5 at $|\rho| = 1$. In comparison, we recall that a single arbitrarily large magnitude data

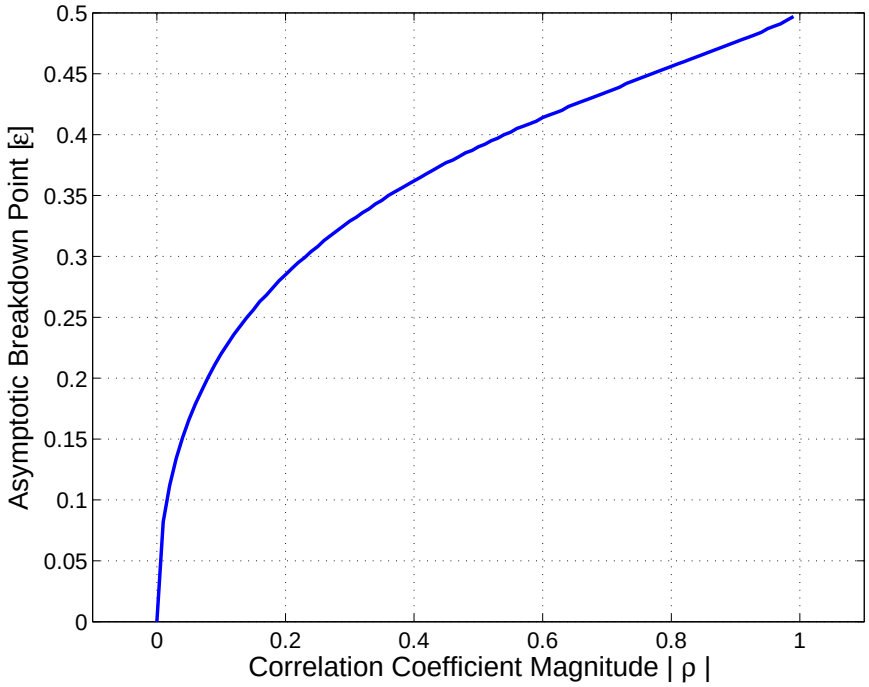

Figure 4.4: The asymptotic breakdown point of $\rho_x(\tau)$ calculated using the phase-phase correlator versus the correlation coefficient magnitude for worst case correlated outlier pairs.

observation forces the sample autocorrelation function to zero fulfilling (3.26). This indicates that the breakdown point of the sample autocorrelation function is zero, regardless of the magnitude of the correlation coefficient. We note that since the amplitude of the data is hard-limited when performing the PPC estimate, this same type of outlier does not have the same catastrophic impact. This observation indicates that the PPC-based estimator offers robustness to impulsive outliers.

Since this analysis is for the worst-case replacement outlier model, we can conclude that no other replacement outlier distribution produces more bias or a lower breakdown point. Thus, we expect the PPC to exhibit smaller bias and lower breakdown points when we introduce uncorrelated, impulsive replacement outliers.

4.3.1 Asymptotic Bias in the Presence of White Complex-Valued Gaussian Outliers

While the maximum asymptotic bias of the PPC is interesting from a theoretical perspective, the worst-case outliers do not match our model for uncorrelated impulsive noise. Additionally, for a more tractable analytical comparison to the MRE, it is useful to compute the asymptotic bias of the PPC in the presence of impulsive noise modeled by replacement outliers following a white complex-valued Gaussian random process.

Assuming W_t is white (Gaussian or not), the second term in (4.13) is zero. Therefore, $E\{\hat{\rho}_y(\tau)\} = (1-\epsilon)^2 \frac{\pi}{4}\rho_x(\tau)\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right)$. This value is now passed into the inverse of the nonlinear function $f[\cdot]$ in (4.7).

The resulting bias is therefore given by

$$b_{PPC}(\epsilon) = \left| |\rho_x(\tau)| - f^{-1}\left[(1-\epsilon)^2 \frac{\pi}{4} |\rho_x(\tau)|\, {}_2F_1\left(\frac{1}{2},\frac{1}{2};2;|\rho_x(\tau)|^2\right)\right]\right|. \tag{4.16}$$

The only restriction is that the replacement outliers are white. Obviously, this bias will be observed when the replacement outliers are complex-valued Gaussian random variables with any variance. We will revisit this when we compare the PPC to the MRE, and will show theoretical bias curves.

4.3.2 Asymptotic Variance Under the Gaussian Assumption

To compute the asymptotic variance of the estimate based on the PPC, we must analyze $\mathrm{VAR}\{\sqrt{N}\hat{\rho}_{y_s}(\tau)\}$ while letting the number of observations N, become large. The estimate $\hat{\rho}_{y_s}(\tau)$ is asymptotically unbiased and is obtained by the sample average, therefore the variance is evaluated via

$$\mathrm{VAR}\{\hat{\rho}_{y_s}(\tau)\} = E\left\{\frac{1}{N}\sum_{j=1}^{N}\left[y^*_{s_j}y_{s_{j+\tau}} - \rho_y(\tau)\right]\frac{1}{N}\sum_{k=1}^{N}\left[y^*_{s_k}y_{s_{k+\tau}} - \rho_y(\tau)\right]^*\right\}. \tag{4.17}$$

Expanding (4.17), distributing the expected value operator, and collecting like terms yields

$$\mathrm{VAR}\{\hat{\rho}_{y_s}(\tau)\} = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} E\left\{y_{s_j}^* y_{s_{j+\tau}} y_{s_k} y_{s_{k+\tau}}^*\right\} - |\rho_y(\tau)|^2. \tag{4.18}$$

This analysis was performed in [39] for independent pairs of correlated observations. In that situation, the expectation in (4.18) equals zero when $j \neq k$. However, for time series analysis, this is not the case. In fact, the expectation must be calculated for all values of j and k. For two- and three-term complex-valued Gaussian integrals, a closed-from solution is possible. However, in [57] Rice indicates that there is no known closed-form solution to the four-term integral

$$\int \frac{x_1^* x_2 x_3 x_4^*}{|x_1||x_2||x_3||x_4|} f_{\mathbf{x}}(\mathbf{x}) dx_1 dx_2 dx_3 dx_4, \tag{4.19}$$

where $f_{\mathbf{x}}(\mathbf{x})$ is defined by (1.1). As a consequence, the variance of the PPC under the Gaussian distribution for time series analysis must be approximated or evaluated analytically. We choose to perform a Monte Carlo analysis in the same manner as our approach for the SHGM estimator. These results will be shown when we compare the variance of both the PPC, MRE and SHGM.

4.4 Median of Ratios Estimator

The complex-valued MRE is an extension of the real-valued MRE proposed and analyzed by Yacine Chakhchoukh [14]. This new analysis was performed after it was observed that the same equation developed by Reed [55] can be used to develop both the PPC and the MRE. We introduce robustness to the MRE estimator by relying on the median operator. Since this analysis is for complex-valued data, the median operator is applied separately to the real and imaginary components of the observed data.

4.4.1 Definition of the Median of Ratios

Using Reed's results [55] to define the MRE requires us to first set $\alpha = -1$, $\beta = 1$ and $m = 1$ in (4.3). Additionally, if we assume that $X_{2,t+\tau} = X_{1,t+\tau}$, which is the case for our time series analysis, then $\sigma_1 = \sigma_2 = \sigma_x$, and (4.3) becomes

$$E\left\{(R_{t+\tau}/R_t)\, e^{j(\theta_{t+\tau}-\theta_t)}\right\} = e^{j\mathrm{Arg}[\rho_x(\tau)]}|\rho_x(\tau)|\left(\sigma_x\sqrt{2}\right)^{-1}\left(\sigma_x\sqrt{2}\right)\cdot\frac{\Gamma(1)\,\Gamma(2)}{1!}{}_2F_1\left(1,0;2;|\rho_x(\tau)|^2\right), \tag{4.20}$$

which reduces to

$$E\left\{(R_{t+\tau}/R_t)\, e^{j(\theta_{t+\tau}-\theta_t)}\right\} = E\left\{X_{t+\tau}/X_t\right\} = \rho_x(\tau), \tag{4.21}$$

since $\Gamma(1)$ and $\Gamma(2) = 1$ and ${}_2F_1\left(1,0;2;|\rho_x(\tau)|^2\right) = 1, \forall\rho_x(\tau)$. In [15], the robustness of the estimator is obtained by replacing the expectation with the sample median. In the complex-valued version of the MRE, we need to take the coordinate-wise median, that is the median of the real and imaginary components separately. However, before we make this substitution, we must verify that the distribution of the ratio of the complex-valued Gaussian processes is in fact symmetric along the real and imaginary axis.

The ratio of correlated complex-valued Gaussian random processes was examined by Baxley *et al.* in [4]. By making the assumption that the two processes are time-shifted versions of a correlated complex-valued Gaussian process with autocorrelation function $\rho_x(\tau)$, the joint probability density function of this ratio is given by

$$f_{x_{t+\tau}/x_t}(z) = \frac{\left(1-|\rho_x(\tau)|^2\right)/\pi}{\left(1+z_r^2+z_i^2-2\rho_{x_r}(\tau)z_r-2\rho_{x_i}(\tau)z_i\right)^2}, \tag{4.22}$$

where $z = z_r + jz_i$. Since (4.22) is a valid probability density function, finding the maximum

in the real or imaginary components requires us to evaluate

$$z_{\max} = \left.\frac{\partial f_{X_{t+\tau}/X_t}(z)}{\partial z_r}\right|_{z_r=0} + j\left.\frac{\partial f_{X_{t+\tau}/X_t}(z)}{\partial z_i}\right|_{z_i=0}. \tag{4.23}$$

Clearly, solving (4.23) produces the result of $z_{\max} = \rho_x(\tau)$. Thus, the mean is equal to the mode.

To determine that the distribution is symmetric about $z_{\max}$, and therefore establishing that the median is equal to the mean and mode, we evaluate (4.22) at the maximum plus some arbitrary offset, that is evaluate $f_{x_{t+\tau}/x_t}(\rho_x(\tau) + \Delta z)$. Doing so reveals that the probability density function becomes a function of the square of the real and imaginary parts of Δz,

$$f_{x_{t+\tau}/x_t}(\rho_x(\tau) + \Delta z) = \frac{\left(1 - |\rho_x(\tau)|^2\right)/\pi}{\left(1 - |\rho_x(\tau)|^2 + \Delta z_r^2 + \Delta z_i^2\right)^2}. \tag{4.24}$$

Thus, the probability density function is symmetric about $\rho_x(\tau)$, justifying the use of the median as a robust replacement for the mean.

We can now define the robust MRE for complex-valued Gaussian processes. Given a finite-length observation vector $\mathbf{x} = \{X_t, t = 1, 2, \ldots, N\}^T$ we define the vector $\mathbf{y}_{\mathrm{r}} = \{Y_{\mathrm{R}_t}, t = 1, 2, \ldots, N - \tau\}^T$ with elements that are realizations of the new random process given by

$$Y_{\mathrm{R}_t} = X_{t+\tau}/X_t, \tag{4.25}$$

consisting of pair-wise ratios of the original process X_t with some time offset τ. The RME is the result of the coordinate-wise sample median, $\hat{\xi}_{\frac{1}{2}}$, of the pair-wise ratios contained in $\mathbf{y}_{\mathrm{r}}$, yielding

$$\hat{\rho}_{x_{\mathrm{MRE}}}(\tau) = \hat{\xi}_{\frac{1}{2}}\left(\mathrm{Re}\{\mathbf{y}_{\mathrm{r}}\}\right) + j\hat{\xi}_{\frac{1}{2}}\left(\mathrm{Im}\{\mathbf{y}_{\mathrm{r}}\}\right). \tag{4.26}$$

4.4.2 Asymptotic Bias of MRE in the Presence of White Complex-Valued Gaussian Outliers

The robustness of the MRE is obtained from the sample median function. Therefore, in order to understand the bias of the MRE with replacement outliers, we need to determine the new marginal cumulative density function (CDF) of the resulting pair-wise ratios for the mixture model. Therefore, we now examine the generalized CDF for the ratio X_t/Y_t of correlated complex-valued Gaussian processes derived by Baxley *et al.* [4]. We first define their λ-function, which is expressed as

$$\lambda\left(z_i, \rho_r, \rho_i, \sigma_x, \sigma_y\right) = \frac{\left(\rho_i \sigma_x + \sigma_y z_i\right)}{\sqrt{\left(1-\rho_r^2\right)\sigma_x^2 + 2\rho_i \sigma_x \sigma_y z_i + \sigma_y^2 z_i^2}}, \tag{4.27}$$

where ρ_r and ρ_i are the real and imaginary parts of the correlation coefficient between X_t and Y_t. The marginal CDFs are

$$F_{x/y_{z_i}}\left(z_i, \rho_r, \rho_i, \sigma_x, \sigma_y\right) = \frac{1}{2}\left(\lambda\left(z_i, \rho_r, \rho_i, \sigma_x, \sigma_y\right) + 1\right), \tag{4.28}$$

and

$$F_{x/y_{z_r}}\left(z_r, \rho_r, \rho_i, \sigma_x, \sigma_y\right) = F_{x/y_{z_i}}(z_r, -\rho_i, -\rho_r, \sigma_x, \sigma_y). \tag{4.29}$$

When the pair-wise ratios are delayed pairs from a single Gaussian time series, X_t, then $F_{z_r}\left(z_r, \rho_{x_r}(\tau), \rho_{x_i}(\tau), \sigma_x, \sigma_x\right)\Big|_{z_r=\rho_r} = F_{z_i}\left(z_i, \rho_{x_r}(\tau), \rho_{x_i}(\tau), \sigma_x, \sigma_x\right)\Big|_{z_i=\rho_i} = 0.5$. This again shows that the coordinate-wise median of the ratio of the two complex-valued Gaussian processes is equal to the correlation coefficient.

Given a replacement outlier signal model, our outlier process, W_t, is a white Gaussian noise process that is uncorrelated to the original time series. The variance, σ_w^2, and contamination rate, ε, are analysis variables. Probabilistically, if the contamination rate is ε, we will observe a CDF of the new process created by the pair-wise ratios, given by

$$F_{g_{t+\tau}/g_t} = (1-\varepsilon)^2 F_{x_{t+\tau}/x_t} + \varepsilon(1-\varepsilon) F_{w_{t+\tau}/x_t} + (1-\varepsilon)\varepsilon F_{x_{t+\tau}/w_t} + \varepsilon^2 F_{w_{t+\tau}/w_t}. \tag{4.30}$$

Our estimate, $\hat{\rho}_{x_{\mathrm{RME}}}(\tau)$, is based on the sample median of the ε-contaminated distribution; therefore to find the bias we need to calculate

$$b_{\mathrm{RME}} = \left|\rho_x(\tau) - F^{-1}_{g_{t+\tau}/g_{t_r}}\left(\tfrac{1}{2}, \rho_r, \rho_i, \sigma_x, \sigma_w\right) - jF^{-1}_{g_{t+\tau}/g_{t_i}}\left(\tfrac{1}{2}, \rho_r, \rho_i, \sigma_x, \sigma_w\right)\right|. \tag{4.31}$$

The bias is seen to be a function of the contamination rate, the correlation coefficients (both magnitude and angle), and the variances of the original time series and the outlier process.

Two examples are now presented and analyzed to highlight some features of the MRE bias characteristics. Firstly, if the bias is calculated as a function of the angle, θ, of the correlation coefficient with all other parameters fixed, the bias will be symmetric about the real and imaginary axes, that is, $b_{\mathrm{MRE}}\left(|\rho|, \frac{n\pi}{2} + \theta\right) = b_{\mathrm{MRE}}\left(|\rho|, \frac{n\pi}{2} - \theta\right), n = 0, 1, 2, 3$. We display the bias curves in Fig. 4.5 for a correlation coefficient with magnitude of 0.5 and angles $\theta = 0, \pm\pi/8$ and $\pm\pi/4$. Secondly, inverting the ratio of the outlier to inlier variances produces the same bias. For instance, the bias is the same if the ratio of the variances of the outliers to the inliers is $\sigma_w^2/\sigma_x^2 = \kappa$ or $\sigma_w^2/\sigma_x^2 = \kappa^{-1}$. This is seen in Fig. 4.6, where the correlation coefficient is held fixed at 0.9, and $\kappa = 1$ and $1/100, 100$ and $1/10000$ and 10000. Also observed in Fig. 4.6, is that as κ deviates from unity, the bias increases for higher contamination rates and lowers slightly for small contamination rates. Finally, we note that maximizing the bias for a fixed correlation coefficient will require an outlier process with different variances for each contamination rate.

4.4.3 Variance of MRE Under the Gaussian Assumption

The distributional robustness of the MRE is obtained through the use of the median, an ordered statistic, ξ_p, where $p = 0.5$. Therefore, we must examine the asymptotic properties of the sample median function, $\hat{\xi}_{\frac{1}{2}}$, to determine its variance. Given a CDF that has a

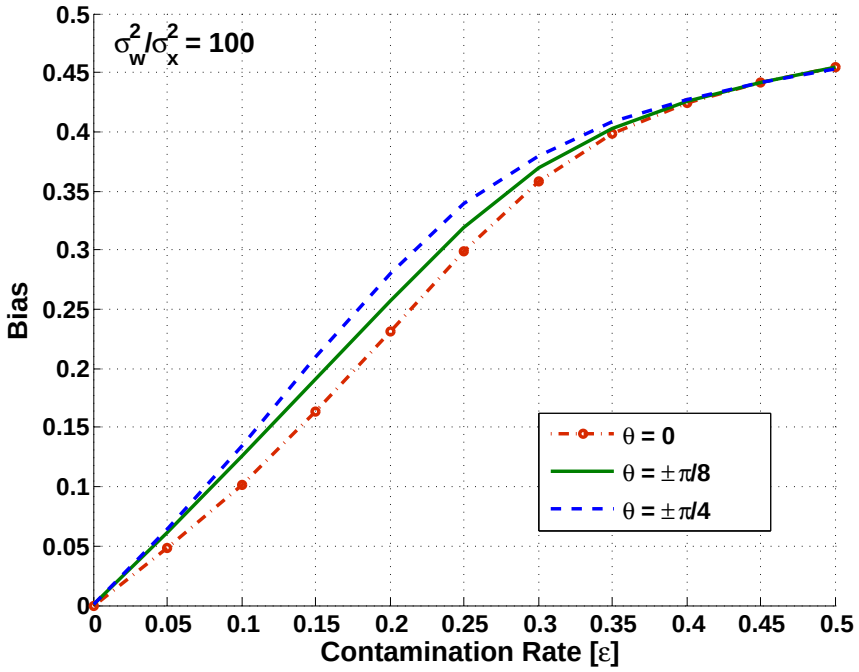

Figure 4.5: These bias curves are for the MRE with replacement outliers taken from a white complex-valued Gaussian process with a variance 100 times larger than the underlying AR(1) model with correlation coefficient magnitude equal to 0.5. The curves indicate a dependency on both the contamination rate and the angle of the correlation coefficient. The bias is a symmetric function of the angle with respect to both the real and imaginary axes.

density f in the neighborhood of ξ_p that is both positive and continuous at ξ_p, then

$$\hat{\xi}_{pn} \sim \text{AN}\left(\xi_p, \frac{p(1-p)}{nf^2(\xi_p)}\right), \tag{4.32}$$

where n is the number of observations [68]. Thus, setting $p = 0.5$, n equal to the number of observed pair-wise ratios, and using the derivatives of (4.28) and (4.29) we can use (4.32) to obtain the estimator's distribution in the real and imaginary components. The total asymptotic variance of the MRE is simply the sum of the real and imaginary variances, that is

$$\text{VAR}\{\sqrt{n}\hat{\rho}_{x_{\text{MRE}}}\} \quad = \quad \frac{1}{4[F'_{z_r}(\rho_{x_r}(\tau))]^2} + \frac{1}{4[F'_{z_i}(\rho_{x_i}(\tau))]^2} \quad = \quad 2\left(1 - |\rho_x(\tau)|^2\right). \tag{4.33}$$

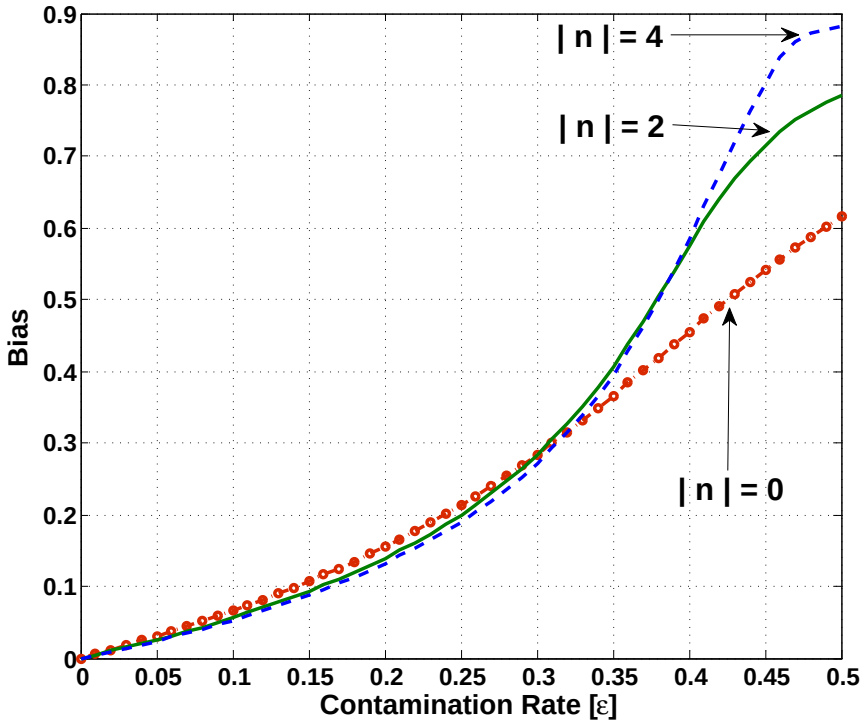

Figure 4.6: The MRE bias is plotted for AR(1) model with $\rho = .9$ and $\sigma_w^2/\sigma_x^2 = 10^n, |n| = 0, 2, 4$. The bias is observed to be the same for n and $-n$.

4.5 Computational Analysis

The computational complexity analysis of the PPC and MRE is complicated by the use of lookup tables, medians and special functions. We briefly assess their complexity with the understanding that hardware architecture may be an additional factor. The PPC requires six multiplies, five additions, two divisions and one square root per pairwise observation, and additionally one final table lookup. A savings can be realized using fast gridded-interpolation algorithms, reducing the cost of the PPC lookup function. We approximate the amount of work required to generate each of the MRE's pairwise ratios to be six multiplies, three additions and two divisions. Fast selection algorithms can be used to implement the MRE's two coordinate-wise medians. In comparison, the sample autocorrelation requires approximately six multiplies and six additions per pairwise observation. Based on these operation counts, and because of the two required medians, it is apparent that the sample autocorrelation

function is more computationally efficient than the PPC, which in turn is faster to execute than the MRE. This is demonstrated in Fig. 4.7, which provides example timing results of the MRE and PPC with respect to the sample autocorrelation function obtained using optimized MATLAB [47] implementations.

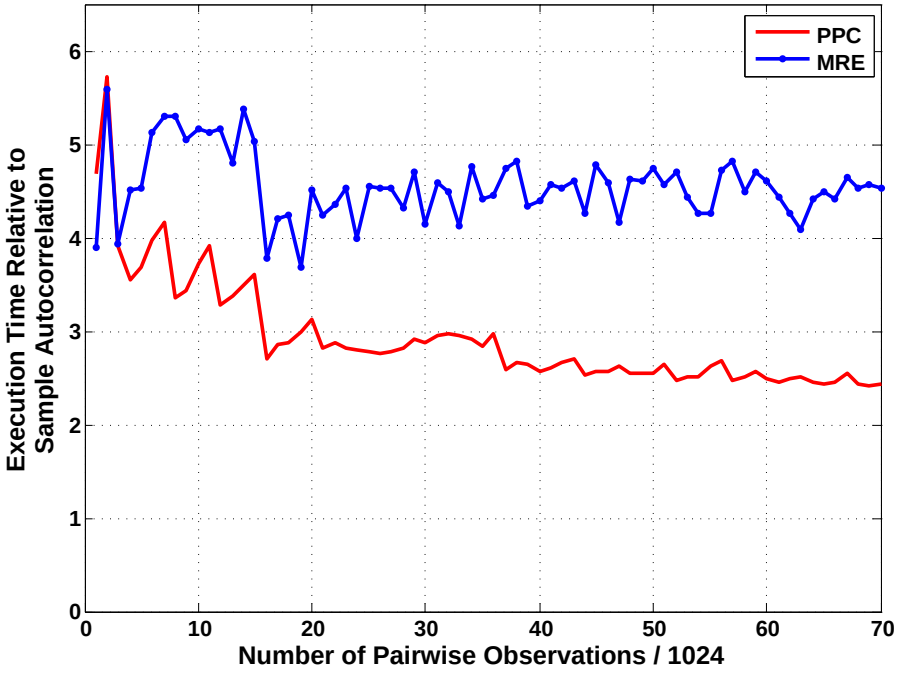

Figure 4.7: Timing comparisions of the PPC and MRE relative to the sample autocorrelation function. The PPC is typically the fastest of the two nonlinear estimators.

An additional computational consideration that one might consider is that the PPC can operate on subblocks of an observation vector independently. That is, data can be divided into smaller blocks of data and the nonlinear function can be applied to each observation. A running sum or mean can be maintained on all of the separate subblocks and combined for a final result. This may allow one to take advantage of efficient data sizes for specific hardware architectures. The MRE will require that the entire data vector of ratios be processed as one due to the median operations. These implementation details might impact which estimator one wishes to implement if computation time is of importance.

4.5.1 Comparison to the Robust Schweppe-Type GM-Estimator With Huber Function

In this section, we compare the performance of the PPC and MRE with the well-known SHGM. Our comparison concentrates on estimator variance at the assumed Gaussian distribution, and bias under the ε-contaminated replacement outlier model. We have derived the bias for both the PPC and MRE in the presence of replacement outliers modeled as a white complex-valued Gaussian process. For the MRE, we must resort to a Monte Carlo technique to obtain the worst-case bias given this contamination model because there are too many parameters involved to yield an exact expression as was the case for the PPC.

Estimator Variance Comparison

Theoretical- and Monte-Carlo-derived estimator variance curves in Fig. 4.8 reveal the cost of robustness, that is, a decrease in estimator efficiency at the assumed Gaussian distribution. As expected, the variance of the sample autocorrelation function is the lowest of all of the examined estimators. Thus, as robustness is added, we will notice an increase in estimator variance. We first discuss the two variants of the robust SHGM estimator. The 'SHGM' curve corresponds to the version of the estimator where $w(\mathbf{h}_i)$ is only a function of the observations in the explanatory variable. This one dimensional weighting is a function of the magnitude of the explanatory variables. The efficiency of this estimator is approximately 0.95. We noticed that robustness, improved bias performance, can be achieved by the SHGM by incorporating pairwise projection statistics. The cost of this is an increased variance as indicated by the 'SHGM Pairwise PS' curve. The efficiency in this case drops to approximately 0.85. We highlight again that this version of the SHGM estimator will completely breakdown when the time series contamination rate equals 0.3, as a result of the breakdown of the projection statistics.

The PPC estimator is observed to be the next best estimator among the three robust esti-

mators investigated from a variance performance perspective. The theoretical variance for independent pairs of correlated observations, labeled 'PPC-Case 1' in Fig. 4.8, is the lower bound for the PPC. We notice through Monte Carlo simulations that a time series application of the PPC estimator, such as an AR(1) model labeled ' PPC-Case 2', equals this curve when the correlation coefficient is zero, an expected result since the observations are uncorrelated. As the magnitude of the correlation coefficient increases, the variance of the PPC estimator for time series increases from this lower bound. At large magnitudes of the correlation coefficient, the variance of the PPC approaches that of the MRE. The efficiency of the PPC ranges from approximately 0.62 at zero correlation to approximately 0.5 at a correlation magnitude of 0.8. At higher correlations, the efficiency appears to be slightly less than the MRE.

The MRE's variance is in general the highest of the examined estimators. From (3.18) and (4.33) we note that the variance of the MRE for the AR(1) model is twice that of the sample autocorrelation function. Therefore, the MRE's efficiency is exactly 0.5 for all values of $\rho_x(\tau)$. The real-valued MRE estimator was shown to have an efficiency of only 0.4 in [15]. We observe that the Monte Carlo results match the theoretical curve for the MRE, validating both (4.33) and our Monte Carlo simulations.

Finally, we see as the correlation coefficient becomes large in magnitude, the difference among the estimators' variances becomes smaller. In fact, the variances of the PPC and MRE for the AR(1) model are essentially the same once the correlation magnitude exceeds 0.8.

Estimator Bias Comparison

Results are shown for two test cases in Fig. 4.9 and Fig. 4.10 where the correlation coefficient is equal to 0.5 and 0.9, respectively. The results are fairly similar at lower contamination rates, and quickly diverge at higher rates. There are a few subtle differences that deserve discussion. Examining the bias results for the SHGM indicates that incorporating pairwise projection statistics improves the bias for low contamination rates. However, this is obtained

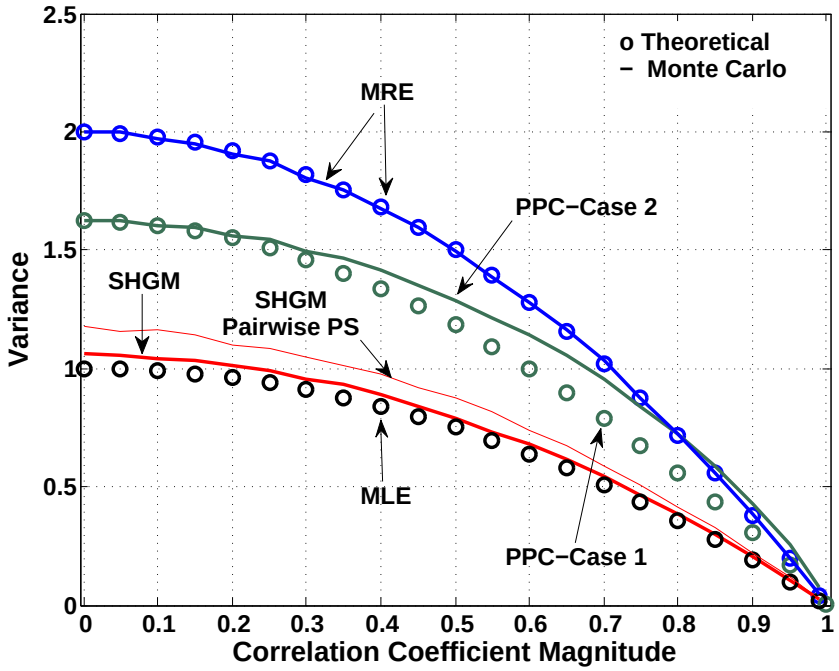

Figure 4.8: Theoretical (circle marker) and Monte-Carlo-derived (solid line) variance curves of normalized estimators are shown for a complex-valued Gaussian AR(1) model with increasing correlation. The results indicate that efficiency is sacrificed for robustness. The MRE and PPC have approximately the same variance when the correlation is strong. The 'PPC-Case 1' curve shows the variance for independent pairs of correlated samples while 'PPC-Case 2' is for the AR(1) time series model. Note that the estimators have been multiplied by $\sqrt{n}$ when assessing their variance

at a cost of estimator breakdown at a time series contamination rate of approximately 0.3. This occurs because half of the pairwise observations used by the projection statistics are contaminated when the time series contamination rate is just 0.3 as discussed earlier. Due to this, we will not include the results for this estimator in further analysis.

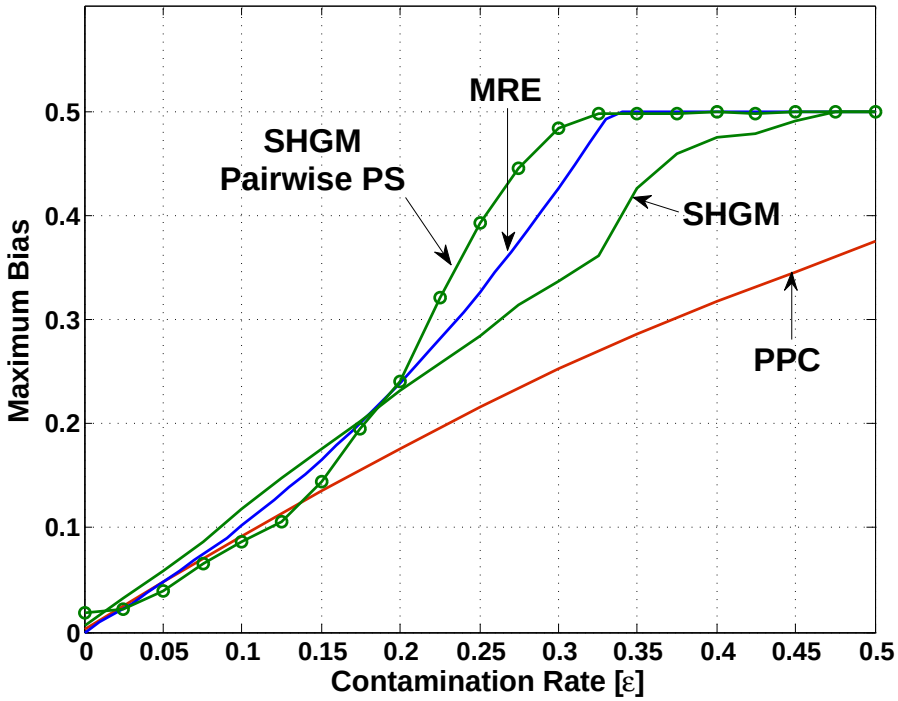

Figure 4.9: Maximum bias curves for AR(1) model with $\rho = 0.5$ using white complex-valued Gaussian replacement outliers. The estimators all have approximately the same bias performance at lower contamination rates ($\varepsilon < 0.15$). At higher contamination rates the PPC estimator performs the best, with the MRE and PS- SHGM estimators breaking down near $\varepsilon = 0.3$. The pairwise PS improve the SHGM until the time series contamination rate reaches 0.3, causing the projections statistics to break down.

As for the MRE estimator, it experiences break down at contamination rates of approximately 0.325 and 0.475 for $\rho = 0.5$ and $\rho = 0.9$ respectively. It should be noted that the ratio of the outlier to inlier variances required for this to happen is either extremely large or extremely small and may not be common in practice. We see that the PPC is the only estimator that does not completely break down under this replacement outlier model. Additionally, the variance of the replacement outlier model has no impact on the PPC's bias.

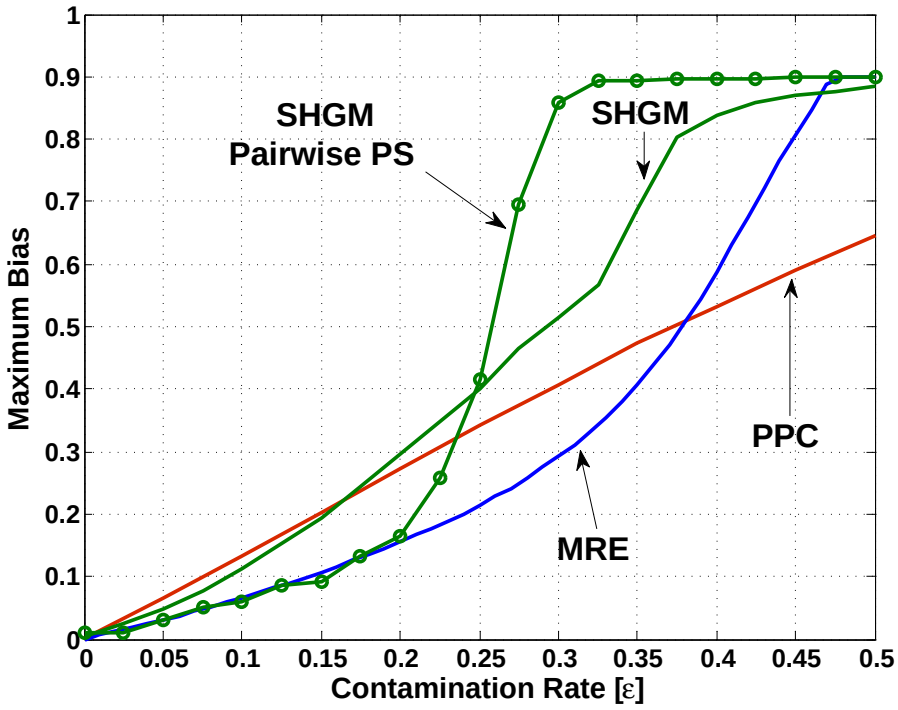

Figure 4.10: Maximum bias curves for AR(1) model with $\rho = 0.9$ using white complex-valued Gaussian outliers. The SHGM using pairwise projection statistics and the MRE perform the best at lower contamination rates. Both versions of the SHGM essentially break down at a contamination rate above 0.35. The PPC is the only estimator to not completely break down.

4.6 Iteratively Reweighted Phase-Phase Correlator and Median of Ratios Estimators

4.6.1 Algorithm Description

The PPC and MRE have been shown to be robust to white impulsive outliers. This robustness is obtained through the use of a memoryless nonlinear function of the observations. The PPC estimator gains robustness by discarding the magnitude and operating only on the phase differences of the observations, whereas the MRE obtains robustness through the median operator. As we have presented, the observed bias of these estimators compares quite favorably to the SHGM estimator.

Our previous asymptotic bias analysis used the replacement outlier model. This was done to enable tractable analysis, and was noted to be a more stressing scenario than the additive outlier model. However, our intent is to use these estimators primarily in the case where additive impulsive noise is present. We therefore need to assess the difference in the maximum bias under the additive model. Figure 4.9 and 4.10 show the worst case bias obtained from the replacement outlier model for Gaussian outliers. The same analysis is presented below, focusing only on the MRE and PPC and now shows the difference between the replacement and additive outlier models. Figure 4.11 reveals that the replacement model induces only slightly more bias than the additive model.

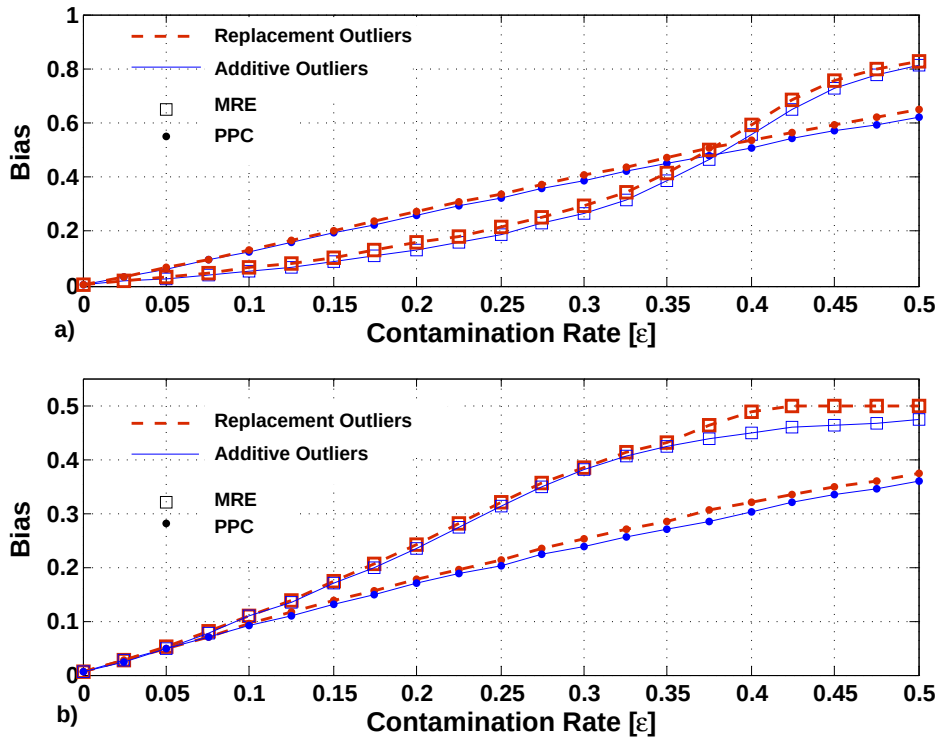

Figure 4.11: A comparison of MRE and PPC maximum bias for an AR(1) model with a) $\rho = 0.9$ and b) $\rho = 0.5$ is shown. The white complex-valued Gaussian outliers follow both the additive and replacement model. The replacement model is only slightly worse than the additive model.

The primary difference in the additive outlier model is that lower variance noise sources impact the estimators less than in the replacement model. Therefore, the estimator is more

susceptible to larger noise sources than smaller ones. Large variance noise sources are typically easier to identify than smaller ones. With this in mind, we have chosen to incorporate a downweighting method to improve these estimators for the additive model, while noticing improvement in the replacement model for all but the highest contamination rates as well.

We know that the sample autocorrelation function is not robust. However, implementing the SHGM with (3.34) illustrates that robustness is obtained when the abnormal observations are downweighted appropriately. Based on this observation, we propose to improve the already robust PPC and MRE estimators by adding an outlier weighting mechanism through an iterative approach. This technique improves bias especially as the variance of the outlier process increases.

The downweighting function used by our new iterative PPC and MRE estimators utilizes robust Mahalanobis distances derived from robust correlation estimates. Using these weights, the sample mean and median functions used by the PPC and MRE estimators are replaced with a weighted sample mean and weighted sample median [22] respectively. Additionally, we estimate the variance of the time series, which is incorporated by the Mahalanobis distances, with a scaled weighted median of the magnitude of the normally distributed complex-valued observations. We start by defining the iteration loop for our new estimators. Individual steps are then discussed in detail.

Let us describe the algorithm shown in Table 4.6.1. Step 0 initializes the weights to one. Steps 1-5 are repeated for as many iterations as desired or until the change of the estimate is smaller than some predefined amount. The first estimate of correlation, which results from the initial execution of Step 1, does not incorporate any downweighting and is equivalent to the definitions of the PPC and MRE estimators in (4.11) and (4.26).

Step 2 performs a robust estimation of the variance of X_t which is needed to generate our estimate of the covariance matrix. It is assumed that X_t is a zero-mean, circularly-symmetric, complex-valued Gaussian random variable, and therefore its magnitude is Rayleigh distributed. If σ is defined to be the standard deviation of both the real and imaginary

Table 4.1: Steps for New Iterative Estimators

0) Initialize $\mathbf{w}_{k=0} = \mathbf{1}$	(4.34)
1) Calculate $\hat{\rho}_{x_k}(\tau) = \hat{\rho}_{x,\text{robust}}(\mathbf{x}, \mathbf{w}_{\mathbf{k-1}})$	(4.35)
2) Calculate $\hat{\sigma}^2_{x_k} = \frac{2}{\ln(4)} \left[\hat{\xi}_{\frac{1}{2}}(\lvert\mathbf{x}\rvert, \mathbf{w})\right]^2$	(4.36)
3) Calculate $\hat{\mathbf{\Sigma}}_k = \hat{\sigma}^2_{x_k} \begin{bmatrix} 1 & \hat{\rho}_{x_k}(\tau) \\ \hat{\rho}^*_{x_k}(\tau) & 1 \end{bmatrix}$	(4.37)
4) Calculate $\text{RMD}_{i_k} = \sqrt{\mathbf{x}_i^H \hat{\mathbf{\Sigma}}_k^{-1} \mathbf{x}_i}$	(4.38)
5) Calculate $w_{i_k} = \min\left[1, \chi^2_{4,0.975} / \left(2\text{RMD}^2_{i_k}\right)\right]$	(4.39)
6) Stop if $(\lvert\hat{\rho}_{x_k}(\tau) - \hat{\rho}_{x_{k-1}}(\tau)\rvert < \delta)$ or if $(k = k_{\max})$; Otherwise $k = k+1$, Go to Step 1	

components, then the median of magnitude of X_t is equal to $\sigma\sqrt{\ln(4)}$. Using this property, we can estimate the variance using the sample median. To improve our estimate, we replace the commonly used sample median with a weighted sample median. The weighted sample median is a weighted ordered statistic defined by

$$\hat{\xi}_{\frac{1}{2}}(\mathbf{x}, \mathbf{w}) = \sum_{x_i < x_k} w_i < 0.5 \text{ and } \sum_{x_i > x_k} w_i \leq 0.5, \tag{4.40}$$

where w_i are the weights associated with the ordered observations x_i.

Our robust estimate of $\mathbf{\Sigma}$ is created from a robust measure of dispersion and robust estimates of the correlation coefficient as

$$\hat{\mathbf{\Sigma}} = \hat{\sigma}^2_{x_k} \begin{bmatrix} 1 & \hat{\rho}_{x_k}(\tau) \\ \hat{\rho}^*_{x_k}(\tau) & 1 \end{bmatrix}. \tag{4.41}$$

Next, we review our construction of the Mahalanobis distance, which uses the robust covariance estimate and is defined by (4.38). It is well-known that the Mahalanobis distance is susceptible to the masking affect of multiple outliers. Therefore, we generate robust Maha-

lanobis distances based on robust estimates of $\mathbf{\Sigma}$.

The covariance matrix captures the linear dependence between the observations found in $\mathbf{x}_i$ given by

$$\mathbf{x}_i = \{x_{\tau+i}, x_i\}^T, i = 1, \ldots, N - \tau, \tag{4.42}$$

which contain pairs of time delayed observations. Given that each of our $\mathbf{x}_i$ vectors are of length $n = 2$, and the assumption that the underlying random process is complex-valued Gaussian, then as discussed in Section 3.5.3, two times the square of the Mahalanobis distances will follow a chi-squared distribution with four degrees of freedom, that is $2MD_i^2 \sim \chi_4^2$. This is slightly different than Mahalanobis distances for real-valued Gaussian random variables, where the squared values follow a Chi-squared distribution with two degrees of freedom. The distribution for the complex-valued case is obtained by noting that a true covariance matrix, $\mathbf{\Sigma}$, is positive semidefinite. This allows us to rewrite the square of the Mahalanobis distance as ${MD_i}^2 = \mathbf{x}_i^H \hat{\mathbf{\Sigma}}^{-1/2} \hat{\mathbf{\Sigma}}^{-1/2} \mathbf{x}_i$. The operation $\mathbf{x}_i^H \hat{\mathbf{\Sigma}}^{-1/2}$ whitens our observations producing unit variance uncorrelated complex-valued Gaussian random pairs. Therefore, the squared Mahalanobis distance is the inner product of n complex-valued unit-variance Gaussian random variables, or the sum of the magnitude squared of $2n$ Gaussian random variables. Multiplying the result by a factor of two transforms the results into a chi-squared distributed random variable with $2n$ degrees of freedom. Observations exceeding the 97.5 percentile are flagged as outliers and downweighted using the weighting function (4.39) generated by Step 5.

The output of a complex-valued Gaussian AR(1) model with feedback coefficient equal to 0.9 was used to verify the distribution of the Mahalanobis distances. Figure 4.12 contains the theoretical probability distribution of a Chi-squared random variable with four degrees of freedom. Additionally, the probability density function estimated from the histogram of two times the square of the robust Mahalanobis distances generated in (4.38) using the IPPC with ten iterations is shown. Very strong agreement between the theoretical and estimated probability density function verify our threshold in (4.39) is set correctly. The results for

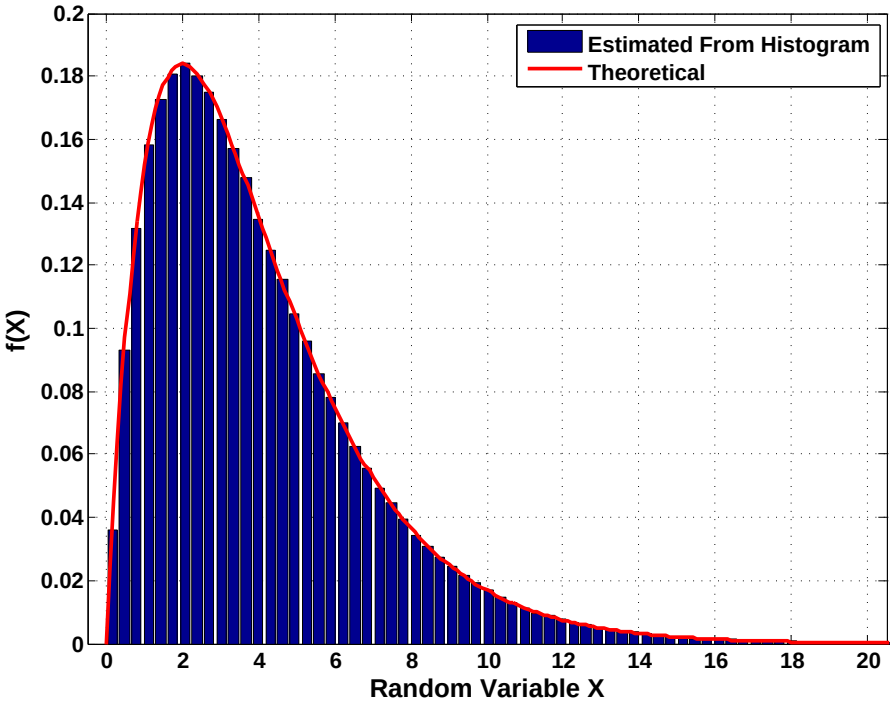

Figure 4.12: Theoretical and estimated probability density functions for twice the robust Mahalanobis distances created by the IPPC and IMRE. The estimated distribution used distances obtained after ten iterations.

the IMRE are similar.

Finally, we highlight that with the gain in robustness, we also gain a significant improvement in execution speed over the SHGM using pair-wise projection statistics. In fact after very careful implementation of the SHGM and robust projection statistics in MATLAB, the IPPC and IMRE (10 iterations) realized approximately 5 and 10, 16 and 20 and approximately 40 times faster execution speed than the SHGM using robust projection statistics for a sample sizes of $N = 1000, 2000,$ and 4000. The improvement trend appears to continue as N grows.

We have found that using pairwise samples is very effective at detecting and downweighting outliers, leading to very good robust estimates of the correlation coefficient. However, if one wishes to use longer vectors of $\mathbf{x}_i$ in (4.38), the same technique may be used. One just needs to adjust the threshold in (4.39) and expand the covariance matrix estimate to include additional lags of the robust correlation coefficient estimate.

Incorporating the weighting function generated in (4.39) into the robust PPC and MRE is straightforward. Downweighting is introduced into the PPC by replacing the sample average defined in (4.10) with a weighted average as follows

$$\hat{\rho}_{y_{\text{s-weighted}}}(\tau) = \frac{\sum_{n=1}^{N-\tau} y^*_{\text{s}_n} y_{s_{n+\tau}} w_n}{\sum_{n=1}^{N-\tau} w_n}. \tag{4.43}$$

The PPC based estimate of $\rho_x(\tau)$ is still obtained using (4.11), but it is now a function of (4.43), yielding

$$\hat{\rho}_{x_{\text{PPC-weighted}}}(\tau) = e^{j\text{Arg}[\hat{\rho}_{y_{\text{s}}\text{-weighted}}(\tau)]} f^{-1}\left[|\hat{\rho}_{y_{\text{s}}\text{-weighted}}(\tau)|\right]. \tag{4.44}$$

The MRE estimator defined in (4.26) is modified such that it employs a coordinate-wise median that is augmented with the weighted coordinate-wise median introduced in (4.40). Formally, we have

$$\hat{\rho}_{x_{\text{MRE-weighted}}}(\tau) = \hat{\xi}_{\frac{1}{2}}\left(\text{Re}\{\mathbf{y}_{\text{r}}\}, \mathbf{w}\right) + j\hat{\xi}_{\frac{1}{2}}\left(\text{Im}\{\mathbf{y}_{\text{r}}, \mathbf{w}\}\right). \tag{4.45}$$

4.6.2 Improved Bias Results

The benefit of the iterative approach is displayed by the following two examples. A unit variance, complex-valued Gaussian AR(1) process is created with a feedback coefficient equal to 0.9. Outliers are generated from a white complex-valued Gaussian random process with a variance equal to 15 dB greater than the AR(1) process with a contamination rate of $\varepsilon = 0.3$. The first case uses the replacement outlier model, while the second uses the additive outlier model. Figure 4.13 shows the MRE and PPC bias and variance estimate at iteration k for both outlier models. This example highlights several features of the iterative estimators. One, significant improvement in the estimator bias is achieved over the non-iterative approach. This is observed by comparing the results from iteration zero (non-iterative result) and ten. Additionally, it is seen that the estimators converge very quickly.

In fact, we see very little improvement after five iterations. Finally, it is noted that this technique produces low-bias estimates even at a high contamination rate for a impulsive noise model with a fairly large variance.

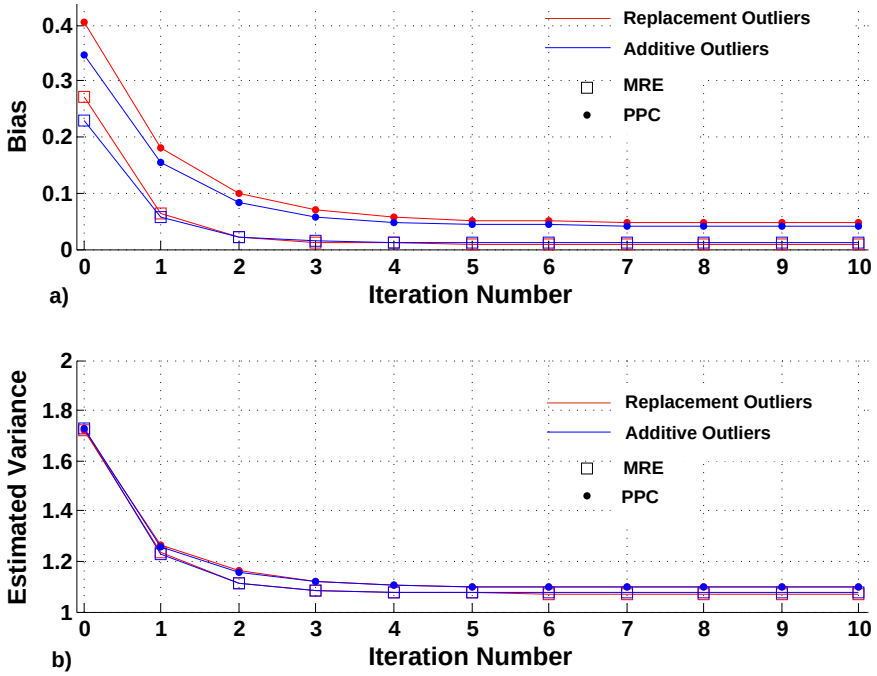

Figure 4.13: The convergence of a) the bias and b) the estimated time series variance is shown per iteration, indicating both improved performance and fast convergence.

A final analysis is performed on these iterative estimators to analyze the worst-case additive-outlier bias given W_t is Gaussian. The Monte Carlo analysis is again leveraged. Figure 4.14 shows the resulting worst-case bias in the presence of additive Gaussian outliers for the non-iterative and iterative PPC and MRE estimators. Additionally, the resulting SHGM bias is shown. The iterative versions of the PPC and MRE have smaller bias than the non-iterative versions. It is also observed that they outperform the SHGM estimator from a bias perspective.

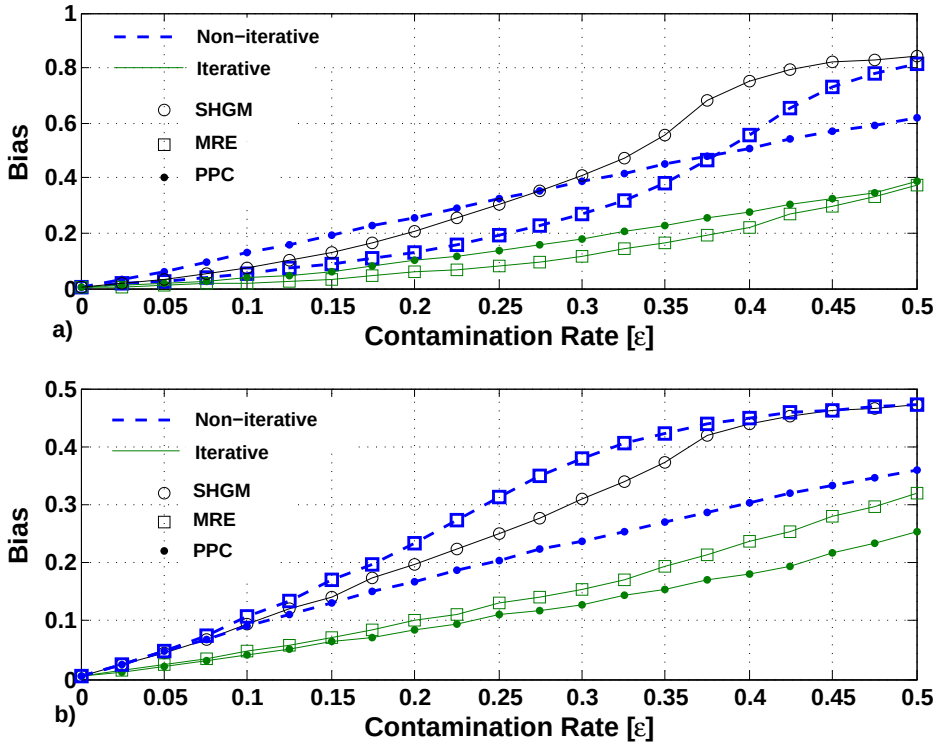

Figure 4.14: Improved bias in the presence of additive outliers is realized for the iterative PPC and MRE with better performance than the SHGM for both a) $\rho = 0.9$ and b) $\rho = 0.5$.

4.6.3 Examination of Sample Support Needed for Robust Mahalanobis Distances

The interest in these new estimators stems from their robustness and fast execution time for large data sizes, for example greater than 1,000 samples. All of the results shown thus far are asymptotic, that is for large N. For our analysis, N was chosen to be 1000 samples. This is not a large observation size given the speed of modern computing processors. In fact, by today's standards one might even consider this to be a small observation length. A crucial step in our iterative estimation process is the creation of robust Mahalanobis distances. It is known that for large sample sizes, that the distribution of Mahalanobis distances based on the sample autocovariance function asymptotically follows a χ^2 distribution. However, for small sample sizes, this is not true. For example, for small sample sizes the square of the Mahalanobis distances follow the non-central F-distribution.

Given that our proposed iterative estimators use nonlinear estimators to generate the covariance matrix, and subsequently the robust Mahalanobis distances, we should not assume that the distributions that describe the sample Mahalanobis distances to be applicable. However, since our robust estimators are asymptotically unbiased, we believe that the Mahalanobis distances generated from these estimators should in fact be χ^2 as the number of observations becomes large. The underlying question is how large does N have to be to give us confidence that the χ^2 distribution is an appropriate choice to indicate outliers.

To provide credence that the χ^2 distribution is a valid approximation for data lengths of interest, we analyzed the observed distribution of the squared Mahalanobis distances generated by our estimators for sample sizes that are smaller than that which we intend to use. Q-Q

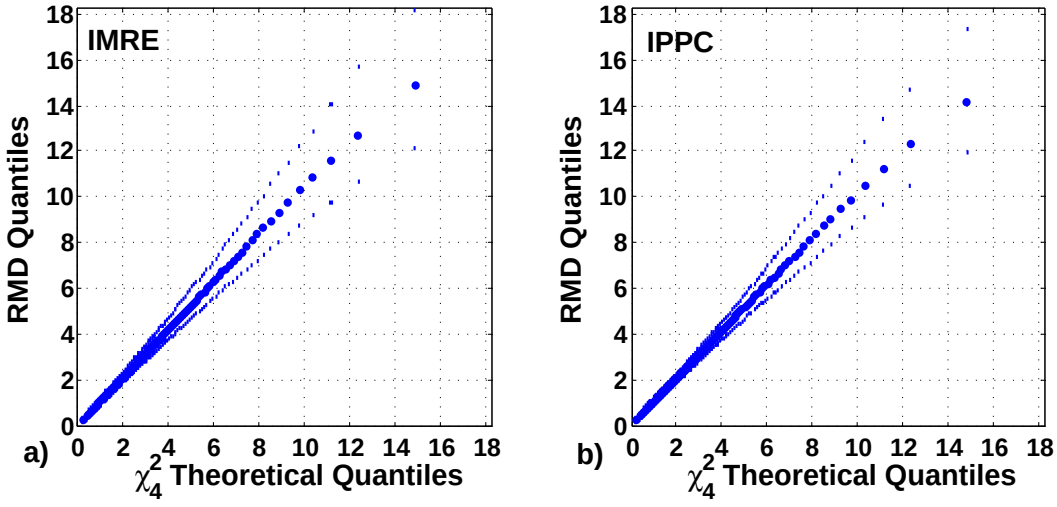

Figure 4.15: Q-Q plots and interquantile ranges of twice the RMD indicating excellent distribution approximation for both a) IMRE and b) IPPC obtained by taking the sample median of 1000 Monte Carlo simulations for data of length $N = 100$, using 5 iterations for correlation coefficient $\rho = 0.85e^{j\pi/4}$.

plots (Q stands for Quantile) in Fig. 4.15 are used to validate that 2RMD_i^2 approximately follow a χ^2_4 distribution. A straight-line, unit-slope Q-Q plot will reveal that the observed

data follows the assumed distribution. To be conservative, we chose to demonstrate this property for sample sizes smaller than our intended application. One thousand Monte Carlo simulations were performed on data of length $N = 100$, generating the RMD from both the IPPC and IMRE. We see that only the last few quantiles deviate from the unit-slope line, indicating that the χ^2_4 distribution is a very good approximation of twice the RMD. We note similar results when $N = 1000$, but with smaller interquantile ranges. All of our asymptotic bias simulation curves used $N = 1000$.

Chapter 5

Robust Parametric Spectral Estimation

From a robustness viewpoint, parametric spectral estimators are attractive because there is no need to perform a data cleaning procedure. Robust estimates of the autocorrelation function or similar quantities, such as reflection coefficients, can be used directly to estimate the system's parameters. To perform parametric spectral estimation, one must first choose a model to explain the observed data. For linear time-invariant systems, the discrete-time difference equation (2.1) is an appropriate choice. The transfer function for this signal model is given by

$$H(z) = \frac{\sum_{q=1}^{Q} \theta_q z^{-q}}{1 - \sum_{p=1}^{P} \phi_p z^{-p}}. \tag{5.1}$$

Given that the input, U_t, is a white random process with finite variance σ_u^2, the power spectral density of the observed output, X_t, is obtained by evaluating and appropriately scaling (5.1) along the unit circle in the complex plane, yielding

$$P(\omega) = \sigma_u^2 \left| H(z) \right|^2_{z=\exp(j\omega)}. \tag{5.2}$$

Given this model, we replace the parameters σ_u^2, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with the estimates $\hat{\sigma}_u^2$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$, which are created from data observations. This yields our parametric spectral estimator expressed as

$$\hat{P}(\omega) = \hat{\sigma}_u^2 \left| \frac{\sum_{q=1}^{Q} \hat{\theta}_q z^{-q}}{1 - \sum_{p=1}^{P} \hat{\phi}_p z^{-p}} \right|^2_{z=\exp(j\omega)} . \tag{5.3}$$

This approach has the advantage of potentially producing super-resolution results, overcoming the spectral leakage problem that plagues Fourier-based spectral estimation techniques caused by data windowing, which can be a significant limitation for small-length data records.

Of course, an appropriate model order must be selected in order to produce accurate power spectral density estimates. There are several well-known order selection methods available in the literature, for example Akaike [3] and Gray [32], and will not be the focus of this research.

5.1 Motivation for Using a Simplified Version of the Discrete Time-Difference Equation

The procedures for estimating the parameters of the ARMA(p,q) model are quite involved. Maximum likelihood estimators of ARMA parameters, which have desirable asymptotic properties, require iterative, nonlinear optimizations. Robust goodness of fit measurements must be made to assure that the correct order, p and q, are chosen to ensure that the data fits the chosen model. Often, for purposes of identifying outliers and creating linear predictors and estimating the power spectral density, we desire a simpler model leading to less complex and faster model estimation methods.

We seek a representable model for an ARMA process, sufficient for a majority of applications. Fortunately, if the autocorrelation function of an ARMA process tends to zero, that is $\rho_x(h) \to 0$ as $\tau \to \infty$, then for any k there is a causal AR(k), $q = 0$, process with autocorrelation function at lags $\{0, 1, \ldots, k\}$ that coincides with $\rho_x(j), j = 0, 1, \ldots, k$. Under

these conditions, we can use an AR model to represent the transfer function of an ARMA process; see Section 3.3 and 8.1 in [9] for more details. In general, it is easier to estimate the parameters of an AR process than the general ARMA process [9]. For this reason, there are several AR spectral estimators used in practice.

The Yule-Walker equations combined with the Levinson-Durbin recursion and the Burg method are two techniques commonly applied to estimate the parameters of an AR(p) model. These are fast algorithms that solve a system of linear equations. The autocorrelation and partial autocorrelation functions are the statistical quantities used by these two methods. The Burg algorithm will be the method of choice in this work. It is often preferred to the Yule-Walker approach because it produces stable model estimates when the pole locations of the observed system are close to the unit circle [10]. Our power spectral density estimates will use a simplified version of (5.3), namely

$$\hat{P}(\omega) = \hat{\sigma}_u^2 \left| \frac{1}{1 - \sum_{n=1}^{p'} \hat{\phi}_n z^{-n}} \right|^2_{z=\exp(j\omega)}, \tag{5.4}$$

where typically it is necessary to choose $p' > p$.

5.2 Classical Implementation of the Burg Method

In Section 2.2.5 the relationships exploited by Burg to develop estimates of the autoregressive coefficients were summarized. In practice, the quantities discussed must be estimated from observed data. Under the assumption of Gaussianity, all of the expectations are replaced with sample averages. The implementation of this method is summarized next.

To estimate the l^{th} reflection coefficient, the previous order, $l-1$, prediction error processes are used. Therefore, to start, we must initialize our estimates of the forward and backward

error processes given our length-N observation vector $x[n]$. We do this by setting

$$\hat{e}_0^f[n] = x[n], n = 1, 2, \ldots, N-1 \tag{5.5}$$

$$\hat{e}_0^b[n] = x[n], n = 0, 1, \ldots, N-2. \tag{5.6}$$

With the prediction errors intialized, we can iteratively update the forward and backward prediction weights of increasing length. To estimate the l^{th} reflection coefficient we use the $l-1$ order prediction error processes expressed as

$$\hat{\kappa}_l = \frac{2\sum_{n=l}^{N-1} \hat{e}_{l-1}^f[n]\hat{e}_{l-1}^{b^*}[n-1]}{\sum_{n=l}^{N-1}\left(\left|\hat{e}_{l-1}^f[n]\right|^2 + \left|\hat{e}_{l-1}^b[n-1]\right|^2\right)}, \tag{5.7}$$

where $\hat{e}_{l-1}^f[n]$ and $\hat{e}_{l-1}^b[n]$ are the current realizations of the estimated prediction error processes $E_{t,l-1}^f$ and $E_{t,l-1}^b$. The prediction weights are given by (2.20). Note that we have $w_{f1}^{(1)} = 1$.

We now need to update the error processes. This is accomplished by

$$\hat{e}_l^f[n] = \hat{e}_{l-1}^f[n] + \hat{\kappa}_l \hat{e}_{l-1}^b[n-1] \qquad n = l+1, l+2, \ldots, N-1, \tag{5.8}$$

$$\hat{e}_n^b[n] = \hat{e}_{l-1}^b[n-1] + \hat{\kappa}_l^* \hat{e}_{l-1}^f[n] \qquad n = 1, l+1, \ldots, N-2. \tag{5.9}$$

We execute this procedure $l = L$ times, which produces our order-L linear forward and backward prediction weights, corresponding to the our estimate of the autoregressive parameters $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_L\}$.

The Burg method is guaranteed to produce pole estimates with magnitude less than or equal to one [42, 10]. This is important for stability of the AR model, and allows reliable spectral estimates when the observed data is from a stable system with poles that are close to the unit circle.

5.2.1 Parametric Spectral Estimation Example Based the Burg Method

We now provide a parametric spectral estimation example based on the Burg method. The signal model is admittedly created to be challenging; one that will stress the estimators. The data contains five band-limited complex-valued Gaussian random processes equally spaced in frequency centered on $f_c = (-0.8, -0.4, 0, 0.4 \text{ and } 0.8) \times \pi$ rad / sample. The power levels vary, with the weakest signal being 54 dB below the strongest signal. Additionally, the two weakest signals are half the bandwidth of the strongest three signals. An example spectral estimate is displayed in Fig. 5.1. The signal model was estimated using the Burg algorithm with an autoregressive model order of $P = 200$. The spectral estimate displayed is an average of thirty Monte Carlo simulations using data vectors of length 2048.

5.2.2 Impact of Outliers

In Chapter 2, we showed that the reflection coefficient was equal to the partial correlation coefficient (2.27). Upon examination, it is observed that this definition is similar to our definition of the general correlation coefficient between two random process given by (2.3) with some differences in indexing. In fact, our definition implies that the conjugate of the reflection coefficient is the correlation between $E^f_{t,L-1}$ and $E^b_{t+1,L-1}$. With this observation, it is not surprising that we can estimate the reflection coefficient with (2.35). The primary difference between this estimate and the sample correlation estimate, other than the conjugate, is the normalization factor. Given the similarities to the sample correlation function, it is reasonable to conjecture that the reflection coefficient estimator is not robust to impulsive observation noise. This will be shown next.

The Burg method relies on estimating the reflection coefficient each iteration. While it is easy to see that a large magnitude observation will bias (5.7) towards zero, there is another deleterious effect of the outliers. For example, suppose the original observation vector con-

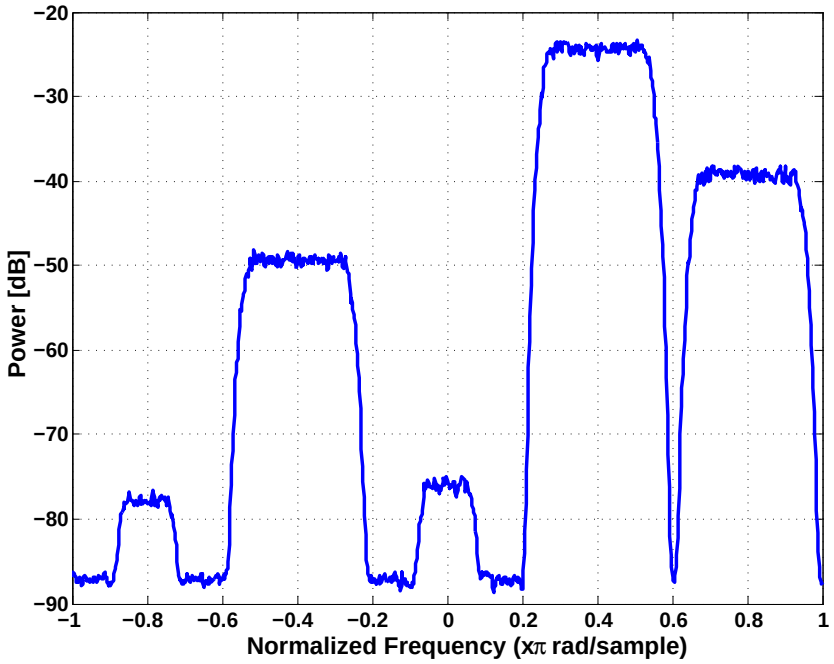

Figure 5.1: Parametric spectral estimation example using the classical Burg method applied to purely Gaussian random process.

tains a single outlier. Not only will it bias the estimate of the reflection coefficient but it will also bias the estimate of the updated forward and backward prediction error processes (5.8) and (5.9). Each update essentially duplicates the outlier. Therefore the number of corrupted samples doubles every time the forward and backward error processes are updated. As such, a small contamination rate can overwhelm the Burg method when the AR order is large.

Suppose that we are able to robustly estimate the reflection coefficient using some robust estimator given some contamination rate. Updating the forward and backward error processes will potentially double the contamination rate of the new processes. Therefore, for each iteration, the robust estimator of the reflection coefficient will have to contend with a higher contamination rate. Therefore, the Burg method is very susceptible to outliers,

requiring very robust estimators of correlation.

To illustrate how susceptible the Burg method is to even a small amount of outliers, our previous example is repeated with an additive replacement outlier model with a contamination rate of only 0.01. The outlier process was a white complex-valued Gaussian process with a variance that was ten times that of the underlying process. Example histograms of the real and imaginary components of the original and contaminated data vector for one realization are displayed in Fig. 5.2 and Fig. 5.3. From the histograms, we observe that the outliers are neither inordinately large in magnitude or quantity. However, we will see that the classic Burg algorithm suffers severe bias.

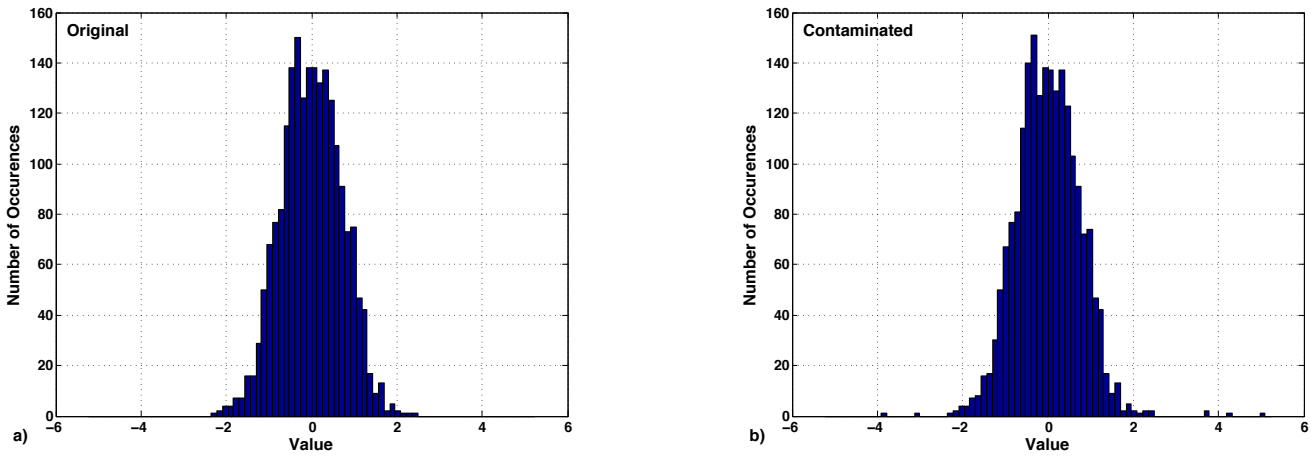

Figure 5.2: Histograms of the real component of a) the original and b) the contaminated test vector.

In Fig. 5.4 the average of the resulting spectral estimates for 30 Monte Carlo simulations under contamination is shown. The noise floor is now elevated, masking the weaker signals, and only two of the strongest channels are still visible. However, these levels are incorrect. We will attempt to make this technique more resistant to outliers using our previously developed robust correlation estimators.

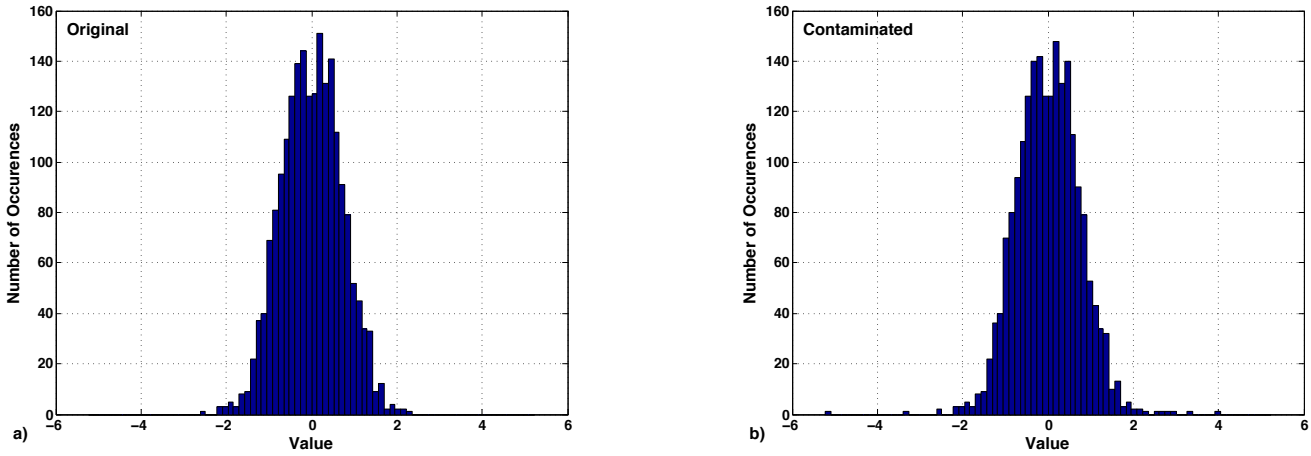

Figure 5.3: Histograms of the real component of a) the original and b) the contaminated test vector.

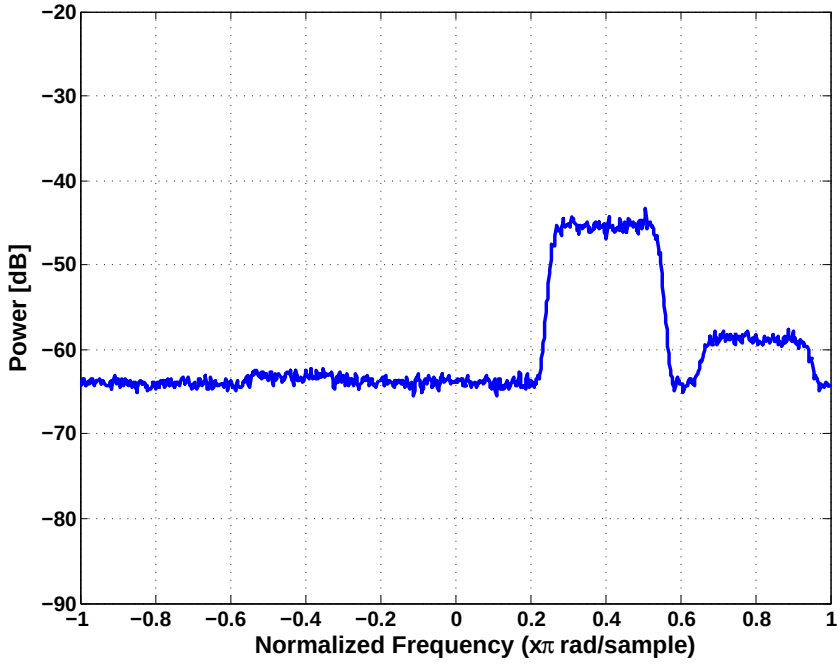

Figure 5.4: Parametric spectral estimation example using the classical Burg method applied to Gaussian data corrupted with replacement outliers have a contamination rate of $\varepsilon = 0.01$ and variance ten times the underlying process.

5.3 Robust Burg Algorithm Using the Robust Estimators of Correlation

A robust implementation of the Burg algorithm replaces the reflection coefficient estimator given by (5.7) with either the IPPC or IMRE. We justify this since the expected value of (5.7) is

$$E\left\{\hat{\kappa}_l\right\} = \frac{2E\left\{E^f_{t,l-1}E^{b*}_{t-1,l-1}\right\}}{\left(\sigma^2_{e^f_{l-1}} + \sigma^2_{e^b_{l-1}}\right)} = \frac{2E\left\{E^f_{t,l-1}E^{b*}_{t-1,l-1}\right\}}{\left(\sigma^2_{e_{l-1}} + \sigma^2_{e_{l-1}}\right)} = \rho^*_{E^f_{t,l-1},E^b_{t-1,l-1}}, \tag{5.10}$$

because $\sigma^2_{e^b_{l-1}} = \sigma^2_{e^f_{l-1}} = \sigma^2_{e_{l-1}}$.

The Monte Carlo procedure is repeated again without contamination to assess that the IPPC and IMRE are suitable replacements for the reflection coefficient estimator. Fig. 5.5 shows the average spectral estimate, indicating performance that is similar to the classic implementation under the Gaussian assumption.

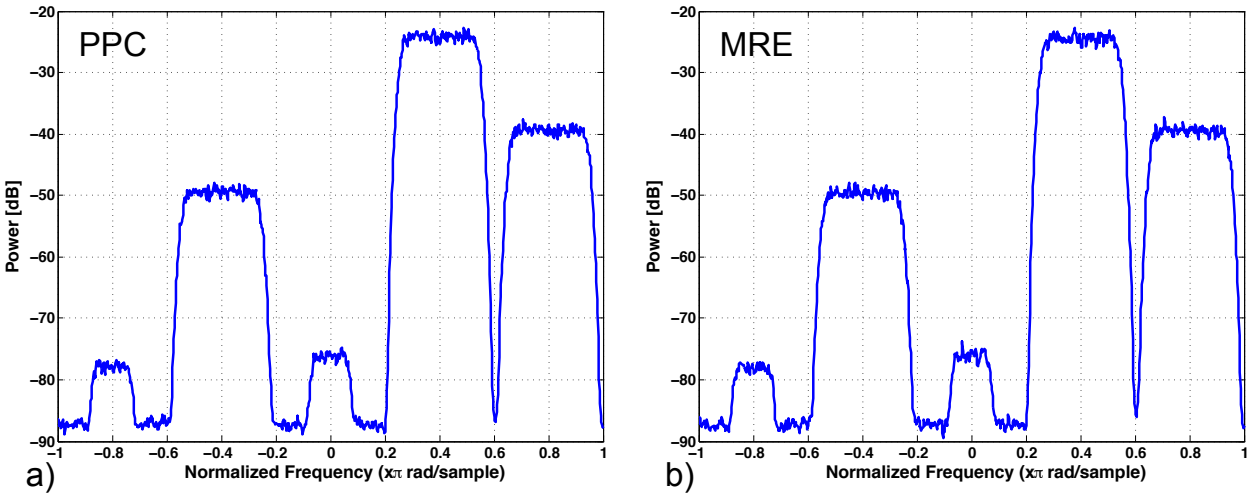

Figure 5.5: Parametric spectral estimation example using the robust a) PPC and b) MRE Burg method applied to purely Gaussian random process. Both robust versions show excellent agreement to the classic Burg method under Gaussianity.

5.4 Results of Robust Burg Algorithms Based on the IPPC and IMRE in the Presence of Additive Outliers

The robustness of the Burg method using the IPPC and IMRE is now examined for the additive outlier model. Again, using the same additive outlier model as previously described, we perform the same type of Monte Carlo analysis. It will be observed that performance will improve as the number of iterations increases. However, as was seen in the bias analysis of these estimators, the algorithms converge very quickly. Typically, using more than two or three iterations offers minimal increased benefit. This is an expected result as we saw an example of the quick convergence for the IPPC and IMRE in Fig. 4.13.

The results in Fig. 5.6 are obtained using the non-iterative PPC and MRE, or the 'zero-iteration' result. This results in spectral estimates revealing three of the five signals. This is an improvement over the classical implementation, but it does not allow detection of the weaker signals.

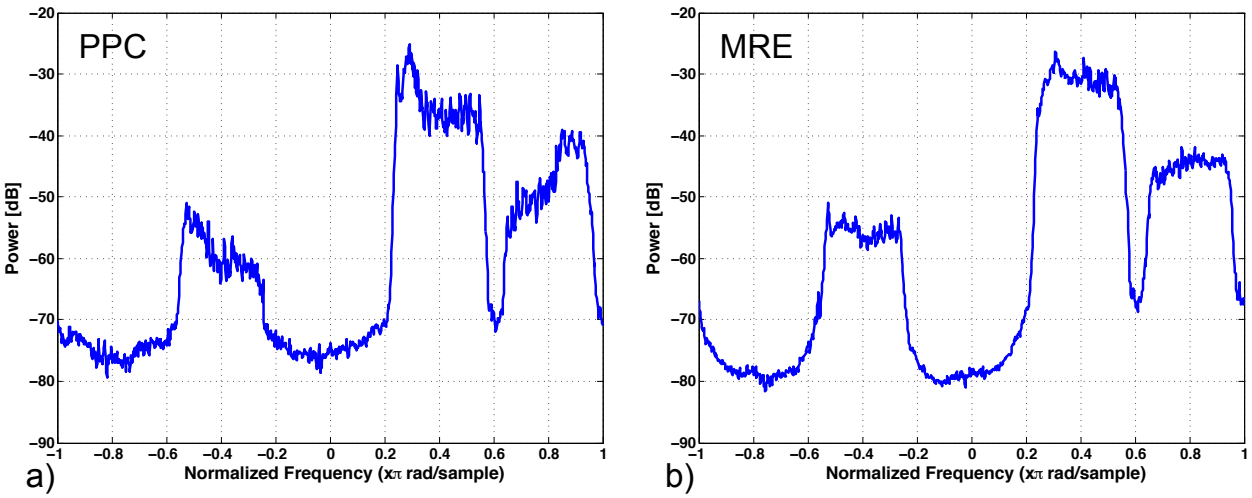

Figure 5.6: Parametric spectral estimation example using the robust a) PPC and b) MRE Burg method applied to $\varepsilon = 0.01$ contaminated Gaussian random process. The two weakest signals are not visible.

The results in Figs. 5.7, 5.8 show improvement as the number of iterations is increased to 1 and 2 respectively. In both cases, the weakest signals are now detectable.

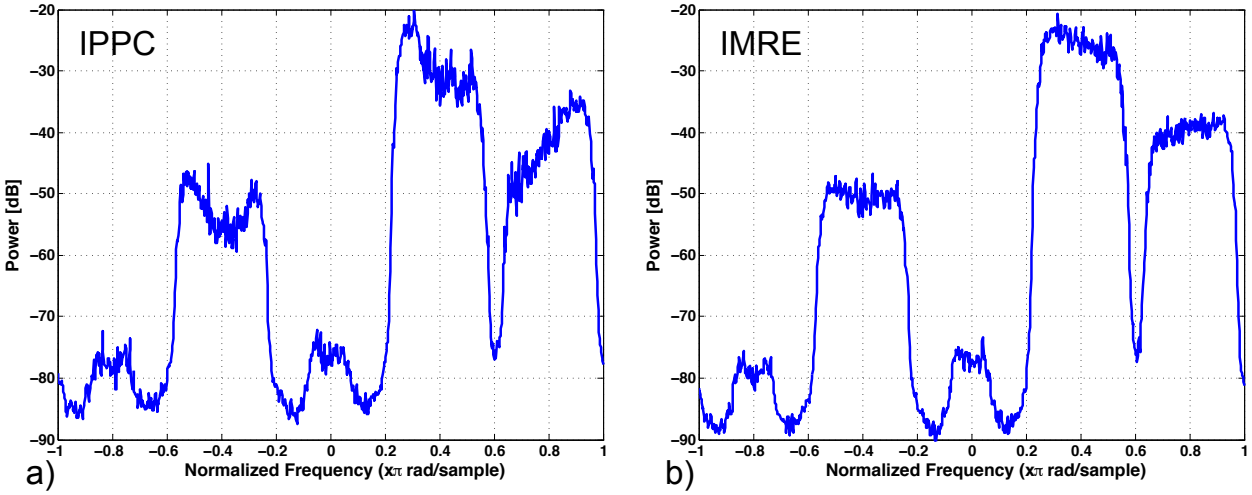

Figure 5.7: Parametric spectral estimation example using the robust a) IPPC and b) IMRE Burg method applied to contaminated data. The estimates improve after an additional iteration, revealing the weaker two signals not seen in Fig. 5.5.

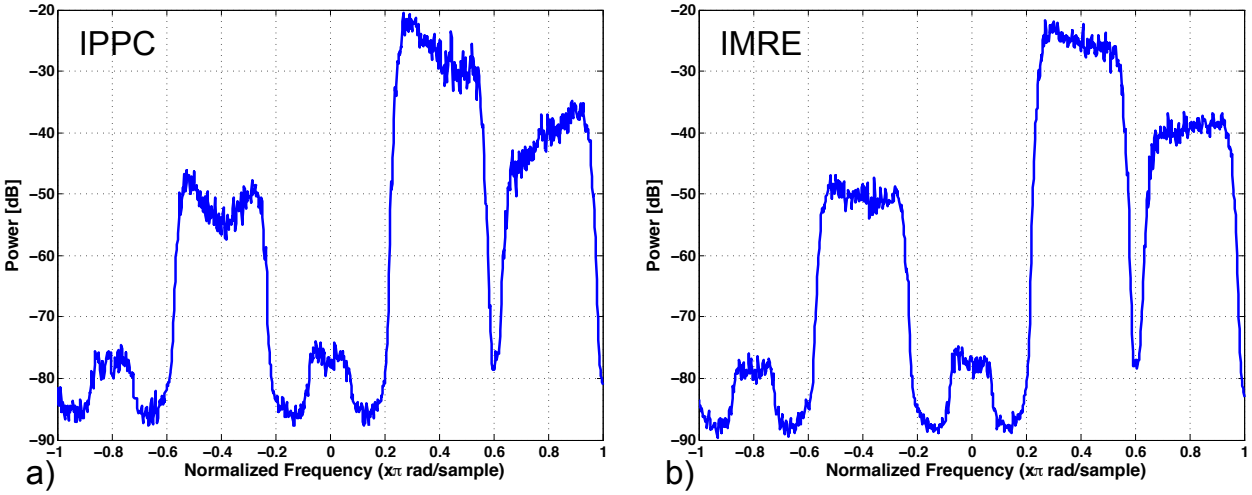

Figure 5.8: Parametric spectral estimation example using the robust a) IPPC and b) IMRE Burg method applied to contaminated data. The estimates improve again after two iterations, improving the estimate further when compared to Fig. 5.7.

Finally, results are shown for the cases where the IPPC and IMRE have converged. This is seen in Fig. 5.9 we we have chosen five and ten iterations. The differences between these two

simulations is minimal as the estimators converged. The resulting spectral estimates show very good agreement with the original case under Gaussianity. All five signals are observed, and the power levels are approximately equal to those seen in Fig. 5.1

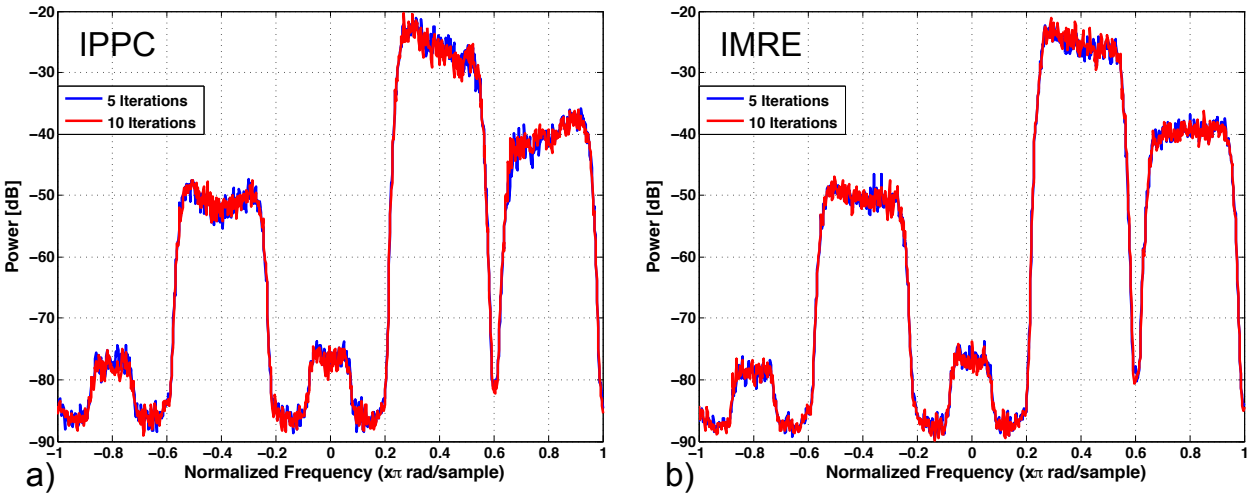

Figure 5.9: A comparison of the IPPC and IMRE based Burg method spectral estimators on contaminated data after five and ten iterations. Very little change is observed, as the iterative estimators have converged.

In conclusion, we observe that the IPPC and IMRE estimators can be used as effective replacements for the estimator of the reflection coefficient in the Burg method. Very good spectral estimates are obtained under additive outlier contamination even for this challenging signal model. Additionally, a significant number of iterations is not required, and the algorithms typically converge after one or two iterations. Results after five and even ten iterations of the robust IPPC and IMRE estimators show no improvement over results obtained using two iterations.

Chapter 6

Robust Non-Parametric Spectral Estimation

Non-parametric spectral estimators are often used in signal processing applications. Unlike the parametric techniques, non-parametric spectral estimators are model agnostic. For example, there is no requirement to fit the data to an ARMA model, and subsequently no need to select or estimate a model order. This allows this same estimator to be applied to different data sets without modification. While this may be attractive, there are of course potential drawbacks. For instance, estimating (2.33) given N observations of a time series presents several problems. Firstly, only a finite number of samples are available. This limitation forces the data to be windowed to a finite length. This windowing decreases the frequency resolution of the estimate. Secondly, since only one realization of the process is observed, the expectation operator embedded in the correlation coefficient must be removed from (2.33). This leads us to the finite-sample PSD estimator, also known as the periodogram spectral estimator defined as

$$\hat{P}_x(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j2\pi nf/f_s}\right|^2, 0 \leq k < N \tag{6.1}$$

where f_s is the sampling frequency. Increasing the number of observations does not decrease the variance of the estimator on a per frequency-bin basis [42].

6.1 Classic Discrete Fourier Transform and the Least Squares Equivalent

When (6.1) is evaluated at N equally spaced frequencies such as, kf_sn/N , $0 \leq k < N$ where f_s is the sampling frequency, the periodogram is often evaluated using the discrete Fourier transform (DFT) given by

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, 0 \leq k < N. \tag{6.2}$$

The DFT is often implemented using the efficient fast Fourier transform. The DFT in (6.2) can be thought of as the solution to a least-squares problem. For instance, suppose we wish to represent the observed data vector $\mathbf{x}$ using an orthonormal basis $\mathbf{D}_{\mathrm{F}}$. We construct an $N \times N$ unitary matrix $\mathbf{D}_{\mathrm{F}}$ which we call a dictionary. The entries of $\mathbf{D}_{\mathrm{F}_{m,n}}$ are given by the complex exponential given by $1/\sqrt{N} e^{j2\pi(m-1)(n-1)/N}$. Our linear system of equations is constructed by

$$\mathbf{D}_{\mathrm{F}} \mathbf{w}_{\mathrm{F}} = \mathbf{x} + \mathbf{e}, \tag{6.3}$$

where $\mathbf{e}$ is our observation noise. If we assume that the entries of $\mathbf{e}$ are independently and identically distributed complex-valued Gaussian random variables, then the least-squares solution is expressed as

$$\hat{\mathbf{w}}_{\mathrm{F}} = \left(\mathbf{D}_{\mathrm{F}}^H \mathbf{D}_{\mathrm{F}} \right)^{-1} \mathbf{D}_{\mathrm{F}}^H \mathbf{x}, \tag{6.4}$$

and since $\mathbf{D}_{\mathrm{F}}^{H}\mathbf{D}_{\mathrm{F}} = \mathbf{I}$, we get

$$\hat{\mathbf{w}}_{\mathrm{F}} = \mathbf{D}_{\mathrm{F}}^{H}\mathbf{x}. \tag{6.5}$$

Note that $\mathbf{D}_{\mathrm{F}}^{H}$ is the Fourier matrix.

6.2 The Classic Thomson Multi-Taper Method

The Thomson Multitaper spectral estimator is an adaptive estimator that is considered a valuable tool for cognitive radio spectrum sensing [35]. It is designed to improve the performance of the periodogram, producing lower-variance estimates. This is achieveed by performing a weighted average of multiple estimates performed on the same N observations, each obtained using a different orthogonal time-domain taper. In a simplified form, the m spectral estimates are produced from

$$\hat{X}_m(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n]\nu^{(m)}[n]e^{-j2\pi kn/N}, \tag{6.6}$$

and the final spectral estimate is the weighted average

$$\hat{S}(k) = \frac{\sum_{m=0}^{M-1} \lambda_m \left|\hat{X}_m(k)\right|^2}{\sum_{m=0}^{M-1} \lambda_m}, \tag{6.7}$$

using λ_m to denote the eigenvalue of the m^{th} taper.

We recognize that (6.6) is a weighted sum of tapered discrete Fourier transforms (DFT). We now use the $N \times N$ Fourier dictionary to indicate how the individual m uncorrelated spectral estimates can be expressed as a linear regression using

$$\mathbf{x} \odot \nu^{(m)} = \mathbf{D}_{\mathrm{F}}\mathbf{w}_m + \mathbf{e}, \tag{6.8}$$

where the symbol $\odot$ indicates the Hadamard product of the data vector and the m^{th} taper. We again use $\mathbf{e}$ as our observation noise vector. Assuming $\mathbf{e}$ is white Gaussian observation noise, the optimal solution for $\mathbf{w}_m$ is obtained by minimizing the sum of the squares of the residuals. This optimal least-squares solution is the Gaussian MLE [42] and can be expressed in matrix form as

$$\hat{\mathbf{w}}_m = \left(\mathbf{D}_{\mathrm{F}}^H \mathbf{D}_{\mathrm{F}}\right)^{-1} \mathbf{D}_{\mathrm{F}}^H \left(\mathbf{x} \odot \nu^{(m)}\right). \tag{6.9}$$

As defined, $\mathbf{D}_{\mathrm{F}}$ is unitary; therefore (6.9) is exactly equal to (6.6).

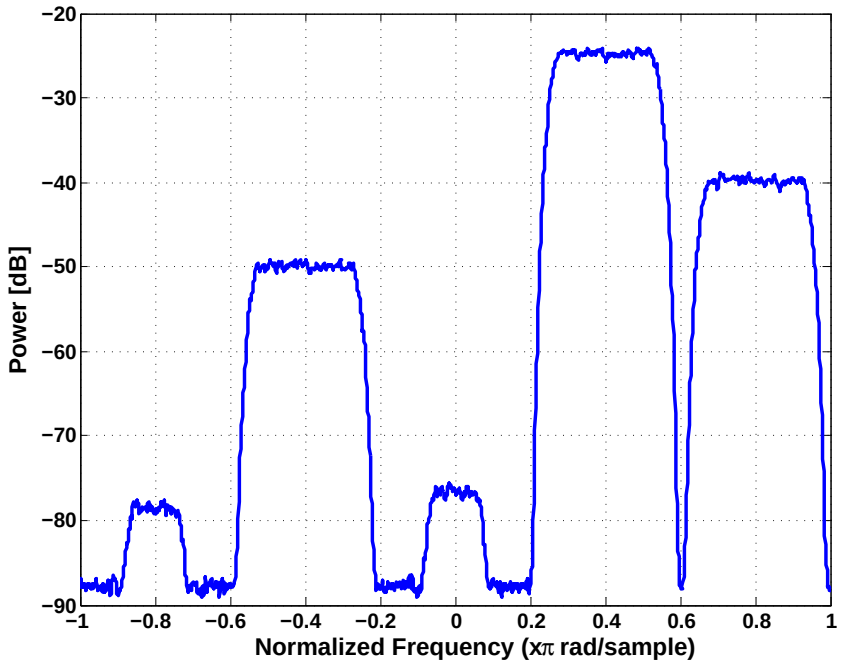

Figure 6.1: Multi-taper method spectral estimation example, with purely Gaussian data and no contamination. All signals are visible and the results are very similar to the results obtained from the Burg method observed in Fig. 5.1.

6.2.1 Impact of Outliers on the Discrete Fourier Transform and the Thomson Multi-Taper Method

We see that (6.3) is equivalent to the DFT-based spectral estimator if the solution is obtained by minimizing a quadratic cost function of the residuals. As such, it is optimal when the observation noise is Gaussian. We highlight that the explanatory matrix, $\mathbf{D}_{\mathrm{F}}$, in (6.8) is non-random, and that non-Gaussian observation noise in $\mathbf{e}$ manifests as non-Gaussian residuals. This has the affect of creating vertical outliers (see Chapter 2 for the definition of a vertical outlier), which can cause severe bias when using (6.9). Because $\mathbf{D}_{\mathrm{F}}$ is square, redundant observations are not available, precluding the use of robust estimators such as M-estimators.

To understand the impact caused by isolated thick-tailed observation noise, we consider the affect of large-magnitude impulsive observations. Since the DFT, and the MTM are both linear estimators, the spectral estimate of the underlying signal and observation noise can be computed separately and summed. For instance, suppose there is an impulsive noise sample that is represented as a scaled Kronecker delta function, $\alpha\delta[n-M]$, where α is an arbitrary complex-valued scalar and M is a delay. The resulting disrete-time Fourier transform is $X(\omega) = \alpha e^{-j\omega M}$. Thus, energy from one corrupted time sample will be distributed to all frequency bins of the resulting DFT. Energy from multiple impulsive noise samples will add non-coherently in the frequency domain. This causes the observed noise floor to be biased high, potentially masking weak peaks.

Fig. 6.2 reveals the bias that occurs when using the MTM on the same corrupted data analyzed in the previous section. At a contamination rate of only $\varepsilon = 0.01$, the MTM suffers from extensive bias. Only two of the five signals are observed, the other three are masked by the elevated noise floor. In light of this, we therefore opt to clean the data prior to using the least-squares-based Thomson Multitaper estimator in an effort to reduce the bias caused by impulsive noise.

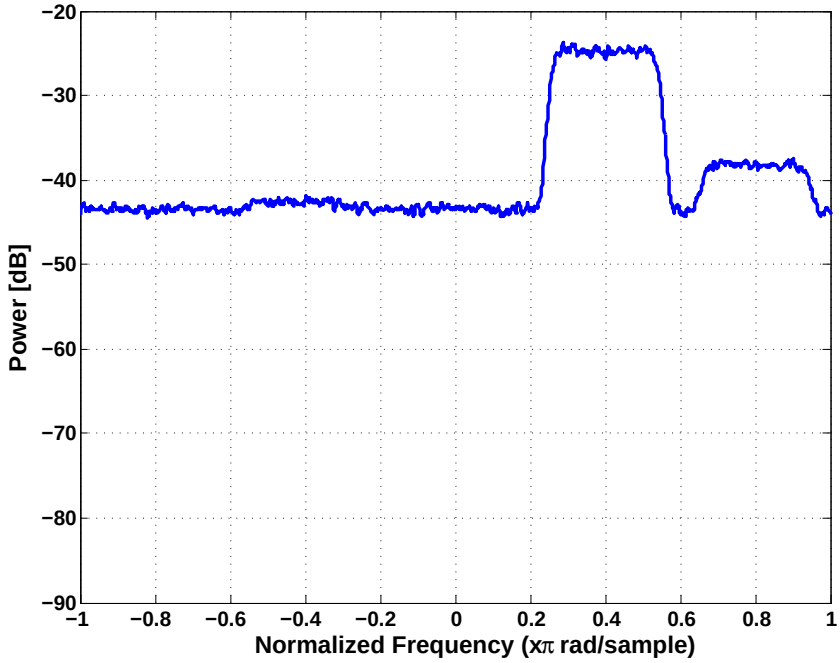

Figure 6.2: Multi-taper method spectral estimation example with $\varepsilon = 0.1$ contamination rate. The estimate is biased, and the artificially high noise floor masks all but the two strongest signals.

6.3 Robust Spectral Estimation Using Multiple-Basis Signal Representation to Remove Impulsive Noise

Data cleaning or filtering is required to remove the impulsive noise prior to spectral estimation. This is an issue that effects non-parametric spectral estimators and was not a concern for the parametric estimator discussed in the previous section. We will briefly discuss the drawbacks of two approaches next, the robust filter cleaner and principal components analysis. The robust filter cleaner has implementation issues that can lead to unstable results. Principal components analysis is not fundamentally robust to outliers. After briefly discussing these two algorithms, a robust implementation of impulsive noise suppression using weighted basis pursuit is developed.

6.3.1 The Robust Filter Cleaner

The robust filter cleaner [46] can be used to clean our data vector. The cleaner requires an estimate of a state transition matrix, populated by the robustly estimated autoregressive parameters that model the observed data. The algorithm uses prediction residuals to test whether data follows the model or is a corrupted observation that is flagged as an outlier. Outliers are replaced with robustly predicted estimates. This algorithm is similar in structure to the Kalman filter [40, 27]. However, the robust filter cleaner is known to be unstable in two situations [8]. The first situation is when the AR coefficients used to model the data are large in magnitude. The second one is when a large number of coefficients are needed. In both situations, the algorithm will perform poorly.

Another practical matter that impacts implementation is the placement of the outliers. The filter cleaner needs to be initialized using uncontaminated observations. If a length-P predictor is used on the observed data, then P continuous uncorrupted samples are required. Assuming that the first P samples used by the filter cleaner are uncorrupted when in fact they are not, can lead to unreliable results. If both forward and backward predictions are used, then the window of uncorrupted samples will increase as illustrated in Fig. 6.3

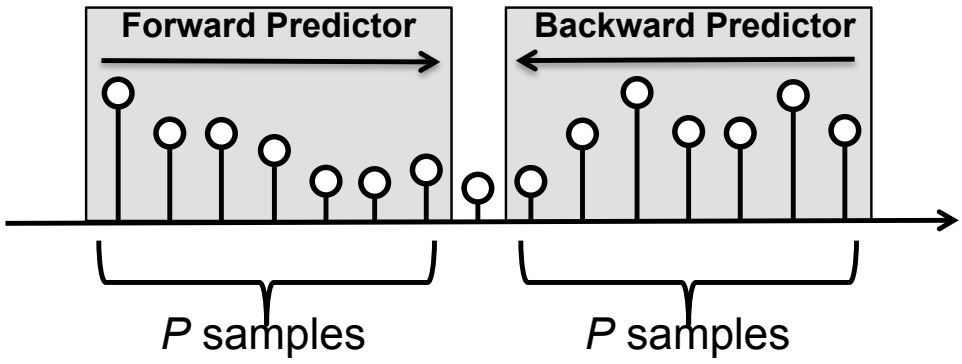

Figure 6.3: The filter cleaner requires P forward and/or P backward outlier-free samples to create an initial prediction to clean data. Without a contiguous block of outlier-free data, the filter cleaner cannot be initialized.

6.3.2 Principal Component Analysis

Principal components analysis (PCA) is a well-known technique to reduce the dimensionality of data [50]. The first principal component is the direction, $\mathbf{a}_1$, that results in the maximum projected variance subject to $||\mathbf{a}_1|| = 1$. The second principal component, $\mathbf{a}_2$, is the direction that maximizes the projected variance of the data subject to $\mathbf{a}_2^H \mathbf{a}_1 = 0$, and so on.

The direction vectors are obtained from the eigen decomposition of the signal's covariance matrix $\mathbf{\Sigma}$. The eigenvectors are the directions of projection, $\mathbf{a}_1, \mathbf{a}_2, \ldots$, and the associated eigenvalues are the resulting variances. Note that the resulting eigenvectors are independent of any scaling of the covariance matrix. Recognizing that the correlation matrix is just a scaled version of the covariance matrix, we would expect that this analysis would not be robust to outliers, such as impulsive noise, if we use a non-robust estimate of the covariance matrix.

Principal component analysis can also be performed using the singular value decomposition (SVD) of any non-square data matrix. The SVD produces the matrices $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}$ such that $\mathbf{X} = \mathbf{USV}^H$, where $\mathbf{S}$ is a diagonal matrix the same size of $\mathbf{X}$ with entries that are the singular values. An approximate reconstruction of the data matrix using the first q principal components requires one to use the q largest singular values and the associated q columns from the matrices $\mathbf{U}$ and $\mathbf{V}$.

A time series analysis example can be constructed by creating a data matrix $\mathbf{X}$ containing N rows where the n^{th} row is $\{x_{t+n-1} \ldots x_{t+n+L-1}\}$, which are time shifted realizations of the process X_t. For example consider a data matrix with 100 rows and 900 columns containing a Gaussian random process from an AR(1) process with feedback coefficient equal to 0.9. PCA is performed on this data matrix for an uncontaminated case, and a case with a single outlier located at sample number 400 having a value of 40. Data cleaning using PCA assumes that the impulsive noise does not corrupt the principal components.

The eigenvalues are plotted in Fig. 6.4 for both the uncontaminated and contaminated

cases. We observe that the single outlier inflates the singular values, especially the lower ones. Next, the data is approximated using the first 20 principal components. Notice in Fig. 6.5 how the representation closely matches the original time series when no outliers are present. However, the representation using the principal components derived from the contaminated data reveals the outlier in sample 400. This is due to the impulsive noise distorting the estimates of the principal components of the underlying uncorrupted data. Classic PCA is not robust to impulsive noise and robust alternatives have been proposed in the literature [13, 18, 46].

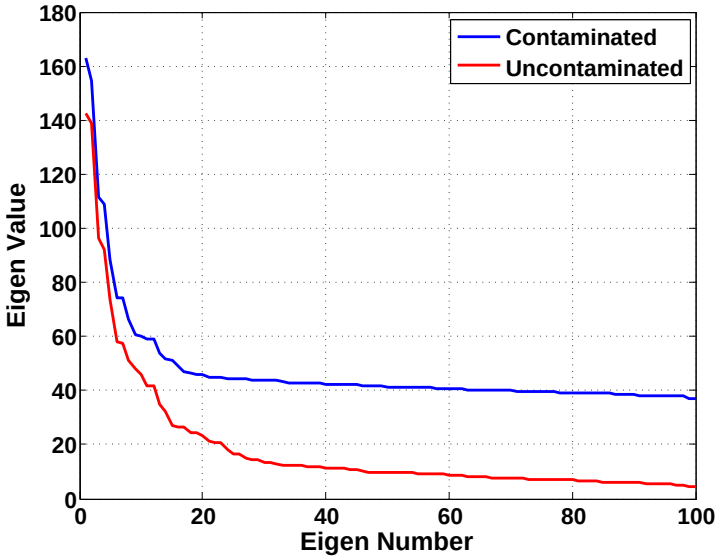

Figure 6.4: The eigen spectra for a time series Principal Components Analysis example reveals inflated lower eigen values when the time series is corrupted with a single outlier.

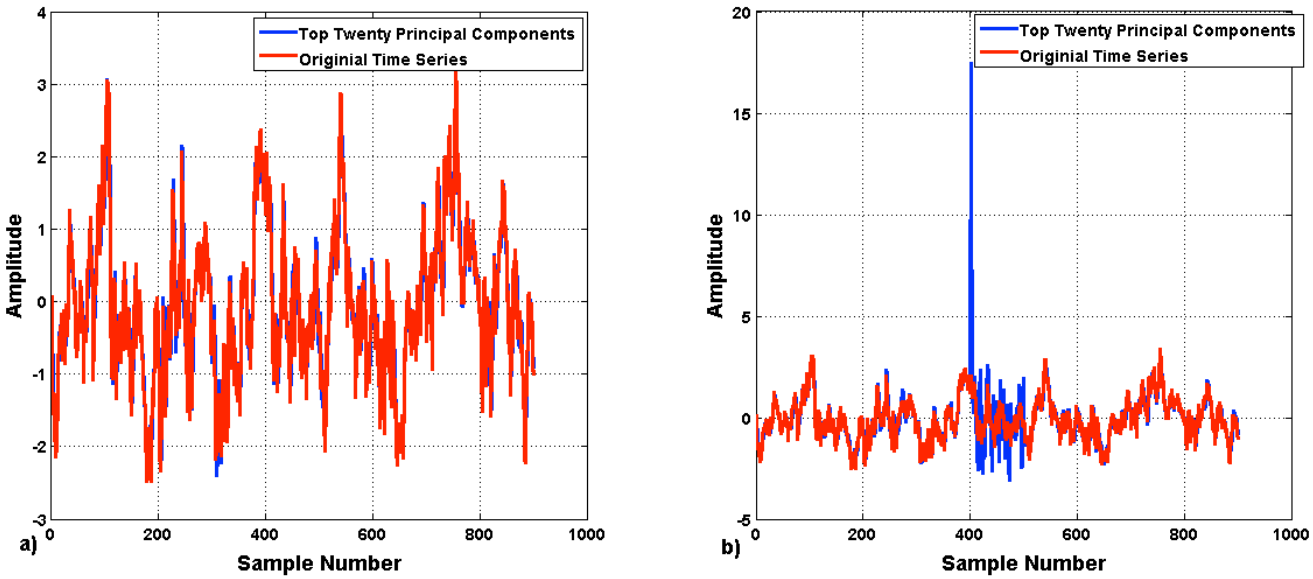

Figure 6.5: The approximation of a time series using PCA reveals that the method is not robust to impulsive noise. Under Gaussianity, a) using the first 20 principals produces a good approximation. However, given a single outlier b), we see that PCA does not produce an accurate representation of the original time series.

6.3.3 Impulsive Noise Removal Using Weighted Multiple-Basis Signal Representation

Multiple-basis signal representation [61] is a powerful signal processing technique that has recently received much attention. In general, two or more $N \times N$ orthonormal bases, or dictionaries, are concatenated to produce an overcomplete dictionary $\mathbf{D}_0 = \mathbf{D}_1 \cup \mathbf{D}_2 \cup \ldots \cup \mathbf{D}_J$, such that a linear combination of its columns, also called atoms, produces the signal $\mathbf{x}$, that is $\mathbf{x} = \mathbf{D}_0 \boldsymbol{\omega}_0$. While it may be possible to represent the signal completely using a single dictionary, the underlying principle is that fewer atoms might be required if the concatenated dictionaries are chosen correctly.

This signal representation can be leveraged to remove impulsive noise from observed data by choosing appropriate dictionaries. The Frequency and Time dictionaries [16] are typical choices when trying to describe data that contains a mixture of both steady-state sinusoidal

components and isolated impulses. We define this composite length-N signal as a linear combination of the columns of these two dictionaries using the notation given by

$$\mathbf{x} = [\mathbf{D}_{\mathrm{F}} \quad \mathbf{D}_{\mathrm{S}}] \begin{bmatrix} \mathbf{w}_{\mathrm{F}} \\ \mathbf{w}_{\mathrm{S}} \end{bmatrix}, \tag{6.10}$$

where $\mathbf{D}_{\mathrm{F}}$ is defined as before, and we introduce the Spike dictionary which is the $N \times N$ identity matrix. Impulsive noise is removed from the observed data by subtracting the contribution from the Spike dictionary as follows

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{D}_{\mathrm{S}}\mathbf{w}_{\mathrm{S}}. \tag{6.11}$$

In practice, $\mathbf{w}_{\mathrm{S}}$ is not known and must be jointly estimated with $\mathbf{w}_{\mathrm{F}}$ from the observed data. The challenge is to obtain the solution to (6.10), which is an overdetermined system of equations.

Borrowing from the field of compressed sensing and sparse signal representation, (6.10) can be solved assuming a sparse condition, a small number of non-zero coefficients, on the vector $\boldsymbol{\omega} = \begin{bmatrix} \mathbf{w}_{\mathrm{F}}{}^{T} & \mathbf{w}_{\mathrm{S}}{}^{T} \end{bmatrix}^{T}$. The sparse solution is obtained as a constrained optimization problem defined as

$$\min \|\boldsymbol{\omega}\|_0 \quad s.t. \, [\mathbf{D}_{\mathrm{F}} \quad \mathbf{D}_{\mathrm{S}}] \, \boldsymbol{\omega} = \mathbf{x}, \tag{6.12}$$

where the l_0-norm defined by Donoho is used. It has been shown that solving (6.12) is an NP hard problem, and is computationally impractical. An iterative nonlinear alternative known as basis pursuit [17] is a similar formulation to (6.12), however it is based instead on the l_1-norm minimization given by

$$\min \|\boldsymbol{\omega}\|_1 \quad s.t. \, [\mathbf{D}_{\mathrm{F}} \quad \mathbf{D}_{\mathrm{S}}] \, \boldsymbol{\omega} = \mathbf{x}. \tag{6.13}$$

The advantage of basis pursuit is that it can be formulated as a convex optimization problem,

solvable by linear programming techniques. However, basis pursuit is only an approximation to (6.12), producing exact solutions depending on the chosen dictionaries, and sparseness of the signal. When it does not produce the exact l_0 solution, it often produces a very good approximation as shown in [21]. Donoho and Huo [21] provide a bound on the maximum number of atoms allowed to define $\mathbf{x}$ such that basis pursuit produces the exact l_0 solution, and is repeated here. Formally we have

$$\|\boldsymbol{\omega}\|_0 < 0.5(1 + M^{-1}). \tag{6.14}$$

This bound includes what is called the Mutual Incoherence, M, of the dictionaries and is defined by

$$M(\mathbf{D}_1, \mathbf{D}_2) = \sup\left\{|\langle \mathbf{d}_1, \mathbf{d}_2\rangle| : \mathbf{d}_1 \in \mathbf{D}_1, \mathbf{d}_2 \in \mathbf{D}_2\right\}. \tag{6.15}$$

Because the chosen dictionaries must contain orthonormal columns, the resulting M must satisfy $1/\sqrt{N} \leq M \leq 1$. Using (6.15), it is seen that the Mutual Incoherence for the multi-bases dictionary in (6.10) is $M = 1/\sqrt{N}$. In practice, many observed signals will contain more atoms than allowed by (6.14), and basis pursuit will only provide an approximation. In these cases, it has been observed that using *a priori* information can improve the sparseness of the resulting approximation [70].

In [31] and [23], the concept of weighted basis pursuit is introduced. Statistical weighting is incorporated into basis pursuit by defining a new vector, $\boldsymbol{\omega}_*$, such that each individual element of $\boldsymbol{\omega}$ is separately scaled by some probability weighting, $\omega_{*_i} = \omega_i/p_i$, and our new l_1 minimization is performed, yielding

$$\min\|\boldsymbol{\omega}_*\|_1 \quad s.t. \begin{bmatrix}\mathbf{D}_{\mathrm{F}} & \mathbf{D}_{\mathrm{S}}\end{bmatrix}\boldsymbol{\omega} = \mathbf{x}, \tag{6.16}$$

or equivalently

$$\min\|\boldsymbol{\omega}\|_1 \quad s.t. \left[\mathbf{D}_{\mathrm{F}}\mathbf{P}_{\mathrm{F}} \quad \mathbf{D}_{\mathrm{S}}\mathbf{P}_{\mathrm{S}}\right]\boldsymbol{\omega} = \mathbf{x}, \tag{6.17}$$

where $\mathbf{P}_{\mathrm{F}}$ and $\mathbf{P}_{\mathrm{S}}$ are diagonal matrices with $\mathbf{P}_{ii} = p_i$.

In our example, $\mathbf{P}_{\mathrm{F}} = \mathbf{I}$ and the elements of the weighting matrix $\mathbf{P}_{\mathrm{S}}$ are indirectly produced from the final weights obtained from (4.39). The *a priori* weighting applied to the Spike dictionary is used to assign a higher probability that certain atoms correspond to impulsive noise samples in the time series. Therefore, weights from (4.39) that are small must be assigned a weighting near one, and large weights must be assigned a weighting that is close zero. This is accomplished by subtracting the weights from one. This results in our pairwise weighting (pw) defined by

$$pw_i = 1 - w_i. \tag{6.18}$$

These pairwise weights must be related to single observations, and subsequently single atoms of the Spike dictionary. An effective method to accomplish this is through

$$p_i = \sqrt{pw_i pw_{i-1}}, \quad 1 < i < N, \tag{6.19}$$

where p_1 and p_N are pw_1 and pw_{N-1}, respectively. Now that we have pairwise weightings, there is one final step that is performed before the elements of $\mathbf{P}_{\mathrm{S}}$ are determined. There is a small probability that the weightings from (4.39) have incorrectly identified a pairwise observation as being contaminated or uncontaminated. Thererefore, to account for this, we should not completely eliminate atoms from the Spike dictionary by assigning a value of zero

to any p_i. Instead, we assign two weightings based on a threshold as follows

$$\mathbf{P}_{ii} = \begin{cases} 0.5, & \text{if } p_i \leq \Omega \\ 1.0, & \text{if } p_i > \Omega \end{cases}, \tag{6.20}$$

where Ω is a tuning parameter. Due to the nonlinear nature of basis pursuit, Ω was chosen empirically through the use of Monte Carlo simulations. A value of 0.5 tended to produce very good results in our simulations.

6.3.4 The Robust Thomson Multi-Taper Method and Simulation Results

In this section we discuss the Thomson multitaper spectral estimation results after impulsive noise is suppressed using (6.11) and (6.17). The weights are computed using the SPGL1 solver [75], an efficient Matlab-based basis pursuit toolbox that can be used to solve several types of L_1 minimization problems. The impulsive-free data is the same as used before, consisting of five narrowband Gaussian processes at five distinct frequencies plus white Gaussian observation noise. Separate results are shown where the weights applied to the spike dictionary are derived from both the IPPC and IMRE estimators as described.

Our first results show the resulting spectral estimate using this technique under the assumed Gaussian distribution. We performed this spectral estimation procedure thirty times, and plotted the median of the resulting spectral estimates. This case allows us to examine any negative impact that this technique imparts on the estimate when no outliers are present. We note two things from Fig. 6.6. First, the resulting performance of the multiple-basis impulsive noise subtraction method matches that of the original MTM under the Gaussian assumption shown in Fig. 6.1. Second, the number of iterations used by both the IPPC and IMRE has little impact on the result. This indicates that the iterative procedures have converged quickly, essentially after one iteration, and have not weighted the atoms of the

the spike dictionary. In this "do no harm test", we see that the algorithms preserved the MTM's uncontaminated performance seen in Fig. 6.1.

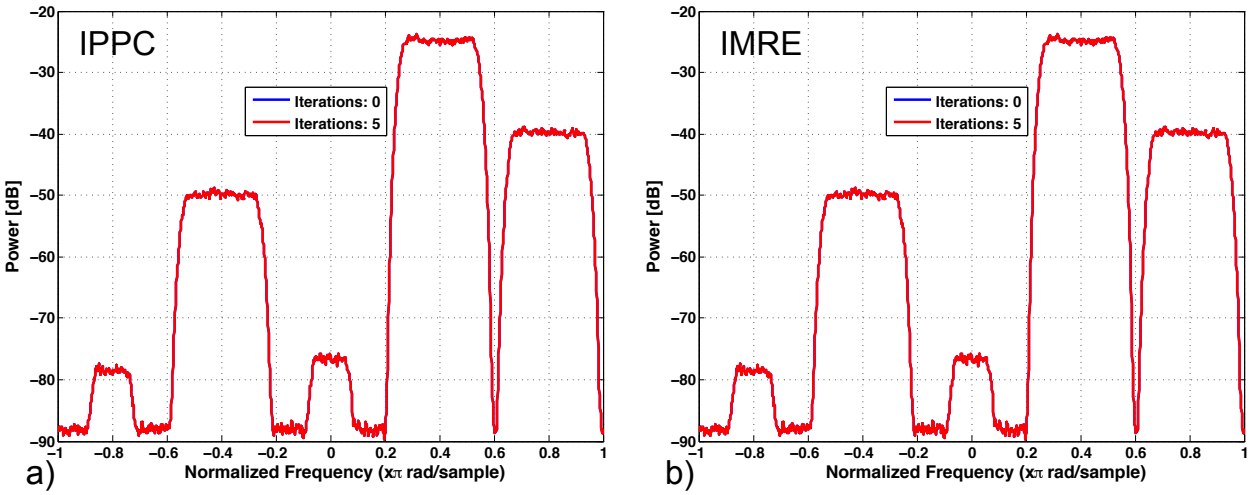

Figure 6.6: MTM spectral estimates made on uncontaminated data after impulsive noise suppression. This example indicates that the technique does not create harmful artifacts in the case of no contamination. These results show that allowing the a) IPPC and b) IMRE to iterate does not affect the estimates, indicating the algorithms have converged.

Additive impulsive noise is now introduced in the same manner as before; it is a white complex-valued Gaussian random process with a variance equal to ten times the variance of the uncontaminated process. The contamination rate is initially set to $\varepsilon = 0.01$, allowing us to compare results from the previous parametric spectral estimation section. We shall see that these new techniques perform very well. We will show additional results for higher contamination rates, $\varepsilon = 0.05$, demonstrating the benefit of the weighted basis pursuit method.

Results are shown as we increase the number of iterations used by the IPPC and IMRE. The first results in the presence of contamination displayed in Fig. 6.7 show that the weighted basis pursuit methods perform very well. These results are obtained after the zeroth iteration of the IPPC and IMRE. In Fig. 6.8 we show the resulting spectral estimates when the iterative estimators use one additional iteration. Using more than one iteration does not noticeabley improve the performance for these contamination parameters producing results

that are nearly identical to the uncontaminated case. Upon inspection, we identify three strong signals at normalized frequencies of -0.2, 0.2 and 0.4. Two weaker signals are present at normalized frequencies -0.4 and 0.0.

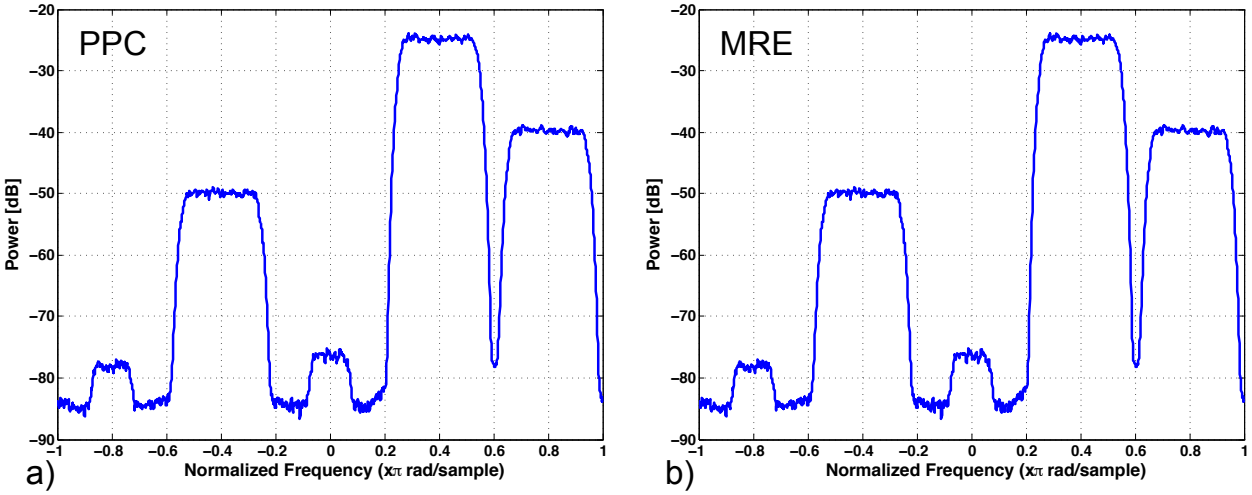

Figure 6.7: The MTM spectral estimator after the zeroth iteration using a) the PPC and b) MRE shows very good results when the data is corrupted at a contamination rate of $\varepsilon = 0.01$.

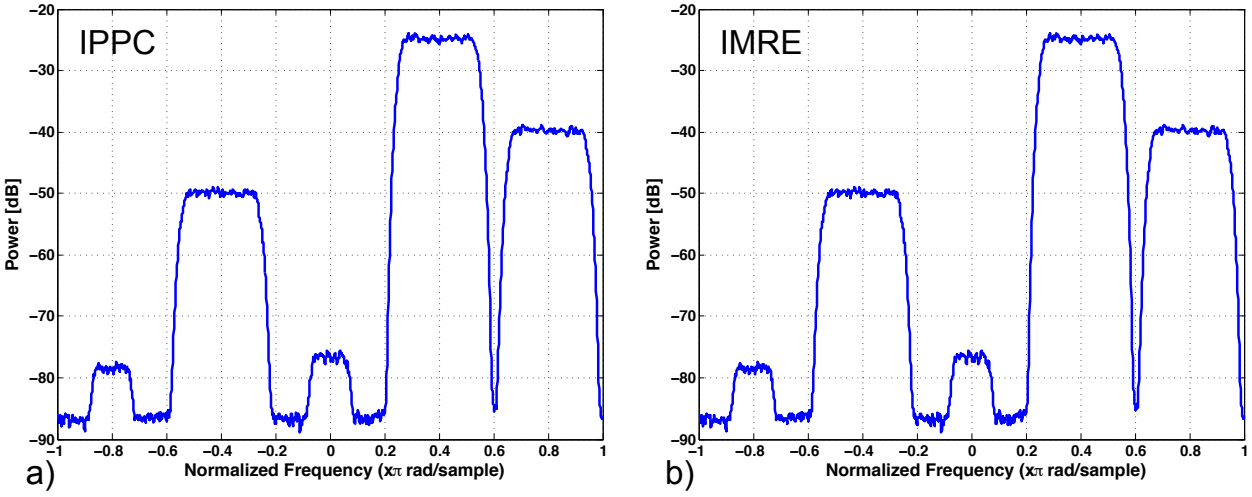

Figure 6.8: The MTM spectral estimator after the first iteration using a) the IPPC and b) IMRE shows very good results when the data is corrupted at a contamination rate of $\varepsilon = 0.01$. These results restore the performance to the case when there is no contamination.

In Fig. 6.9 the spectral estimates using basis pursuit and weighted basis pursuit are com-

pared. The basis pursuit method is effective at preserving all five signals. However, the two weakest signals are not recovered as well as compared to the weighted basis pursuit method that used two iterations of the IPPC and IMRE.

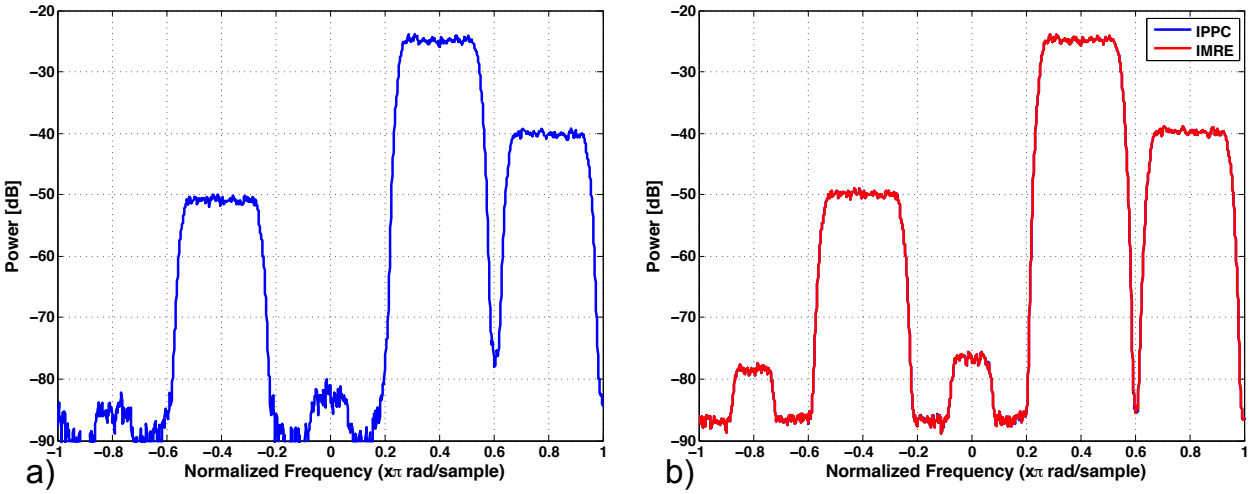

Figure 6.9: A comparison of the a) basis pursuit impulsive noise suppression and b) weighed basis pursuit impulsive noise suppression results. For a contamination level of $\varepsilon = 0.01$, both techniques are effective and all five signals are revealed. The weighted basis pursuit method performs better than the non-weighted version. Two iterations of the IPPC and IMRE were used.

The simulations are now repeated with a higher contamination rate of $\varepsilon = 0.05$. In Fig. 6.10 we see the weighted basis pursuit MTM estimates after the zeroth iteration of the IPPC and IMRE. While the results are better than the original non-robust MTM, the two weaker signals are not visible. This is caused by an elevated noise floor due to improper impulsive noise suppression. Next, we let the IPPC and IMRE to iterate one more time. In Fig. 6.11 the two weaker signals are now just visible over the noise floor. Allowing yet another iteration, the two weakest signals become a more visible and detectable over the noise floor in Fig. 6.12. While we do not recover the exact performance of the uncontaminated case, the two weaker signals are detectable in an extremely stressing example.

We now compare the resulting spectral estimates after standard basis pursuit and weighted basis pursuit impulsive noise suppression. In Fig. 6.13, the weaker signals are not distin-

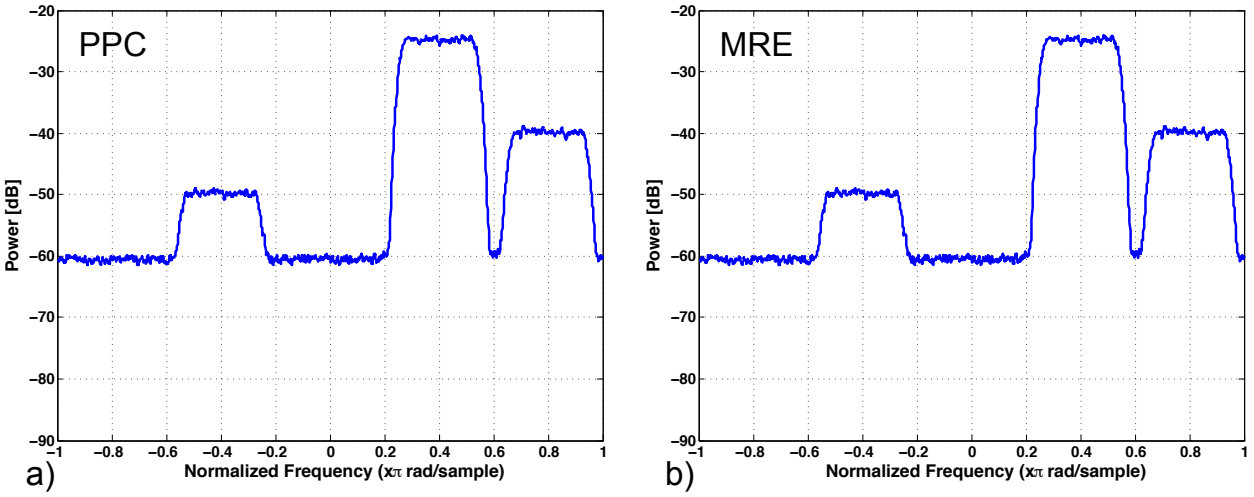

Figure 6.10: Robust MTM spectral estimates after the zeroth iteration using RMDs obtained from the a) IPPC and b) IMRE algorithms. We observe that the two weaker signals are obscured by a raised noise floor with a contamination rate of $\varepsilon = 0.05$, unlike the results seen in Fig. 6.7 that had a contamination rate of $\varepsilon = 0.01$.

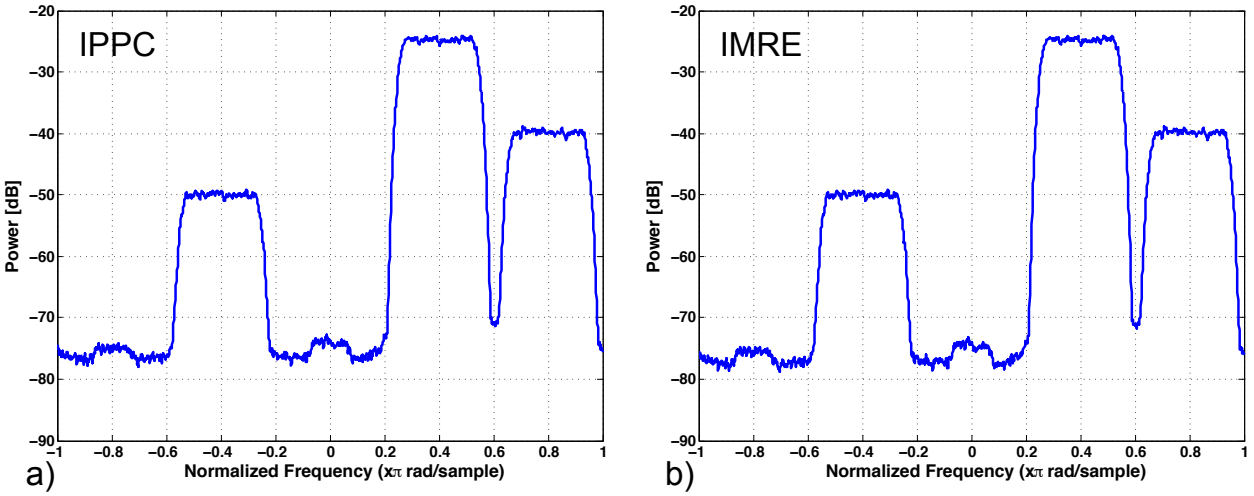

Figure 6.11: Robust MTM spectral estimates after the first iteration using RMDs obtained from the a) IPPC and b) IMRE algorithms. We observe that the two weaker signals are now revealed.

guishable from the noise floor after using the basis pursuit impulsive noise removal method. Additionally, the overall power level of the third loudest signal is slightly less than expected. The weighted basis pursuit method using two iterations of the IPPC and IMRE preserves the spectrum enough to allow us to see the weakest two signals.

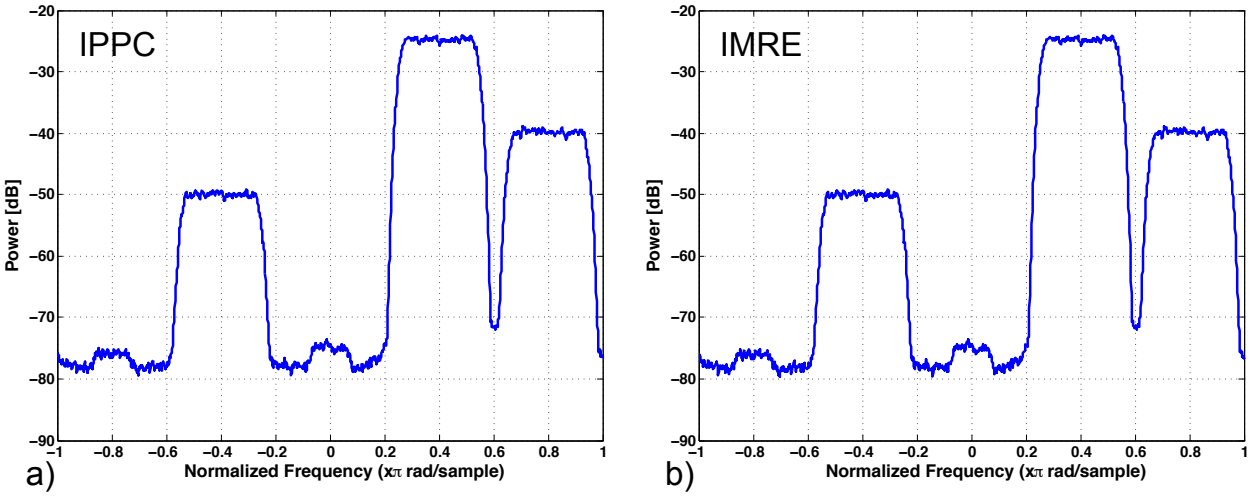

Figure 6.12: Robust MTM spectral estimates after the second iteration using RMDs obtained from the a) IPPC and b) IMRE algorithms. We observe slightly improved performance when only using one iteration of the IPPC and IMRE algorithms when comparing the results to those seen in 6.11.

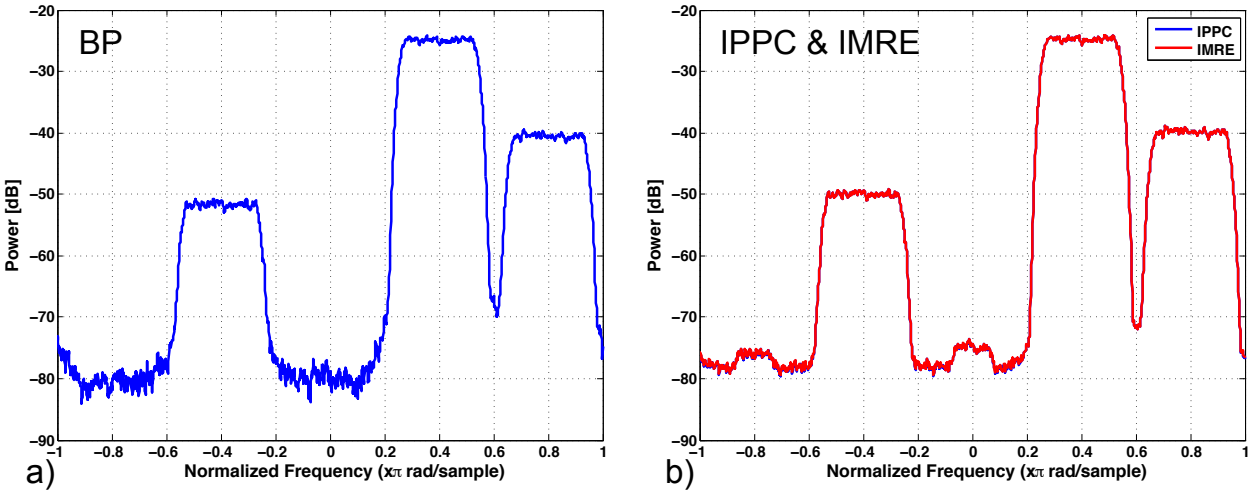

Figure 6.13: Robust MTM estimates after impulsive noise suppression using a) basis pursuit and b) weighted basis pursuit methods. The estimates obtained from basis pursuit do not reveal the very weak signals, but they are seen after using the weighted basis pursuit suppression method and two iterations of the IPPC and IMRE.

To help explain why the weighted basis pursuit algorithm provides better results in the presence of impulsive noise, we examine the residual impulse estimation error obtained using the two approaches. In Fig. 6.14 we plot the magnitude of the impulsive estimation error for

one of the Monte Carlo runs. The residual error for the weighted basis pursuit is observed to be smaller than for basis pursuit. In fact, for this example, the variance of the residual error is approximately five times smaller. Thus, when subtracting the impulsive noise using (6.11), the standard basis pursuit method effectively introduces five times more noise power than the weighted basis pursuit method. The difference in the estimation noise explains why the very weak signals are obscured when not using *a priori* weighting.

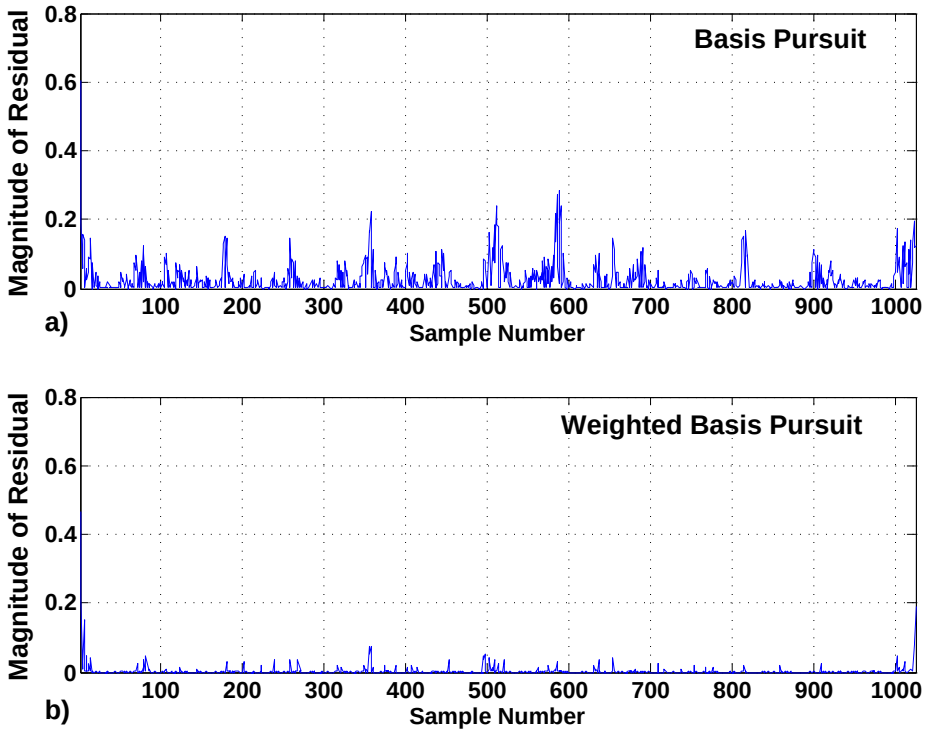

Figure 6.14: The impulsive noise estimation error of a) basis pursuit is approximately five times larger than the b) weighted basis pursuit based algorithm. This noise is inserted into the cleaned data, covering the low-power spectral peaks.

Chapter 7

Spectral Estimation Techniques Applied to Real Data

Standards for wireless regional area networks (WRAN) and wireless local area networks (WLAN) allow devices to transmit in digital television (DTV) white spaces. These standards, IEEE 802.22 and IEEE 802.11af, have been designed to take advantage of cognitive radio spectrum sensing technology. This is used to determine if there is unoccupied spectrum that the end user can use. While this standard potentially makes more spectrum available, the cognitive radios must honor the primary licensed users' rights to that spectrum. Therefore, any user that attempts to use DTV spectrum under these standards must operate in a "not to interfere" manner.

Motivated by this application, we intend to apply our robust estimators to actual recorded DTV data to determine their applicability to data that is not necessarily Gaussian. The Advanced Television Systems Committee standard [24] specifies an eight-level vestigial side-band modulation for the physical level. Each signal has a bandwidth of approximately six megahertz. Our robust parametric and non-parametric spectral estimates are evaluated on these signals. To control the analysis, additive impulsive noise will be synthetically generated in a manner similar to our earlier analysis.

7.1 Description of Recorded Data

The data was recorded in the television band using an Ettus Research^{TM} USRP N210 configured with a WBX receiver. The collected data spans three television channels. In the first channel, a large-power signal is received with strong multipath as evidenced by frequency selective fading. The center channel is known to have no local transmitter and is just noise. Finally, the third channel contains a weak signal from a distant transmitter. While the television signals are not Gaussian processes, their combination, along with multipath copies, tends to Gaussian as a result of the central limit theorem. The choice of this data set was intentional in that the third channel shows a signal with an extremely low power level with respect to the first signal.

One of the underlying assumptions is that the collected data should approximately follow a complex-valued Gaussian distribution. To verify this assumption, representative QQ-plots of the real and imaginary components of the collected data are shown in Fig. (7.1). The nearly linear QQ-plots verify our assumption of approximately Gaussian data.

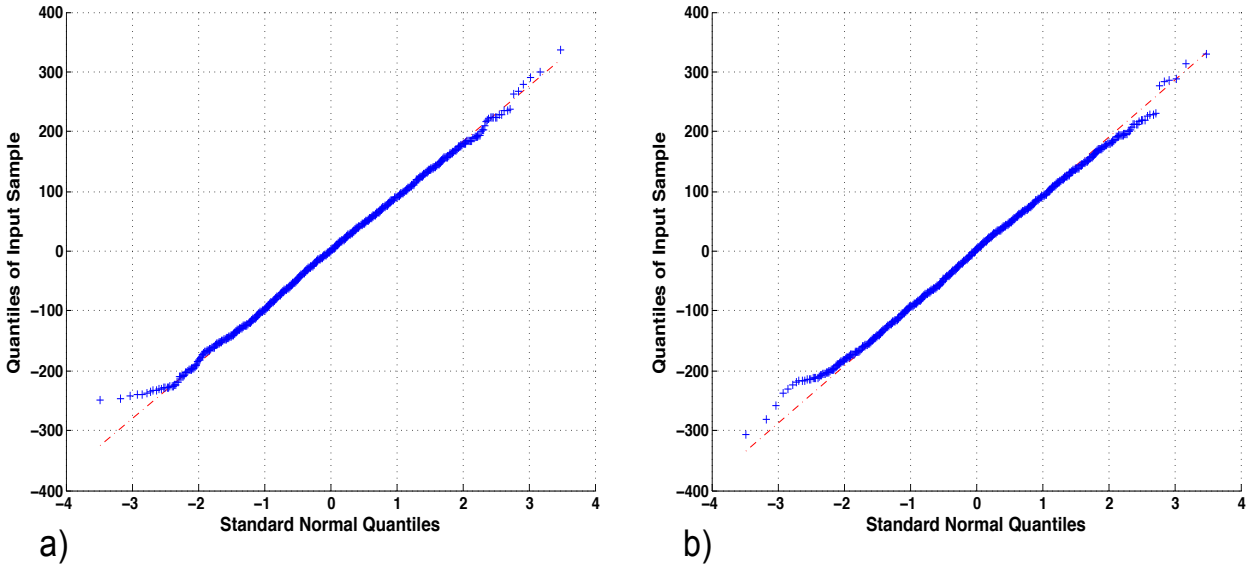

Figure 7.1: The nearly linear QQ-plots of the a) real and b) imaginary components of collected DTV data indicate an approximate Gaussian distribution.

7.2 Results of the Robust Burg Algorithm

We present results of the robust Burg algorithm using both the IPPC and IMRE as a robust replacement to the classic reflection coefficient estimator. First, we show spectral estimates using both the classic Burg method and the robust versions on uncorrupted data. This serves as a baseline of performance. Thirty spectral estimates were performed on data records of length 2048 samples. The data records were taken consecutively from the data acquisition system in non-overlapping segments. In Figs. 7.2 and 7.3, the median of the thirty spectral estimates is plotted for the classic and the robust versions. We observe that the estimates based on the robust estimators performs well, closely aligning with the classic method. There also appears to be a limitation on how well the IPPC- and IMRE-based estimates perform when tones are present in the data set. Observe at approximately 602.3 MHz, the observed power of the pilot tone is approximately 6 dB lower in the robust IPPC estimate and approximately 8 dB less in the IMRE estimate when compared to the classic Burg method. This discrepancy is negligible for the second observed channel at 614.3 MHz. However, this appears to be caused by the observed data and not the algorithms, as the pilot power should exceed the flat part of the data spectrum by approximately 10 dB. We note that there is a spur located in the center of the spectrum. This is a hardware artifact as there is no signal present at that frequency in the external environment.

We now test the estimators in the presence of impulsive noise. We introduce additive outliers in a similar manner to our previous analysis, where the variance of the outlier process is 10 dB greater than the observed data. Results for two different contamination rates, 0.01 and 0.05, are examined.

With an additive outlier contamination rate of 0.01, we observe in Fig. 7.4 that the spectral estimate obtained by the classic Burg algorithm is severely biased. It is difficult to observe the weaker signal over the increased noise floor. Only the upper edge of the weaker signal is observable over the noise floor near 618.5 MHz. Additionally, upon examination of the stronger signal centered on 605 MHz, we notice that the power level of the peaks are sub-

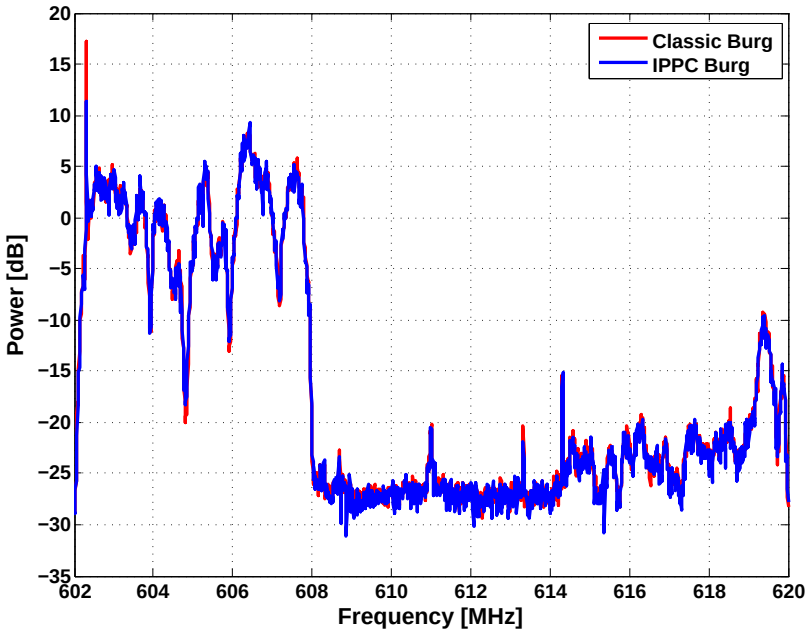

Figure 7.2: The classic Burg and robust Burg method using the IPPC results in spectral estimates that are nearly identical except in regions where tones are present.

stantially lower than in the uncontaminated case. For instance, these peaks appear to be about 7 dB lower as compared to the classical estimate with no contamination.

Next, we examine the performance of the spectral estimates obtained using the robust IPPC and IMRE-based Burg methods. These robust spectral estimates and the original uncontaminated classical Burg method estimates are plotted together. For example, it is seen in Fig. 7.5 that the robust IPPC-based Burg spectral estimator preserves much of the spectrum. The observed peak levels of the stronger signal closely match that of the uncontaminated case. We again see a weaker pilot tone at approximately 602.3 MHz. Unlike the classical estimator, the weaker signal's spectrum is still present using the robust estimator. However, there is some degradation at the lower part of the weaker signal near 614 MHz. We observe the same kind of performance for the IMRE-based spectral estimates in Fig. 7.6.

This analysis revealed two important properties of the proposed robust parametric spectral estimator based on the Burg algorithm. The robust versions of the Burg method using the

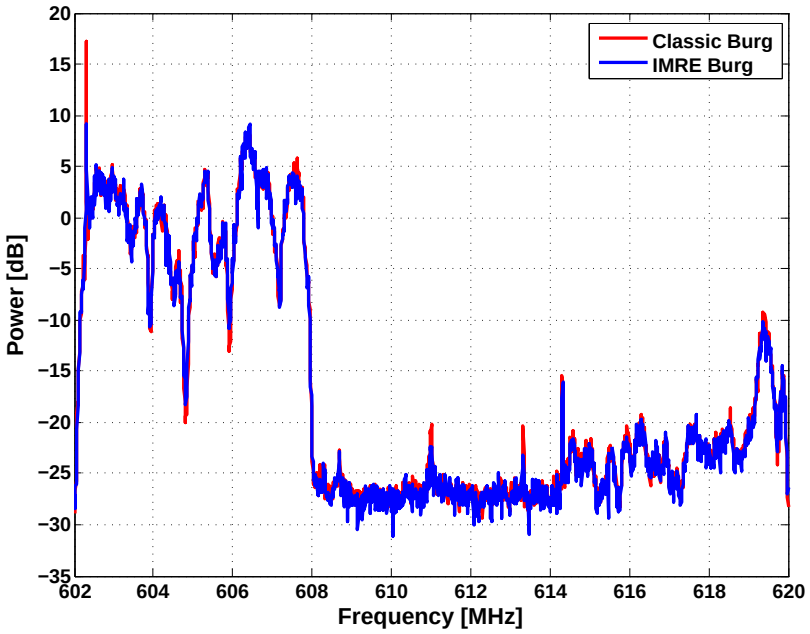

Figure 7.3: The classic Burg and robust Burg method using the IMRE results in spectral estimates that are nearly identical except in regions where tones are present. The IMRE suffers slightly more bias in the regions of tones as compared to the IPPC-based burg algorithm.

IPPC and IMRE as replacements for the reflection coefficient estimator appears to yield robust estimates on collected DTV data. This is important because the DTV signals are not a Gaussian processes. However, multiple DTV signals in the presence of real-world effects such as multi path can be modeled as approximately Gaussian. Our QQ-plots created from collected data demonstrate this is a good model. Also, some degradation is observed with these proposed estimators in sections of the spectrum where narrowband tones are present. However, in general this estimator preserves much of the spectrum as compared to the conventional Burg method implementation in the presence of impulsive noise.

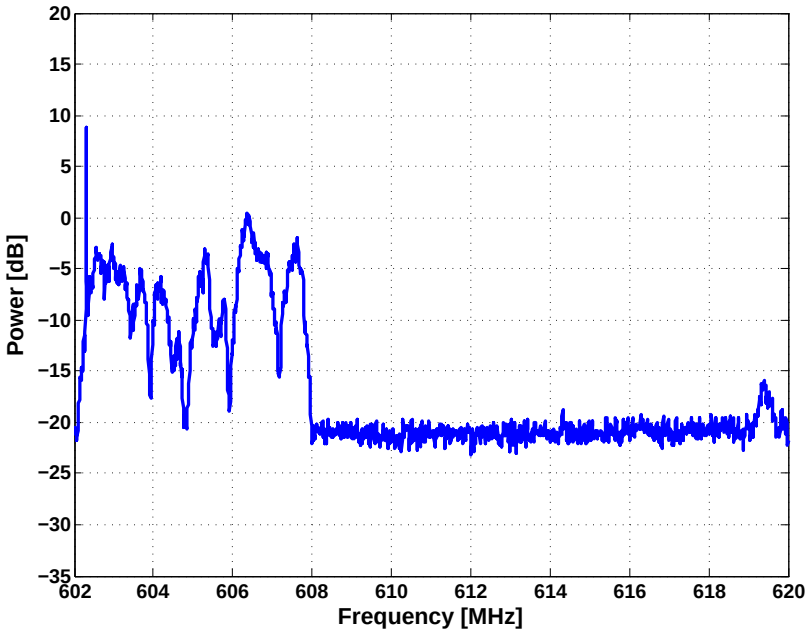

Figure 7.4: The classic Burg method produces a biased spectral estimate when performed on the DTV data with artificially inserted impulsive noise. The second channel centered at 617 MHz is not visible except for a peak in the very upper region of the spectrum.

7.3 Results of the Multi-Taper Method Applied to Robustly Cleaned Data Using Weighted Multiple-Basis Signal Representation

The non-parametric spectral estimators are now applied to the same collected data described in Section 7.1. The data records and contamination processes are identical to those generated in Section 7.2. Two separate results are obtained for the weighted basis pursuit impulsive noise suppression techniques, one using the robust Mahalanobis distances obtained from the IPPC and the other from the IMRE. Both estimates will be shown on the same plots, and are nearly indistinguishable.

To establish a performance baseline, spectral estimates are performed on data without impulsive noise. Similar to before, this tests the impulsive noise removal part of the algorithm.

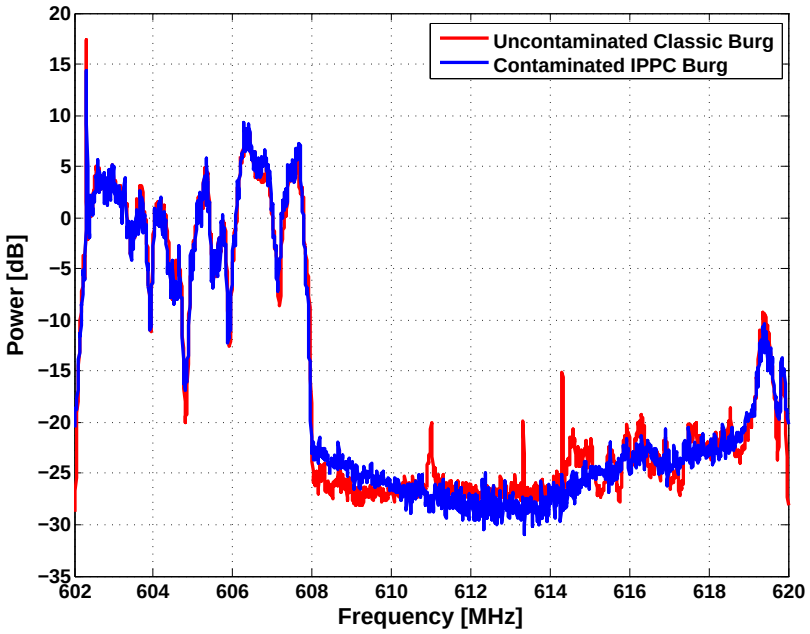

Figure 7.5: The robust IPPC-based Burg method applied to the $\varepsilon = 0.01$ contaminated DTV data preserves much of the signal. The weaker signal is mostly visible with only the lower extreme part of the spectrum biased low.

We wish to preserve the performance of the MTM when the data is uncontaminated. Observe in Fig. 7.7 that the classical implementation of the MTM and the one using the impulsive noise suppression technique based on basis pursuit produce different spectral estimates. The two methods produce similar results for the regions where the stronger signal is present. However, observe that the estimators produce different results for the weaker signal. The levels observed for the weaker signal are less for the basis pursuit-based algorithm except for the higher frequencies near 619 MHz. Also, the center of the spectrum is much lower for the basis pursuit-based estimator. Examining the results in Fig. 7.8, we see that the MTM based on the weighted basis pursuit noise suppression algorithms produce nearly identical results over the entire spectrum. The weighted basis pursuit noise suppression technique preserves the performance of the MTM in the absence of impulsive noise. This is a desirable result.

Now, we test the estimators in the presence of impulsive noise. We start our analysis with a

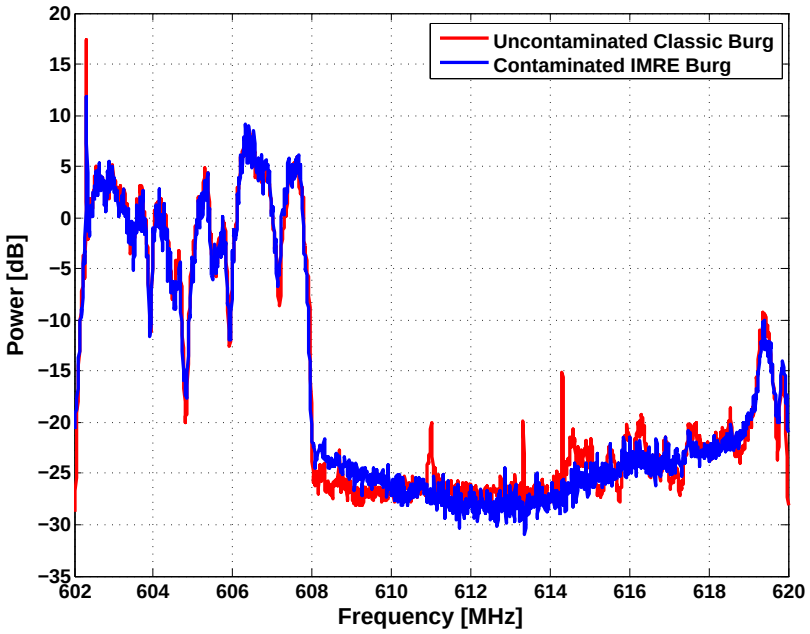

Figure 7.6: The robust IMRE-based Burg method applied to the $\varepsilon = 0.01$ contaminated DTV data preserves much of the signal. Again, as with the IPPC based algorithm, the weaker signal is mostly visible with only the lower extreme part of the spectrum biased low.

contamination rate of $\varepsilon = 0.01$ First, the MTM estimator without impulsive noise suppression is seen to produce a severely biased estimate of the spectrum. Examining Fig. 7.9 reveals that there is an elevated noise floor. While the peaks of the stronger signal match the uncontaminated case, we can see that the nulls in the spectrum are not as deep, for example examine the null at approximately 604.8 MHz. In addition, only the very upper edge of the weaker signal is visible above the elevated noise floor.

We now examine the results of the MTM after using the basis pursuit-based impulsive noise suppression technique. Once again, while not as pronounced, the spectral nulls are not as deep in the stronger signal. In general, the lower part of the weaker signal is biased low, and only the very highest part of the spectrum of the weaker signal matches the original MTM estimate. Additionally, the noise level in the center channel is biased low by more than 5 dB. Inspecting the MTM estimates after using the weighted basis pursuit noise suppression technique reveals excellent results. In fact, the difference between these results and the

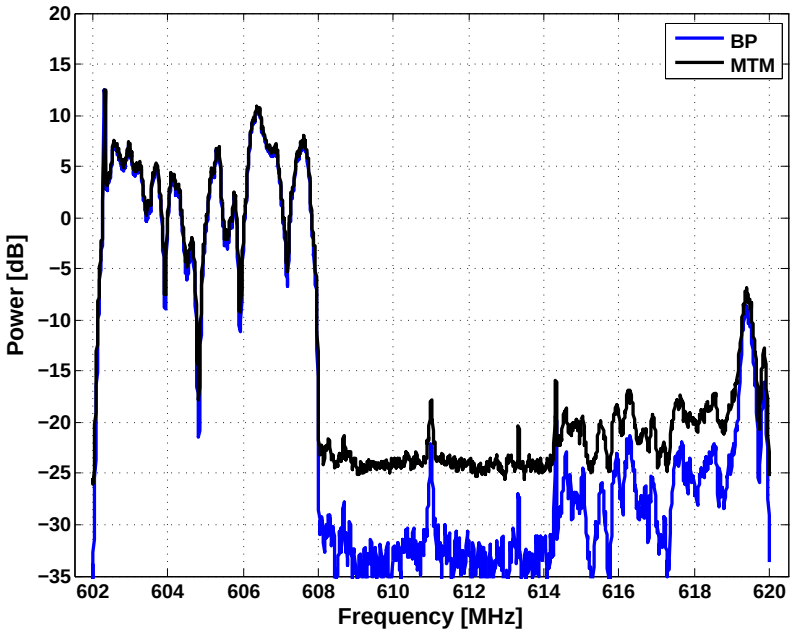

Figure 7.7: The MTM estimator is applied to uncontaminated DTV data after basis pursuit impulsive noise suppression is performed. We observe that the spectral estimate is biased low as compared to the classic MTM estimator.

original MTM without contamination are negligible.

We exercise the estimators further, now increasing the contamination rate to $\varepsilon = 0.05$. Similar results are once again produced. Once again, we observe that the MTM experiences severe bias under contamination. Now, we see an even higher noise floor. Under this contamination level, it would be difficult to even detect the weaker channel.

The MTM estimate after using the basis pursuit noise suppression technique again performs better than the classic MTM estimator, but it is biased low. For instance, the lower half of the weaker signal's spectrum is lower than the original uncontaminated MTM estimate. The MTM estimate after impulsive noise suppression using weighted basis pursuit performs very well again. Only a slight bias is noticed as compared to the original MTM estimate. Once again, we find that the MTM estimate using weighted basis pursuit impulsive noise suppression is closer to the original uncontaminated MTM estimate.

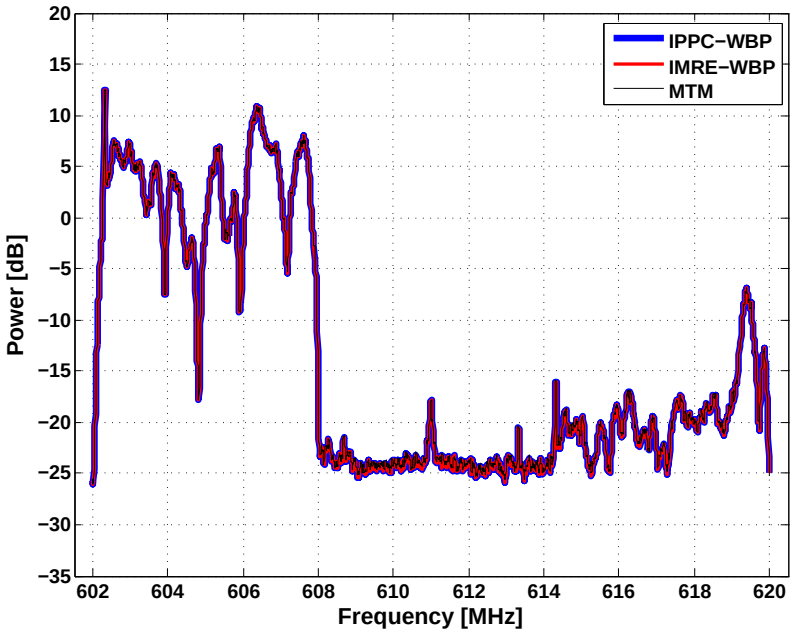

Figure 7.8: The MTM estimator is applied to uncontaminated DTV data after weighted basis pursuit impulsive noise suppression is performed. We observe that the spectral estimate is nearly identical to the classic MTM estimator. The weighted basis pursuit MTM estimator performs better than the basis pursuit MTM estimator.

Clearly, using multiple-basis signal representation to suppress impulsive noise is a very powerful signal processing technique. It was shown that the use of proper weighting of the spike dictionary using the robust Mahalanobis distances obtained from the IPPC or IMRE can improve the performance of this noise suppression technique. Additionally, we observed better performance from a bias perspective using these techniques when compared to the Burg method results shown in the previous section.

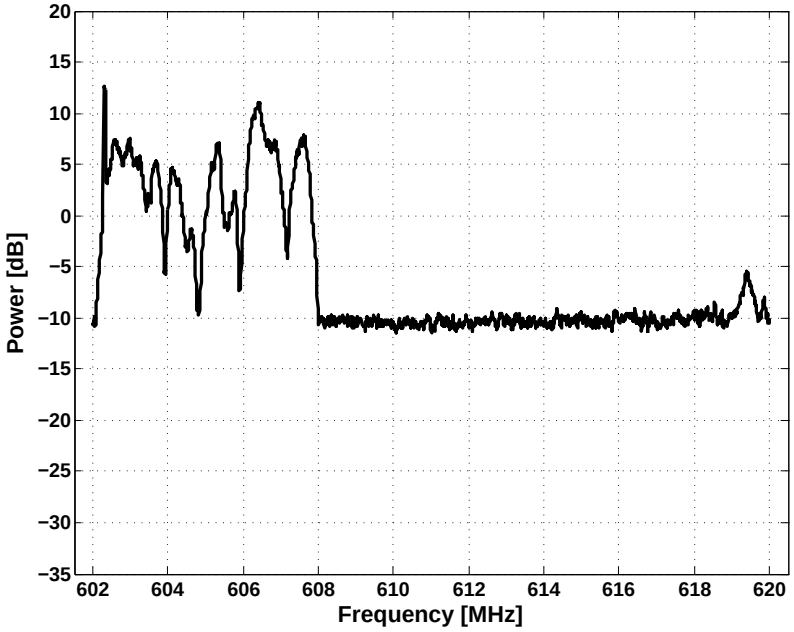

Figure 7.9: The classic MTM estimator is applied to $\varepsilon = 0.01$ contaminated DTV data, producing a biased estimate. The nulls in the spectrum of the stronger signal are not as deep as in the uncontaminated case, and the second signal is not visible expect for the very upper part of the spectrum.

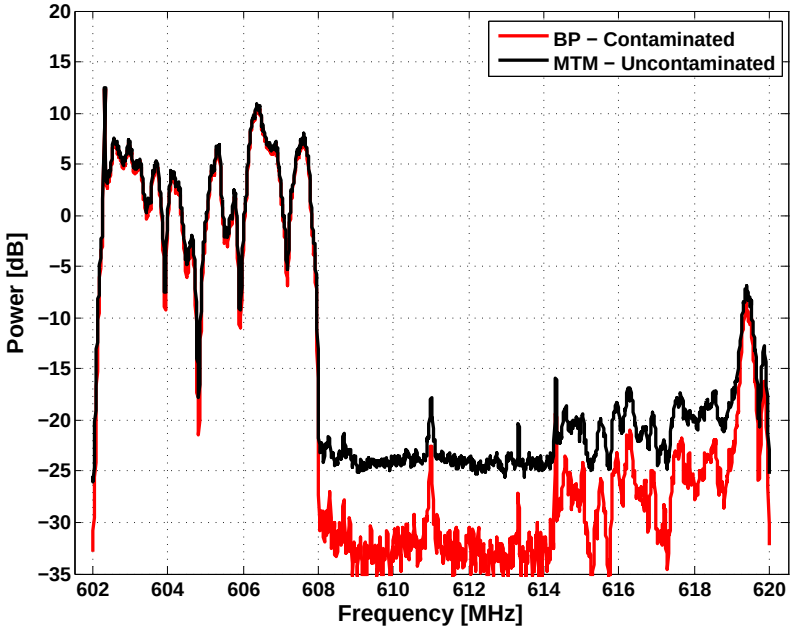

Figure 7.10: The MTM estimate after basis pursuit impulsive noise suppression reduces the bias as compared to the classic implementation. Notable bias exists for the weaker signal.

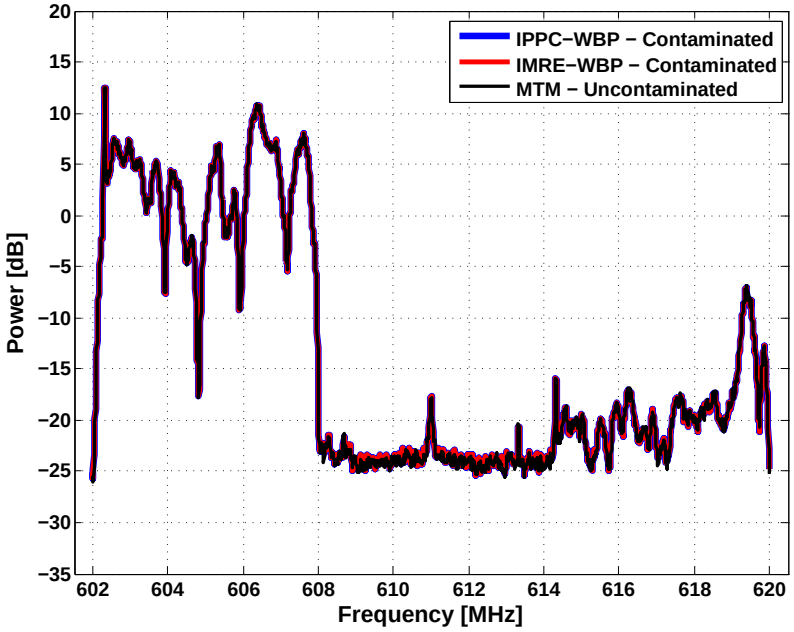

Figure 7.11: The MTM estimate after weighted basis pursuit impulsive noise suppression restores the performance to the uncontaminated case. We observe no noticeable bias for this contamination rate.

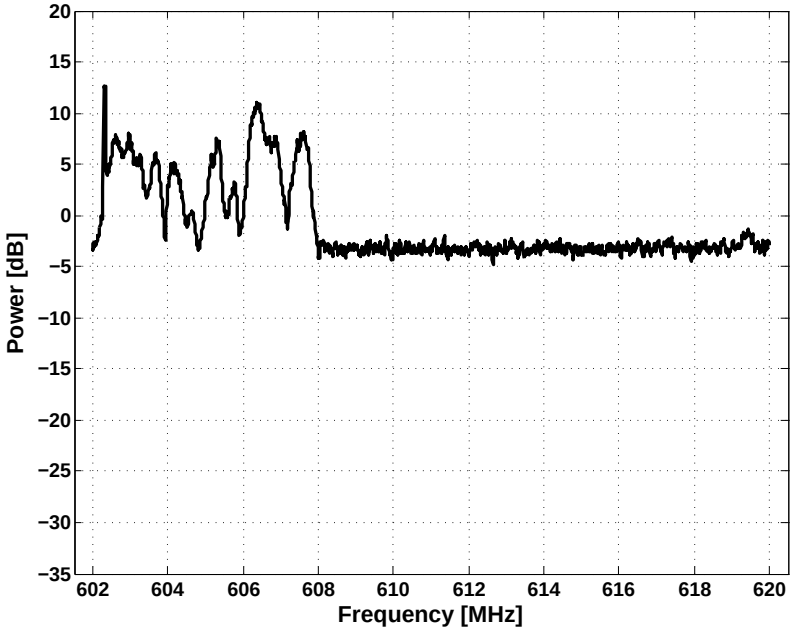

Figure 7.12: The MTM estimator applied to $\varepsilon = 0.05$ contaminated data is biased, displaying a higher noise floor than in the $\varepsilon = 0.01$ case seen in Fig. 7.9.

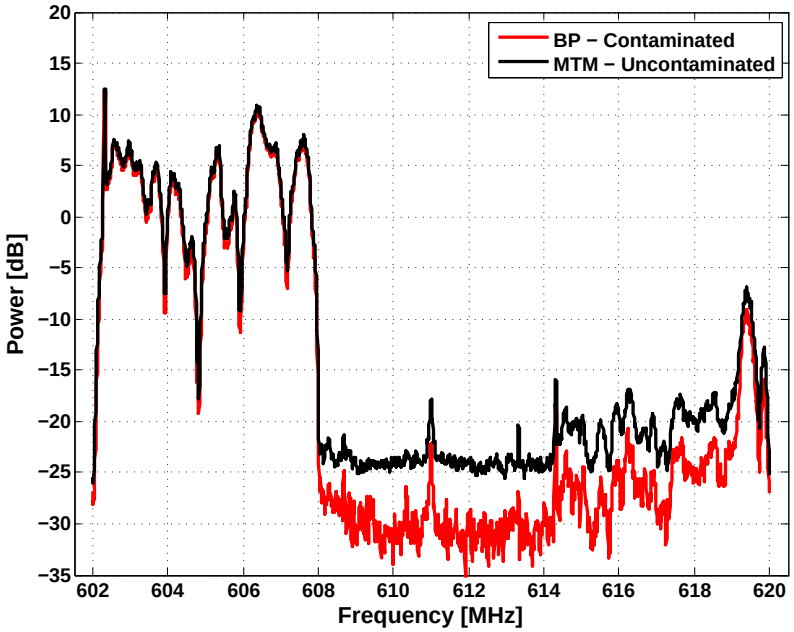

Figure 7.13: The MTM estimator applied to $\varepsilon = 0.05$ contaminated DTV data produces estimates that are very similar to the $\varepsilon = 0.01$ case.

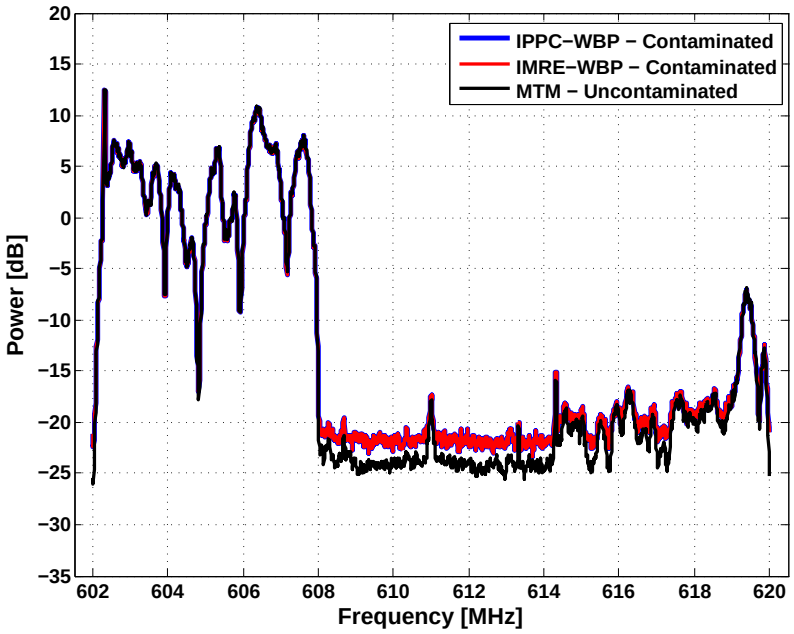

Figure 7.14: Applying the MTM estimator to $\varepsilon = 0.05$ contaminated DTV data performs very well, exhibiting less bias than after basis pursuit impulsive noise suppression. The performance is not as good as observed in Fig. 7.11, but the weaker signal is visible with much less bias in the spectral estimate than seen in Fig. 7.10.

Chapter 8

Conclusions and Future Work

This work provides a robustness analysis of two nonlinear estimators of correlation for ε-contaminated complex-valued Gaussian random processes. These estimators are attractive from both a bias and computational complexity perspective as compared to the robust SHGM using robust projection statistics. Iterative versions of these estimators were developed. These estimators are very applicable to time series analysis. Furthermore, a parametric spectral estimation procedure was developed using the Burg method to estimate the autoregressive parameters of a time series. Additionally, a non-parametric spectral estimation technique was developed which incorporated robust Mahalanobis distances, a byproduct of the iterative versions of these estimators, with weighted basis pursuit to improve impulsive noise suppression using multiple basis signal representations. Both of these robust spectral estimation techniques proved to be effective when applied to data containing multiple signals with drastically differing power levels. This research has revealed several interesting areas of future research which will be discussed next.

The linear regression model can be used to estimate the correlation coefficient of a time series for different time lags. Robust estimates are obtained using the robust SHGM. A requirement of the robust SHGM is the identification of bad leverage points in order to provide unbiased estimates. If the underlying process is known to be Gaussian, the robust

Mahalanobis distances obtained from the IPPC or IMRE could be used as replacements for the robust projection statistics. This would provide a faster method to identify bad leverage points. Additionally, since the robust projection statistics are known to breakdown at a contamination rate of approximately 0.3 for time series, replacing them with the RMD from the IPPC and IMRE may improve the bias performance of the SHGM at the same time decreasing the computation time.

In this analysis, we have assumed that the observed process is a corrupted complex-valued Gaussian random process. If the underlying distribution of the observed process was not Gaussian, we would speculate that in general the estimators analyzed in this work would suffer some kind of bias. The observed bias would probably depend on how the process deviated from Gaussianity. Future work might focus on the analysis of the usefulness of these estimators when applied to different distributions. A reason to investigate this is to determine if these estimators may be suitable to speech processing. Speech is modeled as a Gaussian process over small segments, but it is known that it closely follows a Laplacian distribution in general.

The Student's t-distribution appears to be an interesting distribution. While not a time series model we have considered, if one were to assume a two-dimensional complex-valued Student's t-distribution, then the PPC would be an excellent estimator of the correlation coefficient between the jointly distributed data. The reason for this is that one generates complex-valued data, $\mathbf{t}$, following a t-distribution by dividing a unit variance complex-valued Gaussian random variable by the square root of a normalized Chi-squared process with N "complex degrees of freedom", that is

$$\mathbf{t} = \frac{\mathbf{z}}{\sqrt{\mathbf{q}/N}}, \tag{8.1}$$

where $\mathbf{z} \sim N_c(0, \Sigma)$, $\Sigma = \begin{bmatrix} 1 & \rho_{z_1,z_2} \\ \rho^*_{z_1,z_2} & 1 \end{bmatrix}$ and $\mathbf{q} \sim \chi^2$ with N complex degrees of freedom. The data, $\mathbf{t}$, is a complex-valued Guassian random process with a magnitude that has been

altered on a sample by sample basis. Note that the phase of the original Gaussian process has been preserved since $\sqrt{\mathbf{q}/N}$ is real-valued and non-negative. Since the PPC discards the magnitude and operates on only the phase, this sample by sample scaling of the magnitude, which causes the thick tails, is removed. The MRE, however, would not produce unbiased estimates in this case.

We would speculate that the performance of the MRE and PPC would "degrade" for other thick-tailed distributions. Quantifying this would obviously have to be made on a case-by-case basis. This degradation may however, only be a bias, much like the bias observed in our analysis when we corrupted the Gaussian distribution with outliers. Perhaps, the bias might be predictable. If a closed form solution of the bias did not exist for the new distributions, one could obtain them empirically through simulations, in a manner similar to our maximum bias simulations. In fact, one can almost view the PPC as having a correctable magnitude bias under Gaussianity, one that is corrected using the inverse of the Gaussian Hypergeometric function.

It was demonstrated that the impulsive-noise suppression method using weighted basis pursuit is an effective data cleaning technique. The method implemented to weight the individual atoms in the spike dictionary is a relatively simple approach. It might be of interest to examine additional weighting strategies to determine if the impulsive noise suppression performance can be improved.

If one wishes to use these analyzed estimators for data records that are not sufficiently large, it might be of interest to examine their small-sample properties. This also would require the analysis of the distribution of the RMDs for small data records. One should verify that our proposed outlier thresholds are still valid. The challenge with this is that the RMDs are obtained using iterative nonlinear methods.

The non-negative definite property of the autocorrelation function might not be guaranteed by the robust estimators analyzed in this work. It is believed that the PPC may be constructed to guarantee this property, but this has not been proven. Additionally, the iterative

versions of these estimators may prove to be more difficult to construct with this property in mind. Additional analysis could be done in this area.

Finally, it would be of interest to incorporate these estimators into other signal processing areas such as speech processing and Wiener filtering. Speech often follows a Laplacian distribution, not Gaussian. Understanding if these estimators would work well for corrupted speech signals would have benefit. Also, these estimators may be useful in generalized sidelobe canceller applications where one is trying to cancel non-Gaussian jamming noise.

Bibliography

[1] F. Abdelkefi, P. Duhamel, and F. Alberge. Impulsive noise cancellation in multicarrier transmission. *IEEE Transactions on Communications*, 53(1):94–106, January 2005.

[2] T. Adali and S. Haykin. *Adaptive Signal Processing: Next Generation Solutions*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley, 2010.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado.

[4] R.J. Baxley, B.T. Walkenhorst, and G. Acosta-Marum. Complex gaussian ratio distribution with applications for error rate calculation in fading channels with imperfect CSI. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5, 2010.

[5] H. Ben Mâad, A. Goupil, L. Clavier, and G. Gelle. Robust clipping demapper for ldpc decoding in impulsive channel. In *Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on*, pages 231–235, sep 2010.

[6] P. Billingsley. *Probability and Measure*. A Wiley-Interscience publication. Wiley, 2012.

[7] Manuel Blum, Robert W. Floyd, Vaughan Pratt, Ronald L. Rivest, and Robert E. Tarjan. Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461, August 1973.

[8] Andreas Brandt and Hans R. Kunsch. On the stability of robust filter-cleaners. *Stochastic Processes and their Applications*, 30(2):253 – 262, 1988.

[9] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, 1991.

[10] P.M.T. Broersen. Finite-Sample Bias in the Yule-Walker Method of Autoregressive Estimation. In *IEEE Instrumentation and Measurement Technology Conference Proceedings*, pages 342–347, May 2008.

[11] J. Burg. *Maximum Entropy Spectral Analysis*. PhD thesis, Stanford University, 1975.

[12] J. J. Bussgang. Cross-correlation function of amplitude-distorted Gaussian signals. Technical Report 216, Res. Lab. Elec., Mass. Inst. Technol., Cambridge, MA, Mar. 1952.

[13] N. Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):231–237, 1980.

[14] Y. Chakhchoukh. A New Robust Estimation Method for ARMA Models. *IEEE Transactions on Signal Processing*, 58(7):3512–3522, July 2010.

[15] Y. Chakhchoukh, P. Panciatici, and P. Bondon. Robust estimation of sarima models: Application to short-term load forecasting. In *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, pages 77–80, 2009.

[16] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[17] Shaobing Chen and David Donoho. Basis pursuit. Technical report, Stanford University, 1994.

[18] C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

[19] D. L. Donoho. *Breakdown properties of multivariate location estimators.* PhD thesis, Harvard University, 1982.

[20] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 12 1992.

[21] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845 –2862, nov 2001.

[22] F.Y. Edgeworth. On a new method of reducing observations relating to several quantities. *Philosophical Magazine Series 5*, 25(154):184–191, 1888.

[23] O.D. Escoda, L. Granai, and P. Vandergheynst. On the use of a priori information for sparse signal approximations. *IEEE Transactions on Signal Processing*, 54(9):3468–3482, 2006.

[24] FCC. ATSC Digital Television Standard A/53, 1996.

[25] A. J. Fox. Outliers in time series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):350–363, 1972.

[26] L. Galin and H. Messer. Robust processing of heavy tails signals-comparison of approaches. In *Statistical Signal and Array Processing, 1996. Proceedings., 8th IEEE Signal Processing Workshop on (Cat. No.96TB10004*, pages 230–233, jun 1996.

[27] M.A. Gandhi and L. Mili. Robust kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing*, 58(5):2509–2520, May 2010.

[28] M.G. Genton and A. Lucas. Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1):81–94, 2003.

[29] M. Ghosh. Analysis of the effect of impulse noise on multicarrier and single carrier QAM systems. *IEEE Transactions on Communications*, 44(2):145–147, feb 1996.

[30] N. R. Goodman. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1):152–177, 1963.

[31] L. Granai and P. Vandergheynst. Sparse decomposition over multi-component redundant dictionaries. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*, pages 494–497, 2004.

[32] H. L. Gray, G. D. Kelley, and D. D. Mc Intire. A new approach to ARMA modeling. *Communications in Statistics - Simulation and Computation*, 7(1):1–77, 1978.

[33] Xin Guo, Hongbo Sun, and Y. Lu. Joint suppression of radio frequency interference and lightning impulsive noise in hfswr. In *Radar Conference - Surveillance for a Safer World, 2009. RADAR. International*, pages 1–6, Oct 2009.

[34] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, New York, 1st edition edition, January 1986.

[35] S. Haykin, D.J. Thomson, and J.H. Reed. Spectrum sensing for cognitive radio. *Proceedings of the IEEE*, 97(5):849–877, May 2009.

[36] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):pp. 73–101, 1964.

[37] Peter J. Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):pp. 1041–1067, 1972.

[38] Peter J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, 1981.

[39] G. Jacovitti and A. Neri. Estimation of the autocorrelation function of complex Gaussian stationary processes by amplitude clipped signals. *IEEE Transactions on Information Theory*, 40(1):239–245, Jan. 1994.

[40] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[41] S.A. Kassam and H.V. Poor. Robust techniques for signal processing: A survey. *Proc. IEEE*, 73(3):433–481, March 1985.

[42] S.M. Kay. *Modern Spectral Estimation: Theory and Application*. Prentice Hall, New Jersey, 1988.

[43] M.W. Legg, A. Zaknich, A.J. Duncan, and M.V. Greening. Analysis of impulsive biological noise due to snapping shrimp as a point process in time. In *OCEANS 2007 - Europe*, pages 1–6, June 2007.

[44] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.

[45] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[46] R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. J. Wiley, New York, 2006.

[47] MATLAB. *version 7.13.0.564 (R2010b)*. The MathWorks Inc., Natick, Massachusetts, 2011.

[48] J.M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Pearson Education, Upper Saddle River, NJ, 1995.

[49] L. Mili, M.G. Cheniae, N.S. Vichare, and P.J. Rousseeuw. Robustification of the least absolute value estimator by means of projection statistics [power system state estimation]. *Power Systems, IEEE Transactions on*, 11(1):216–225, feb 1996.

[50] Todd K. Moon and Wynn C. Stirling. *Mathematical methods and algorithms for signal processing*. Prentice Hall, Upper Saddle River, NJ, 2000.

[51] Y. Morishita, Y. Tsuda, T. Fujii, and T. Shimamura. An LMS adaptive equalizer using threshold in impulse noise environments. In *Proc. IEEE Int. Conf. Telecommunications*, volume 1, pages 578–582, February 2003.

[52] C.R. Obranovich, J.M. Golusky, R.D. Preuss, D.R. Fabbri, D.R. Cruthirds, E.M. Aylward, J.A. Freebersyser, and S.R. Kolek. 300 bps Noise Robust Vocoder. In *Military Communications Conference, 2010 - MILCOM 2010*, pages 298–303, November 2010.

[53] E. Ollila and V. Koivunen. Robust antenna array processing using m-estimators of pseudo-covariance. In *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, volume 3, pages 2659–2663 vol.3, Sept 2003.

[54] C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley-Interscience, New York, 1973.

[55] I. Reed. On the use of laguerre polynomials in treating the envelope and phase components of narrow-band Gaussian noise. *IEEE Transactions on Information Theory*, 5(3):102–105, September 1959.

[56] R. Remmert. *Theory of Complex Functions*. Graduate Texts in Mathematics / Readings in Mathematics. Springer New York, 1991.

[57] S. O. Rice. Mathematical analysis of random noise. Technical Report 3, Bell Labs Technical Journal, New York, Mar. 1944.

[58] Ethan H Roth, John A. Hildebrand, Sean M. Wiggins, and Donald Ross. Underwater ambient noise on the Chukchi sea continental slope from 2006-2009. *The Journal of the Acoustical Society of America*, 131(1):104–110, 2012.

[59] Peter J. Rousseeuw and Bert C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):pp. 633–639, 1990.

[60] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. J. Wiley, New York, 1987.

[61] F. G. Safar. Signal compression and reconstruction using multiple bases representation. Master's thesis, Virginia Polytechnic Institute and State University, 1990.

[62] M.G. Sanchez, L. de Haro, M.C. Ramon, A. Mansilla, C.M. Ortega, and D. Oliver. Impulsive noise measurements and characterization in a UHF digital TV channel. *IEEE Transactions on Electromagnetic Compatibility*, 41(2):124–136, May 1999.

[63] I. Santamaria, P.P. Pokharel, and J.C. Principe. Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization. *IEEE Transactions on Signal Processing*, 54(6):2187–2197, June 2006.

[64] A.H. Sayed. *Fundamentals of Adaptive Filtering*. Wiley, New Jersey, 2003.

[65] G. Schoenig. *Contributions to Robust Adaptive Signal Processing with Application to Space-Time Adaptive Radar*. PhD thesis, Virginia Polytechnic Institute and State University, 2007.

[66] G.N. Schoenig, M.L. Picciolo, and L. Mili. Improved detection of strong nonhomogeneities for STAP via projection statistics. In *2005 IEEE International Radar Conference*, pages 720–725, May 2005.

[67] P.J. Schreier and L.L. Scharf. *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge University Press, 2010.

[68] R.J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley, New York, 2002.

[69] W. A. Stahel. Breakdown of covariance estimators. Technical Report 31, Fachgruppe fur Statistik, Zurich, Mar. 1981.

[70] Christoph Studer and Richard G. Baraniuk. Stable restoration and separation of approximately sparse signals. *CoRR*, abs/1107.0420, 2011.

[71] P. Tamburello and L. Mili. Robustness analysis of the phase-phase correlator to white impulsive noise with applications to autoregressive modeling. *IEEE Transactions on Signal Processing*, 60(11):6053 –6058, November 2012.

[72] P. Tamburello and L. Mili. New robust estimators of correlation and weighted basis pursuit. *IEEE Transactions on Signal Processing*, 63(4):882–894, Feb 2015.

[73] Howard G. Tucker. A generalization of the glivenko-cantelli theorem. *The Annals of Mathematical Statistics*, 30(3):828–830, Sep 1959.

[74] P.P. Vaidyanathan. *The Theory of Linear Prediction*. Synthesis Lectures on Engineering Series. Morgan & Claypool, 2008.

[75] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[76] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley-Interscience, 1 edition, September 2001.

[77] B. Widrow. A study of rough amplitude quantization by means of nyquist sampling theory. *IRE Transactions on Circuit Theory*, 3(4):266–276, dec 1956.

[78] M. Zimmermann and K. Dostert. Analysis and modeling of impulsive noise in broadband powerline communications. *IEEE Transactions on Electromagnetic Compatibility*, 44(1):249–258, Feb 2002.

[79] Y. Zou, S.C. Chan, and T.S. Ng. A recursive least M-estimate (RLM) adaptive filter for robust filtering in impulse noise. *IEEE Signal Processing Letters*, 7(11):324–326, November 2000.

[80] Yuexian Zou, Shing-Chow Chan, and Tung-Sang Ng. Least mean M-estimate algorithms for robust adaptive filtering in impulse noise. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(12):1564–1569, December 2000.